Rethinking Exploration in RLVR: From Entropy Regularization to Refinement via Bidirectional Entropy Modulation

Anonymous ACL submission

Abstract

Reinforcement learning with verifiable rewards (RLVR) has significantly advanced the reasoning capabilities of large language models (LLMs). However, it faces a fundamental limitation termed *restricted exploration*, where the policy rapidly converges to a narrow set of solutions. While entropy regularization is a popular approach used to sustain exploration, it often proves unreliable for LLMs, suffering from high hyperparameter sensitivity and yielding only marginal performance gains. Motivated by these inefficiencies, we propose to rethink the relationship between policy entropy and exploration. By deriving a parametric formulation of group-relative advantage estimation and analyzing entropy dynamics, we conceptually decompose policy entropy into *informative entropy*, which preserves diverse solution paths, and *spurious entropy*, which erodes reasoning patterns. Our analysis reveals that, in contrast to blind maximization, effective exploration requires *entropy refinement*—a mechanism implicitly embedded in group-relative advantage estimation that sustains informative entropy on positive rollouts while suppressing spurious entropy on negative ones. Guided by this insight, we propose **AsymGRPO**, an exploratory framework that explicitly decouples the modulation of positive and negative rollouts. This allows for independent control over the preservation of informative entropy and the suppression of spurious noise. Extensive experiments demonstrate that AsymGRPO achieves superior performance compared to strong baselines and exhibits the potential to synergize with existing entropy regularization methods.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has recently emerged as a promising post-training paradigm (Zhang et al., 2025; Lambert et al., 2024; Wen et al., 2025; Mroueh, 2025; Wen et al., 2025; Lv et al., 2025). By leveraging programmatic feedback via automated verifiers, RLVR effectively alleviates reward-model overoptimization ("reward hacking") (Miao et al., 2024; Gao et al., 2023) and enables verification-guided solution exploration for large language models (LLMs) (Setlur et al., 2024b; Wang et al., 2025b), thereby improving performance on challenging reasoning tasks, such as mathematics and coding (Gehring et al., 2024; Setlur et al., 2024a).

Despite its success, RLVR faces a fundamental limitation termed *restricted exploration*, often manifesting as *entropy collapse* (Cui et al., 2025; Yu et al., 2025; Yue et al., 2025): In the early stage of training, the policy becomes overconfident in a narrow set of solutions, causing its entropy to drop sharply. This suppression of alternative reasoning strategies inevitably leads to premature performance saturation. To mitigate this, most studies propose enforcing entropy regularization in the training objective (Wang et al., 2025c; He et al., 2025), attempting to artificially raise policy entropy with the expectation of sustaining exploration.

However, recent studies have revealed that entropy regularization is less effective for LLM-RL than in conventional RL (Haarnoja et al., 2018; Schulman et al., 2017). It is highly hyperparameter-sensitive, prone to entropy explosion that yields near-uniform and semantically uninformative policies, and often provides only marginal performance gains (Jiang et al., 2025; Shen, 2025; He et al., 2025), rendering it an unstable and unreliable intervention. Given these pervasive inefficiencies, a critical yet overlooked question arises:

Does simply increasing policy entropy truly guarantee improved exploration, or is a more nuanced mechanism required?

To answer this question, we conduct a rigorous analysis of entropy dynamics during RL training. Using group-relative advantage estimation (Shao et al., 2024) as a probe, we derive its continuous, parametric formulation to enable fine-

grained control and ablation of policy update dynamics. Through systematic performance comparisons, mechanistic analysis, and adversarial entropy flipping experiments, we conceptually decompose policy entropy into two distinct types: *informative entropy*, which facilitates effective exploration by preserving diverse solution paths, and *spurious entropy*, which tends to erode salient reasoning patterns by introducing unnecessary noise. With this distinction, we reveal that group-relative advantage estimation functions as an implicit ***entropy refinement*** mechanism: it sustains informative entropy on positive rollouts while suppressing spurious entropy on negative ones, synergistically driving higher performance. This finding clarifies that:

Effective exploration requires precise entropy refinement rather than the blind maximization inherent in naïve entropy regularization.

Guided by this insight, we propose an exploratory framework termed Asymmetric Group-Relative Policy Optimization **(AsymGRPO)** to investigate precise entropy refinement. Formulated as a parametric generalization of group-relative estimation, AsymGRPO explicitly decouples the modulation of positive and negative rollouts, allowing for independent control over the intensity of informative entropy sustainment and spurious entropy suppression. Experiments on five mathematical reasoning benchmarks demonstrate that AsymGRPO achieves highly competitive performance compared to strong baselines and exhibits the potential to collaborate with existing entropy-regularized methods for further performance gains.

2 Mechanistic Analysis of Entropy Dynamics in Group-Relative Policy Optimization

In this section, we formulate the RLVR framework and deconstruct the Group-Relative Policy Optimization (GRPO) algorithm. By generalizing standard GRPO into a parametric form and analyzing it as a reweighting mechanism, we uncover its inherent capability for bidirectional entropy modulation.

2.1 RLVR Formulation and Parametric Generalization of Group-Relative Advantages

Reinforcement learning with verifiable rewards (RLVR) encourages models to develop long, deliberate chains of thought, thereby substantially improving reasoning accuracy. Given an LLM policy π_θ, the standard objective is to maximize the expected reward of sampled responses:

$$\max_\theta \mathcal{J}_{\mathrm{RLVR}}(\theta) \;=\; \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|x)}\big[\, r(x,y)\,\big], \tag{1}$$

where x is a prompt sampled from dataset $\mathcal{D}$, y is a rollout generated by π_θ, and $r(x,y)\in\{0,1\}$ is a binary verifiable reward indicating correctness.

PPO-style surrogate objective. To optimize (1), RLVR methods typically employ a PPO-style clipped surrogate objective (Schulman et al., 2017):

$$\mathcal{J}_{\mathrm{PPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_{\theta_{\mathrm{old}}}(\cdot|x)} \left[\frac{1}{T}\sum_{t=1}^{T}\min\Big(\rho_t(\theta)\,A_t,\, \mathrm{clip}\big(\rho_t(\theta),1-\epsilon,1+\epsilon\big)A_t\Big)\right]. \tag{2}$$

where $\rho_t(\theta)=\frac{\pi_\theta(y_t|x,y_{<t})}{\pi_{\theta_{\mathrm{old}}}(y_t|x,y_{<t})}$ is the importance ratio, T is the length of rollout y, and ϵ is the clipping hyperparameter. A_t denotes the token-level advantage, which is typically estimated by a value network in standard PPO. In reasoning tasks with sparse rewards, the outcome reward is typically assigned to all tokens in the trajectory (Guo et al., 2025; Liu et al., 2025), such that A_t takes the value of the rollout-level advantage A_{rollout} for all t.

Entropy Regularization. Standard RL methods often augment the PPO objective with an entropy bonus to encourage exploration. Mathematically, the entropy of the current policy π_θ over the vocabulary $\mathcal{V}$ at timestep t is defined as:

$$\mathcal{H}_t(\pi_\theta) \;=\; -\sum_{v\in\mathcal{V}}\pi_\theta(v\mid x,y_{<t})\log\pi_\theta(v\mid x,y_{<t})\,. \tag{3}$$

Group Relative Policy Optimization (GRPO). GRPO (Shao et al., 2024) estimates the rollout-level advantage A_{rollout} using group statistics without a value network. For each prompt x, it samples a group of G rollouts $\{y_i\}_{i=1}^G$ from the old policy and computes the advantage by standardizing rewards against the group statistics. This significantly reduces memory and computational costs:

$$A_i^{\mathrm{GRPO}} = \frac{r(x,y_i) \;-\; \mathrm{mean}\big(\{\,r(x,y_j)\,\}_{j=1}^G\big)}{\mathrm{std}\big(\{\,r(x,y_j)\,\}_{j=1}^G\big)}\,. \tag{4}$$

We refer to A_i^{GRPO} as the *group-relative advantage*, as it evaluates the quality of each rollout relative to its peers within the same prompt-level group.

The foundational REINFORCE formulation. The most elementary form of advantage estimation, the REINFORCE algorithm (Williams, 1992),

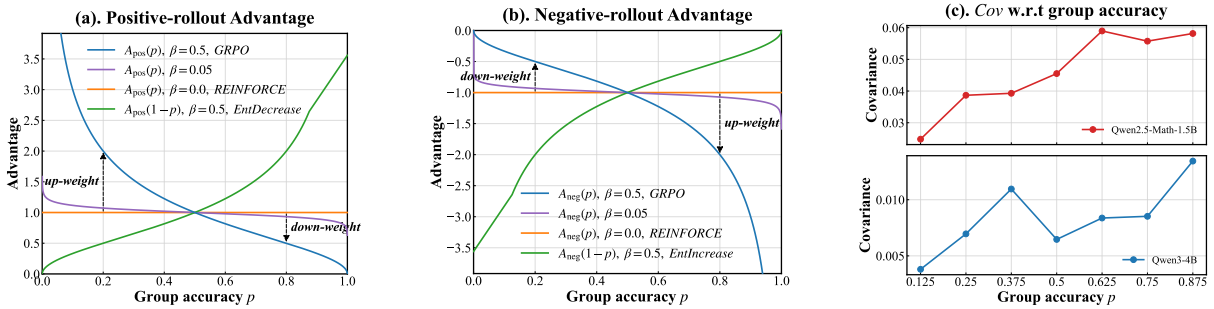

Figure 1: (a) Positive rollout advantage w.r.t. group accuracy. (b) Negative rollout advantage w.r.t. group accuracy. (c) Estimated sample covariance w.r.t. group accuracy collected during the first 40 training steps.

employs a group-independent constant baseline. Standard implementations typically assign a fixed scalar value based solely on the binary outcome. A standard practice in RLVR is to set the baseline $b = 0.5$ and rescale the rewards (Zhu et al., 2025; Peng et al., 2025), yielding:

$$A_i^{\text{REINFORCE}} \;=\; 2\,r(x, y_i) - 1 \;\in\; \{+1, -1\}. \tag{5}$$

In this setting, positive rollouts consistently contribute $+1$ and negative rollouts contribute -1, regardless of the model's current performance. This provides a neutral reference point for analyzing the dynamic properties of group-relative estimators.

GRPO from group accuracy. Under binary rewards $r \in \{0, 1\}$, the group mean equals the in-group accuracy $p = \frac{1}{G}\sum_{j=1}^{G} r(x, y_j)$, and the standard deviation becomes $\sqrt{p(1-p)}$. Consequently, the advantages for positive ($r = 1$) and negative ($r = 0$) rollouts calculated by Eq. (4) can be expressed solely as functions of p:

$$A_{\text{pos}}^{\text{GRPO}}(p) = \sqrt{\frac{1-p}{p}}, \;\; A_{\text{neg}}^{\text{GRPO}}(p) = -\sqrt{\frac{p}{1-p}}. \tag{6}$$

We note that while Eq. (6) is undefined at the boundaries $p \in \{0, 1\}$, these singularities are benign: when $p = 0$, no positive rollouts exist to instantiate A_{pos}, and conversely for $p = 1$.

Parametric generalization of group-relative advantages. To unify the fixed-magnitude advantages (± 1) and the dynamic, accuracy-dependent advantages of GRPO, we introduce a continuous β-parametrized family of advantage functions:

$$A_{\text{pos}}^{(\beta)}(p) = \left(\frac{1-p}{p}\right)^{\beta}, \;\; A_{\text{neg}}^{(\beta)}(p) = -\left(\frac{p}{1-p}\right)^{\beta}. \tag{7}$$

This formulation generalizes the advantage estimation: setting $\beta = 0.5$ recovers the standard GRPO scaling in Eq. (6), while $\beta = 0$ collapses to the constant-magnitude REINFORCE regime in Eq. (7). This parametric view allows us to analyze and control the intensity of the advantage signal based on group accuracy.

2.2 Bidirectional Entropy Modulation via Group-Accuracy Dependent Reweighting

Gradient Reweighting View. We now examine the mechanistic impact of group-relative advantages by shifting focus from variance reduction to gradient reweighting. Omitting the clipping operation for clarity, the effective rollout-level policy gradient derived from Eq. (2) can be expressed as:

$$\nabla_\theta \mathcal{J}(\theta) \;\approx\; \mathbb{E}_{x\sim\mathcal{D},\,\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|x)} \left[\frac{1}{G}\sum_{i=1}^{G} A_i \cdot \left(\frac{1}{T}\sum_{t=1}^{T} \rho_{i,t}(\theta)\,\nabla_\theta \log \pi_\theta(y_{i,t})\right)\right]. \tag{8}$$

Eq. (8) reveals that the magnitude $|A_i|$ functions as a scalar weight scaling the gradient update of rollout i, while the sign determines the direction. Thus, GRPO essentially implements a **rollout reweighting mechanism** dependent on group accuracy p. Compared to a constant baseline (where $|A_i|$ is fixed), Fig. 1(a)-(b) illustrates how GRPO dynamically modulates these weights: for *positive* rollouts, the weight $|A_i|$ decreases as p increases; for *negative* rollouts, the weight $|A_i|$ increases as p rises. The hyperparameter β explicitly controls the intensity of this relative deviation from the constant baseline.

Bidirectional Entropy Dynamics. To link this reweighting to entropy, we consider the entropy change under natural policy gradient in a single-step bandit approximation (Kakade, 2001; Proof in Cui et al., 2025). For a given prompt x, the change is governed by the covariance:

$$\Delta\mathcal{H}(\pi(\cdot|x)) \;\approx\; -\eta \cdot \text{Cov}_{y\sim\pi(\cdot|x)}\big(\log \pi(y|x),\, A(y)\big). \tag{9}$$

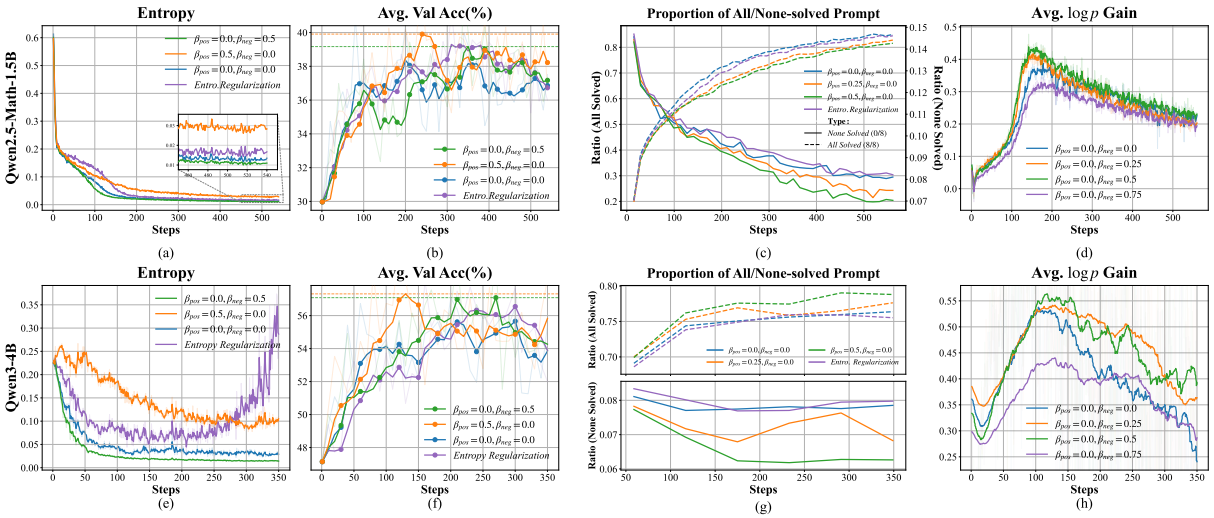

Figure 2: **Evolution of training dynamics and mechanism analysis.** The top row presents results on **Qwen2.5-Math-1.5B**, while the bottom row corresponds to **Qwen3-4B**. **(a, e)** Policy entropy over training steps. **(b, f)** Average validation accuracy. **(c, g)** The epoch-wise proportion of prompts categorized as "all-solved" and "none-solved". **(d, h)** The average log probability increment of positive samples after each update.

We estimated the average sample covariance for prompts with different accuracies during RL training (details in Appendix B). As shown in Fig. 1(c), the covariance correlates positively with p, confirming that high group accuracy implies a strong natural tendency for entropy reduction.

Combining this with the reweighting analysis reveals a **bidirectional entropy modulation**: (1) on positive rollouts, the decaying advantage weight opposes the increasing covariance trend, effectively **sustains policy entropy**; and (2) on negative rollouts, the amplifying penalty aligns with the covariance trend, actively **drives entropy reduction**. We empirically verify these distinct entropy dynamics in the following section.

3 Group-Relative Policy Optimization as a Mechanism for Entropy Refinement

Building on our theoretical analysis that group-relative advantages drive entropy in opposite directions on positive and negative rollouts, we now present empirical evidence demonstrating how this mechanism instantiates *informative* and *spurious* entropy in practice. We examine how these distinct entropy dynamics influence reasoning performance through controlled ablation studies using Qwen2.5-Math-1.5B (Yang et al., 2024) and Qwen3-4B (Yang et al., 2025a). Our experiments track entropy trends and average validation accuracy across multiple mathematical reasoning benchmarks during RL training. To systematically isolate these effects, we instantiate Eq. (7) with separate coefficients for successful and failed rollouts, denoted β_{pos} and β_{neg}, generating a family of advantage variants by varying $(\beta_{\text{pos}}, \beta_{\text{neg}})$. Detailed experimental settings are provided in Appendix C.

3.1 Validating the Refinement Hypothesis: Disentangling Informative and Spurious Entropy

To investigate the fine-grained entropy dynamics and performance effects induced by GRPO, we decompose its group-relative modulation into two parametric variants compared against a constant baseline:

1. **Pos-Only Modulation** $(\beta_{\text{pos}} = 0.5, \beta_{\text{neg}} = 0)$: Applies group-relative reweighting exclusively to positive rollouts.
2. **Neg-Only Modulation** $(\beta_{\text{pos}} = 0, \beta_{\text{neg}} = 0.5)$: Restricts group-relative reweighting solely to negative rollouts.

We employ **REINFORCE** as the reference constant baseline $(\beta_{\text{pos}} = 0, \beta_{\text{neg}} = 0)$ and also evaluate REINFORCE with **Entropy Regularization** using tuned hyperparameters. This decomposition allows us to explicitly disentangle the impact of sustaining entropy on successful rollouts from the impact of suppressing entropy on failed rollouts.

1. Pos-Only Modulation: Sustaining Informative Entropy

► **Observation:** Compared to the REINFORCE baseline, the Pos-Only variant maintains substantially higher policy entropy throughout training (Fig. 2(a,e)). This aligns with the gradient reweighting view established in Sec. 2.2, which posits that

the modulation on positive rollouts **opposes** the natural trend of entropy reduction. Consequently, this mechanism explicitly weakens the force of entropy collapse, effectively reserving exploration budget for uncertain regions. Crucially, this sustained entropy is accompanied by a clear improvement in validation accuracy (Fig. 2(b,f)), indicating that the preserved variability facilitates productive exploration rather than mere noise. We therefore regard the entropy maintained by Pos-Only modulation as **informative entropy**.

↪ **Mechanism Analysis:** To understand how this retained entropy aids reasoning, we track the epoch-wise proportion of "all solved" and "none solved" groups during training (Fig. 2 (c,g)). Across both models, increasing β_{pos} consistently leads to a significant reduction in the fraction of "none-solved" groups, suggesting that the maintained entropy allows the policy to **expand the solvable boundary** into previously intractable regions.

Regarding "all-solved" groups, we observe distinct patterns: while Qwen3-4B shows an increase, Qwen2.5-Math-1.5B exhibits a decrease. For the latter case, this reduction indicates a resistance to overfitting on easy prompts. This behavior aligns with recent findings that over-reinforcing easy instances can induce negative interference that hinders generalization to harder tasks (Nguyen et al., 2025; Yao et al., 2025; Dong et al., 2025). By preventing premature convergence on simple problems, the modulation mitigates such interference and effectively channels the learning budget into productive exploration on difficult queries.

2. Neg-Only Modulation: Pruning Spurious Entropy

► **Observation:** Compared to the REINFORCE baseline, the Neg-Only variant exhibits a marked reduction in policy entropy throughout training, particularly for Qwen3-4B (Fig. 2(a,e)). This observation validates the theoretical insight from Sec. 2.2: the reweighting on negative rollouts **aligns with** the natural tendency for entropy reduction, thereby accelerating the decrease in policy uncertainty. In parallel, validation accuracy improves noticeably (Fig. 2(b,f)), suggesting that the discarded uncertainty serves no functional role in reasoning and does not support productive exploration. We therefore regard the entropy pruned by Neg-Only modulation as **spurious entropy**.

↪ **Mechanism Analysis:** To understand how this entropy pruning affects learning, we track the **average log probability increment** of positive samples after each policy update (calculated as $\mathbb{E}[\log \pi_{\text{new}}(y) - \log \pi_{\text{old}}(y)]$ across all successful rollouts). We observe that increasing the modulation strength from $\beta_{\text{neg}} = 0$ to 0.5 consistently elevates the curve of these likelihood gains. This suggests that GRPO's targeted suppression of spurious entropy mitigates *Lazy Likelihood Displacement* (Deng et al., 2025), where indiscriminate negative gradients on incorrect samples **hinder the effective exploitation of correct solutions**. Such interference arises because incorrect trajectories often share long reasoning prefixes with positive rollouts within the same group; consequently, uniform penalties on failures can inadvertently dampen the probability growth of valid paths (Razin et al., 2024; Ren and Sutherland, 2024). By reducing this destructive interference, negative modulation allows the probability of correct reasoning paths to grow more robustly.

However, a distinct pattern emerges when β_{neg} is further increased to 0.75: the curve drops below the baseline level. We hypothesize that with such an excessively high β_{neg}, the penalties on common error patterns (i.e., groups with low accuracy) become insufficient. This causes the model to settle into overly rigid behaviors on difficult problems (Zhu et al., 2025), which subsequently suppresses the likelihood gains for novel solutions. These results suggest that the negative modulation strength requires careful tuning to effectively prune spurious entropy, thereby avoiding the introduction of harmful or non-functional uncertainty without freezing the model's capacity for improvement.

3. Naïve Entropy Regularization: The Suboptimality of Blind Entropy Inflation

► **Observation:** While the Entropy Regularization baseline successfully raises policy entropy, even with hyperparameter tuning, it fails to match the reasoning accuracy of the Pos-Only modulation (Fig. 2 (b,f)). Examining the group composition metrics, we find that the proportions of "all-solved" and "none-solved" groups remain at similar levels to those in the REINFORCE baseline. This suggests that blindly injecting entropy fails to substantially enhance exploration or extend the solvable boundary, underscoring the need for targeted entropy refinement that treats different sources of entropy separately rather than through a uniform regularization term.

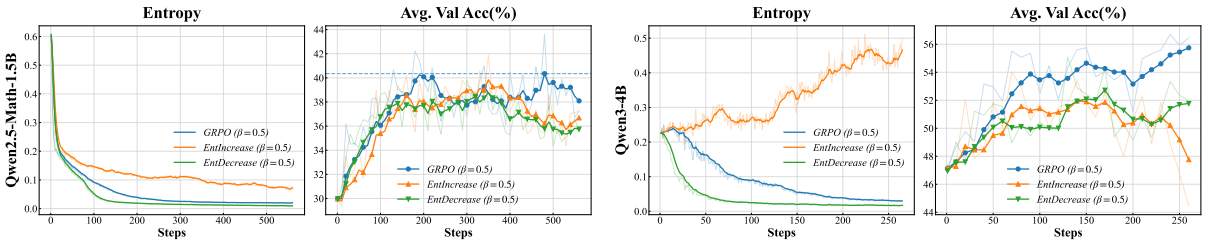

Figure 3: **Adversarial entropy flipping experiments. (a, c)** Policy entropy. **(b, d)** Average validation accuracy.

3.2 Adversarial Analysis: The Necessity of Bidirectional Entropy Modulation

To further verify the existence of informative and spurious entropy, and to assess the necessity of applying opposite modulation to positive and negative rollouts in GRPO, we design an adversarial "flipping" experiment. Based on the parametric advantage formulation in Eq. (7), we construct flipped versions of the advantage curves to **reverse the original reweighting trends** (Fig. 1(a)–(b)). Mathematically, this is achieved by reflecting the advantage function around $p = 0.5$, such that $\tilde{A}(p) = A(1-p)$.[1] With $(\beta_{\text{pos}}, \beta_{\text{neg}})$ fixed at $(0.5, 0.5)$, this construction yields two adversarial variants:

1. **EntDecrease:** Flips the positive-advantage curve while keeping the negative curve unchanged (Fig. 1(a)). By reversing the weighting on positive rollouts, this variant drives a **consistent entropy reduction**.
2. **EntIncrease:** Flips the negative-advantage curve while leaving the positive curve intact (Fig. 1(b)). By reversing the weighting on negative rollouts, this variant promotes a **consistent entropy increase**.

This unification of entropy dynamics allows us to isolate and examine whether the directional entropy modulation inherent to GRPO is indeed critical for performance.

► **Observation:** Compared to GRPO, **EntDecrease** induces a clear reduction in policy entropy throughout training, while **EntIncrease** produces a marked increase in entropy (Fig. 3(a, c)), yet both variants exhibit lower validation accuracy than GRPO and show late-stage degradation in performance (Fig. 3(b, d)). This pattern indicates that suppressing the entropy associated with positive rollouts in EntDecrease removes the informative variability that GRPO maintains, whereas inflating the entropy on negative rollouts in EntIncrease injects additional harmful uncertainty. Taken together, these adversarial flips confirm that GRPO's original design—preserving entropy on successes while reducing entropy on failures—aligns with our notion of *informative* versus *spurious* entropy, and that reversing these roles leads to unstable training and inferior reasoning performance.

TAKEAWAY

- GRPO refines entropy by increasing it on positive rollouts and decreasing it on negative rollouts (Section 3.1).
- Both sustaining informative entropy and pruning spurious entropy improve performance (Section 3.1)
- Indiscriminate entropy modulation that suppresses informative entropy or amplifies spurious entropy degrades accuracy and destabilizes training (Section 3.2).
- **Conclusion**: Blindly maximizing or minimizing entropy is suboptimal. Effective post-training requires **entropy refinement** strategies.

4 Asymmetric Group-Relative Policy Optimization

Motivated by the objective of **entropy refinement**, we move from analysis to algorithmic formulation. While GRPO inherently performs this refinement by applying opposing forces to successful and failed rollouts, enforcing a fixed, symmetric coupling between these forces may limit the flexibility needed for optimal training dynamics.

In this section, we propose an exploratory framework, called **Asymmetric Group-Relative Policy Optimization (AsymGRPO)**, to explicitly decouple the modulation of positive and negative rollouts. Rather than introducing a radically new optimization paradigm, AsymGRPO serves as a parametric generalization of GRPO, enabling more pre-

[1] For completeness, advantages at boundary cases are handled by linearly extending the final segment of the curve; full implementation details are provided in Appendix D.

Table 1: Main experimental results on mathematical reasoning benchmarks. The best result in each column is shown in **bold**, and the second-best is underlined.

Method	MATH-500	AIME24	AIME25	AMC23	Olympiad	Avg.
Qwen3-4B	81.60	21.67	20.00	63.75	47.52	46.91
RLVR Baselines						
REINFORCE	86.60	28.67	24.67	73.75	54.86	53.71
GRPO (Guo et al., 2025)	88.20	31.00	27.33	78.25	57.74	56.50
GRPO w/ Entro.Regularization	88.20	38.33	28.33	75.50	57.24	57.52
GRPO w/ Clip-higher (Yu et al., 2025)	**90.07**	34.67	32.33	78.50	58.18	58.75
GRPO w/ Entro.Adv (Cheng et al., 2025)	86.73	32.00	25.33	77.25	54.46	55.16
Dr.GRPO (Liu et al., 2025)	88.87	36.33	30.00	78.25	57.24	58.14
Pass@K Training (Chen et al., 2025)	86.33	27.67	31.00	74.00	55.06	54.81
Our Methods						
Pos-Only Modulation (§ 3.1)	87.13	27.33	28.00	76.75	57.34	55.31
Neg-Only Modulation (§ 3.1)	87.00	26.00	27.00	78.00	54.46	54.49
EntIncrease (§ 3.2)	85.60	26.00	23.33	71.75	53.03	51.94
EntDecrease (§ 3.2)	83.73	25.00	23.67	71.50	50.74	50.93
AsymGRPO ($\beta_{\text{pos}} = \beta_{\text{neg}}$)	88.53	32.00	29.33	78.50	57.34	57.14
AsymGRPO	89.33	**39.33**	28.67	81.00	58.48	59.36
AsymGRPO w/ Clip-higher	89.73	33.67	**36.00**	**83.25**	**58.93**	**60.32**

cise control over the intensity of entropy refinement—sustaining informative exploration while precisely pruning spurious noise.

4.1 Decoupled Advantage Formulation

To break the fixed and symmetric reweighting constraints of the standard formulation (Eq. 6), we introduce two independent hyperparameters, β_{pos} and β_{neg}. These parameters govern the reweighting intensity for positive and negative samples, respectively. For a group of rollouts $\{y_i\}_{i=1}^{G}$ with group accuracy p, the decoupled token-level advantage estimates $A_{i,t}$ are defined as:

$$A_{i,t}(p) = \begin{cases} A_{\text{pos}}^{(\beta_{\text{pos}})}(p) = \left(\frac{1-p}{p}\right)^{\beta_{\text{pos}}} & \text{if } r(x, y_i) = 1, \\ A_{\text{neg}}^{(\beta_{\text{neg}})}(p) = -\left(\frac{p}{1-p}\right)^{\beta_{\text{neg}}} & \text{if } r(x, y_i) = 0. \end{cases} \tag{10}$$

This formulation recovers the standard REINFORCE baseline when $(\beta_{\text{pos}}, \beta_{\text{neg}}) = (0, 0)$ and the standard GRPO when $\beta_{\text{pos}} = \beta_{\text{neg}} = 0.5$. By setting $\beta_{\text{pos}} \neq \beta_{\text{neg}}$, the algorithm enables an asymmetric modulation strategy, e.g., maintaining a high β_{pos} to boost exploration on rare successes while calibrating β_{neg} to appropriately penalize errors without causing collapse.

4.2 Asymmetric Policy Gradient

To explicitly reflect this separation in the optimization landscape, we decompose the policy gradient into two distinct components summed over the subsets of correct rollouts ($\mathcal{I}^+$) and incorrect rollouts ($\mathcal{I}^-$). The resulting policy gradient (simplified without PPO clipping) is given by:

$$\nabla_\theta \mathcal{J}_{\text{Asym}} = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \left(\underbrace{\sum_{i\in\mathcal{I}^+} \sum_{t=1}^{|y_i|} \nabla_\theta \log \pi_\theta(y_{i,t}|x, y_{i,<t}) \cdot A_{\text{pos}}^{(\beta_{\text{pos}})}}_{\text{Positive Rollout Gradient}} + \underbrace{\sum_{j\in\mathcal{I}^-} \sum_{t=1}^{|y_j|} \nabla_\theta \log \pi_\theta(y_{j,t}|x, y_{j,<t}) \cdot A_{\text{neg}}^{(\beta_{\text{neg}})}}_{\text{Negative Rollout Gradient}} \right) \right]. \tag{11}$$

By decoupling the advantage terms, this formulation allows the optimizer to independently scale the learning signals from informative successes and spurious failures using the **group-level** advantages (A_{pos} and A_{neg}), thereby facilitating the targeted entropy refinement strategy verified in our analysis.

4.3 Main Experimental Results and Analysis

We evaluate the proposed methods by training the **Qwen3-4B** model on the MATH dataset (Hendrycks et al., 2021). To ensure robust evaluation, we report the **Avg@5** accuracy for large datasets (MATH-500, OlympiadBench) and **Avg@10** accuracy for small datasets (AIME 2024, AIME 2025, AMC 2023) with temperature $= 0.4$ and Top-p $= 1.0$. MATH-500 serves as the validation set: for each run, we select the checkpoint achieving the highest validation accuracy and evaluate it on all mathmatical benchmarks. Detailed

hyperparameters and experimental settings are provided in Appendix E.

Table 1 presents the main and ablation comparison results while Fig. 4 presents the visualization of training dynamics of entropy and validation accuracy. Due to space limitations, we provide additional visualizations of the training dynamics in Appendix F. Based on these results, we summarize our key findings as follows:

1. AsymGRPO significantly outperforms baselines by optimizing refinement intensity. AsymGRPO achieves an average accuracy of 59.36%, outperforming the standard GRPO baseline (56.50%) by a substantial margin of 2.86%. Notably, AsymGRPO maintains a policy entropy level comparable to GRPO (Fig. 4), suggesting that the performance gain stems not from simply increasing entropy, but from achieving a superior *entropy refinement*—**effectively allocating training pressure**. Furthermore, AsymGRPO surpasses the strongest baseline, Dr.GRPO (58.14%), by 1.22%, and consistently outperforms various entropy-modified GRPO variants (e.g., Entro.Regularization, Clip-higher). Critically, compared to its own symmetric ablation (symmetric but variable modulation by setting $\beta_{\text{pos}} = \beta_{\text{neg}}$), the decoupled AsymGRPO yields a 2.22% improvement. This result empirically validates **the necessity of the asymmetric formulation**: the optimal intensities for sustaining informative entropy and suppressing spurious entropy are indeed distinct.

2. Our reweighting GRPO variants further confirm the necessity of directional entropy modulation. The ablation results in Table 1 corroborate our empirical analysis in Section 3. Pos-only and Neg-only modulations both outperform the REINFORCE baseline but fall short of the full GRPO, indicating that simultaneous (but opposing) modulation is beneficial. Conversely, the adversarial variants EntIncrease and EntDecrease significantly underperform GRPO. This pattern confirms that GRPO's effectiveness originates from its inherent directional modulation—increasing entropy on successes and decreasing it on failures—and that AsymGRPO amplifies this benefit by **granting greater flexibility to the modulation intensity**.

3. Clip-higher implicitly filters for informative entropy. Among the existing entropy-regularized variants of GRPO, Clip-higher (Yu et al., 2025) demonstrates the strongest performance (58.75%), surpassing naive entropy regularization (57.52%) by 1.23%. We attribute this to its selective nature: unlike naive regularization which indiscriminately inflates global entropy, Clip-higher leverages the positive advantage signal—encouraging only actions with positive advantages as they alone trigger the clipping upper bound—to **filter out unreasonable actions**, thereby concentrating the increase on *informative entropy* rather than spurious noise.

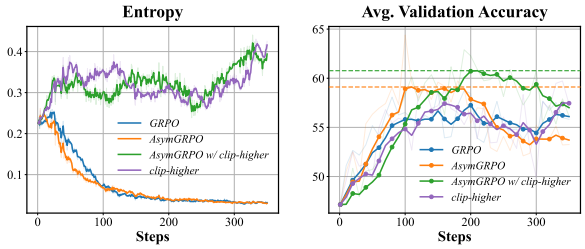

Figure 4: **Entropy Dynamics** and **Validation Accuracy**

4. Synergistic gains with AsymGRPO and Clip-higher. AsymGRPO and Clip-higher operate through orthogonal mechanisms and can be effectively combined. AsymGRPO w/ Clip-higher achieves a remarkable average accuracy of 60.32%. Analysis of the training dynamics (Fig. 4) reveals that compared to GRPO w/ Clip-higher, the combined method maintains similar entropy levels throughout the training process. This sustained uncertainty translates into improved exploration and significantly better generalized performance: AsymGRPO w/ Clip-higher outperforms GRPO w/ Clip-higher (58.75%) by 1.57%. This suggests that AsymGRPO serves as a robust backbone, effectively refining the learning signal while Clip-higher provides a complementary exploration mechanism, allowing the model to leverage higher entropy for better optimization without collapsing.

5 Conclusion

This work addresses the critical limitation of restricted exploration in RLVR. By conceptually decomposing policy entropy into *informative* and *spurious* forms, we identify that group-relative estimation functions as an implicit *entropy refinement* mechanism—sustaining useful diversity while suppressing noise. Building on this, we propose **AsymGRPO**, a parametric framework that explicitly decouples these modulation effects to optimize the exploration-exploitation trade-off. Experiments confirm its superior performance and synergistic potential with existing entropy-based regularizers. We thus advocate a paradigm shift from indiscriminate entropy maximization toward targeted refinement strategies to better guide complex reasoning.

Limitations

Our work establishes a novel framework for understanding and manipulating entropy dynamics in RLVR. Building on these findings, we summarize several limitations to guide future research:

- **Granularity of Entropy Modulation:** While utilizing group accuracy as a proxy effectively distinguishes entropy types for reweighting, future research could design more fine-grained measurable metrics (e.g., rollout-level, token-level) to identify the specific optimization elements driving different types of entropy dynamics, and achieve more precise, targeted entropy refinement.

- **Hyperparameter Optimization:** AsymGRPO relies on two decoupled hyperparameters (β_{pos} and β_{neg}) to achieve modulation flexibility. Currently, these coefficients remain static throughout the training process and require manual tuning. Future investigations could explore heuristic optimal correlations between these parameters to reduce the search cost. Additionally, developing adaptive scheduling mechanisms that dynamically adjust the modulation intensity across different training stages—rather than using fixed values—represents a promising direction to further optimize the trade-off between exploration and exploitation.

Ethical Considerations

This research focuses exclusively on computational methodologies for model reasoning and involves no human subjects, animal testing, or sensitive data. Consequently, we anticipate no ethical risks or conflicts of interest. We adhere to the highest standards of scientific integrity to ensure the validity and reliability of our findings.

References

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

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, and 1 others. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*.

Wenlong Deng, Yi Ren, Muchen Li, Danica J Sutherland, Xiaoxiao Li, and Christos Thrampoulidis. 2025. On the effect of negative gradient in group relative deep reinforcement optimization. *arXiv preprint arXiv:2505.18830*.

Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, Zhi Jin, Fei Huang, Yongbin Li, and Ge Li. 2025. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *Preprint*, arXiv:2508.00222.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Jonas Gehring, Kunhao Zheng, Jade Copet, Vegard Mella, Quentin Carbonneaux, Taco Cohen, and Gabriel Synnaeve. 2024. Rlef: Grounding code llms in execution feedback with reinforcement learning. *arXiv preprint arXiv:2410.02089*.

John C Gittins. 1979. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(2):148–164.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr.

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.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, and 1 others. 2025. Skywork open reasoner 1 technical report. *arXiv preprint arXiv:2505.22312*.

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, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*.

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*.

Yuxian Jiang, Yafu Li, Guanxu Chen, Dongrui Liu, Yu Cheng, and Jing Shao. 2025. Rethinking entropy regularization in large reasoning models. *arXiv preprint arXiv:2509.25133*.

Sham M Kakade. 2001. A natural policy gradient. *Advances in neural information processing systems*, 14.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

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*.

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Lixuan He, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, and 1 others. 2025. Towards a unified view of large language model post-training. *arXiv preprint arXiv:2509.04419*.

Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37:134387–134429.

Youssef Mroueh. 2025. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*.

Phuc Minh Nguyen, Chinh D La, Duy MH Nguyen, Nitesh V Chawla, Binh T Nguyen, and Khoa D Doan. 2025. The reasoning boundary paradox: How reinforcement learning constrains language models. *arXiv preprint arXiv:2510.02230*.

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

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. 2024. Unintentional unalignment: Likelihood displacement in direct preference optimization. *arXiv preprint arXiv:2410.08847*.

Yi Ren and Danica J Sutherland. 2024. Learning dynamics of llm finetuning. *arXiv preprint arXiv:2407.10490*.

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

Amrith Setlur, Saurabh Garg, Xinyang Geng, Naman Garg, Virginia Smith, and Aviral Kumar. 2024a. Rl on incorrect synthetic data scales the efficiency of llm math reasoning by eight-fold. *Advances in Neural Information Processing Systems*, 37:43000–43031.

Amrith Setlur, Chirag Nagpal, Adam Fisch, Xinyang Geng, Jacob Eisenstein, Rishabh Agarwal, Alekh Agarwal, Jonathan Berant, and Aviral Kumar. 2024b. Rewarding progress: Scaling automated process verifiers for llm reasoning. *arXiv preprint arXiv:2410.08146*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Han Shen. 2025. On entropy control in llm-rl algorithms. *arXiv preprint arXiv:2509.03493*.

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.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

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*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, and 1 others. 2025c. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, and 1 others. 2025. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Yiwei Wang, Xiaodan Liang, and Jing Tang. 2025b. Depth-breadth synergy in rlvr: Unlocking llm reasoning gains with adaptive exploration. *arXiv preprint arXiv:2508.13755*.

Xinhao Yao, Lu Yu, Xiaolin Hu, Fengwei Teng, Qing Cui, Jun Zhou, and Yong Liu. 2025. The debate on rlvr reasoning capability boundary: Shrinkage, expansion, or both? a two-stage dynamic view. *arXiv preprint arXiv:2510.04028*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, and 1 others. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

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*.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, and 1 others. 2025. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*.

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 Related Work

A.1 Reinforcement learning for LLMs

Recently, post-training research has increasingly focused on reinforcing Large Language Models (LLMs) in complex domains such as mathematics and programming using outcome-level verifiable rewards (Jaech et al., 2024; Guo et al., 2025; Team et al., 2025). This paradigm, often termed RLVR, is designed to incentivize extended Chain-of-Thought (CoT) reasoning (Wei et al., 2022), thereby enabling models to solve highly complex problems through scaled test-time computation (Wang et al., 2022). Notably, DeepSeek-R1 (Guo et al., 2025) demonstrated that reinforcement learning can effectively scale reasoning capabilities, and further revealed the spontaneous emergence of advanced behaviors such as self-reflection and branching during RLVR training. In practice, the prevailing approach is to optimize PPO-style policy-gradient surrogate objectives (Schulman et al., 2017) while leveraging a range of value-free advantage estimation methods to simplify reward-baseline computation, such as GRPO (Shao et al., 2024), which exploits group statistics, and REINFORCE++ (Hu et al., 2025), which incorporates global advantage normalization for stabilized updates. Despite these advances, RLVR still faces substantial challenges in exploration (Cui et al., 2025; Yu et al., 2025; Yue et al., 2025): insufficient exploration often manifests as entropy collapse and premature performance saturation, ultimately limiting its ability to unlock more robust and generalizable reasoning.

A.2 Exploration in RLVR

Effective exploration presents a unique challenge in RLVR compared to traditional RL settings (Xie et al., 2025; Yang et al., 2025b). While standard entropy regularization under the maximum-entropy RL view is often sufficient to maintain stochasticity and encourage exploration in conventional RL benchmarks (Haarnoja et al., 2018; Schulman et al., 2017), it faces difficulties in the vast vocabulary and long-horizon generation of LLM policies (Shen, 2025; Jiang et al., 2025). To address this, recent work has pursued two primary directions. One line of research focuses on maintaining policy entropy at a global level, enforcing target entropy constraints to prevent premature convergence (Yu et al., 2025; Cui et al., 2025). A second perspective investigates the non-uniform value of tokens, finding that RLVR gains are driven primarily by specific "forking" tokens—critical decision points in reasoning. Consequently, methods in this area employ token pruning (Wang et al., 2025b) or advantage shaping (Cheng et al., 2025; Wang et al., 2025a) to concentrate exploration credits specifically on these high-impact moments. Most relevantly, concurrent works (Jiang et al., 2025; Shen, 2025) introduce selective regularization. By limiting entropy maximization to the top-p nucleus or adapting it based on confidence, these methods attempt to filter out noise. This aligns with our objective: to amplify *informative entropy* while suppressing spurious uncertainty.

B Details on Covariance Estimation

The theoretical analysis of entropy evolution presented in Section 2.2 relies on a simplified approximation within the RL bandit setting (Gittins, 1979), where the prompt x is regarded as the state and the complete response y as the action.

During training, we calculate the group-wise covariance for each prompt, and average across a batch of prompts. Following Cui et al., 2025, we normalize the log-probability by the length of the response to mitigate the confounding effect of varying sequence lengths. We define the length-normalized log-probability, denoted as $\log \bar{\pi}_\theta(y_i \mid x)$, as:

$$\log \bar{\pi}_\theta(y_i \mid x) \triangleq \frac{1}{|y_i|} \sum_{t=1}^{|y_i|} \log \pi_\theta(y_{i,t} \mid x, y_{i,<t}). \tag{12}$$

For a specific group of G rollouts $\{y_i\}_{i=1}^{G}$ generated from prompt x, we estimate the covariance between the policy's confidence and the advantage signal $A(x, y_i)$ as:

$$\text{Cov}_{\text{group}} = \frac{1}{G} \sum_{i=1}^{G} \left(\log \bar{\pi}_\theta(y_i \mid x) - \mu_\pi\right) \left(A(x, y_i) - \mu_A\right), \tag{13}$$

where μ_π and μ_A denote the mean length-normalized log-probability and mean advantage within the group, respectively.

To analyze the relationship between optimization dynamics and problem difficulty, we aggregate these covariance estimates based on the group accuracy p. The reported metric $\text{Cov}(p)$ is the average covariance over the set of all groups $\mathcal{D}_p$ with accu-

racy p:

$$\mathrm{Cov}(p) = \frac{1}{|\mathcal{D}_p|} \sum_{g \in \mathcal{D}_p} \mathrm{Cov}^{(g)}_{\text{group}}. \tag{14}$$

In our analysis, we compute this metric using data exclusively from the first 40 training steps. This restriction is necessary because policy entropy tends to rapidly decrease and then stabilize in the initial training phase; focusing on the early steps allows us to capture optimization signals more obviously before the policy distribution approaches a relatively deterministic state.

C Experimental Settings for Entropy Analysis (Section 3)

This section details the experimental setup used to examine the impact of entropy dynamics on reasoning performance. All RL experiments are implemented using the verl (Sheng et al., 2025) framework on a single node equipped with 4 × NVIDIA H100 GPUs.

We conduct ablation studies on two base models: Qwen2.5-Math-1.5B (Yang et al., 2024) and Qwen3-4B (Yang et al., 2025a). For Qwen3-4B, we specifically utilize its non-thinking mode for training. The models are trained on the MATH dataset (Hendrycks et al., 2021), which contains 7,500 problems spanning diverse mathematical areas and difficulty levels.

We employ the AdamW optimizer with a learning rate of 2×10^{-6} for both models. Following Yu et al., 2025, we apply token-level loss aggregation for all settings. For each query, the policy generates $G = 8$ rollouts. Regarding model-specific configurations, the Qwen2.5-Math-1.5B experiments use a global batch size of 512, a mini-batch size of 128, and a maximum response length of 2,560 tokens. Conversely, the Qwen3-4B experiments utilize a global batch size of 128, a mini-batch size of 64, and a maximum response length of 4,096 tokens.

To monitor performance, we report the Avg. Val Acc, calculated as the mean accuracy across five mathematical reasoning benchmarks: AIME 2024, AIME 2025, MATH-500 (Lightman et al., 2023), AMC 2023 and OlympiadBench (He et al., 2024). Validation is performed every 10 training steps and the temperature is set to 0 to ensure the fast and reliable evaluation of model capabilities. To clearly visualize training trends, we apply Exponential Moving Average (EMA) smoothing with a factor of 0.7 to all validation accuracy curves.

D Implementation Details for Flipped Advantage Curves

To investigate the necessity of the proposed reweighting strategy, we construct flipped versions of the advantage curves to **reverse the original reweighting trends**. Mathematically, this is achieved by reflecting the advantage function around $p = 0.5$, such that $\tilde{A}(p) = A(1-p)$.

However, this reflection introduces numerical singularities at the boundaries. To handle these cases for a group size of G, we employ a **linear extension strategy** that extrapolates the trend from the penultimate feasible data points.

Positive Advantage Boundary ($p \to 1$). For positive samples (visualized as the **EntDecrease** curve in Fig. 1(a)), the flipped function is valid up to $p = \frac{G-1}{G}$. We replace the curve segment on the interval $[\frac{G-1}{G}, 1]$ with a linear function connecting the last valid point to an extrapolated boundary value. Specifically, we define the boundary value $V^{\text{end}}_{\text{pos}}$ at $p = 1$ by replicating the advantage increment from the previous step:

$$V^{\text{end}}_{\text{pos}} \triangleq \tilde{A}_{\text{pos}}\left(\tfrac{G-1}{G}\right) + \left[\tilde{A}_{\text{pos}}\left(\tfrac{G-1}{G}\right) - \tilde{A}_{\text{pos}}\left(\tfrac{G-2}{G}\right)\right]. \tag{15}$$

Negative Advantage Boundary ($p \to 0$). Similarly, for negative samples (visualized as the **EntIncrease** curve in Fig. 1(b)), the flipped function implies a singularity at $p = 0$. We linearly extend the curve on the interval $[0, \frac{1}{G}]$ based on the slope between $p = \frac{2}{G}$ and $p = \frac{1}{G}$. The boundary value $V^{\text{end}}_{\text{neg}}$ at $p = 0$ is derived as:

$$V^{\text{end}}_{\text{neg}} \triangleq \tilde{A}_{\text{neg}}\left(\tfrac{1}{G}\right) - \left[\tilde{A}_{\text{neg}}\left(\tfrac{2}{G}\right) - \tilde{A}_{\text{neg}}\left(\tfrac{1}{G}\right)\right]. \tag{16}$$

Summary of Piecewise Formulation. Combining the reflected core and the boundary extensions, the final flipped advantage functions are defined as:

$$\tilde{A}_{\text{pos}}(p) = \begin{cases} A^{(\beta)}_{\text{pos}}(1-p) & \text{if } 0 < p \le \frac{G-1}{G}, \\ \text{Linear}\left(p; \frac{G-1}{G}, 1\right) & \text{if } \frac{G-1}{G} < p \le 1, \end{cases} \tag{17}$$

$$\tilde{A}_{\text{neg}}(p) = \begin{cases} \text{Linear}\left(p; 0, \frac{1}{G}\right) & \text{if } 0 \le p < \frac{1}{G}, \\ A^{(\beta)}_{\text{neg}}(1-p) & \text{if } \frac{1}{G} \le p < 1, \end{cases} \tag{18}$$

where $\text{Linear}(p; a, b)$ represents the linear interpolation function connecting the derived boundary values at the interval endpoints a and b.

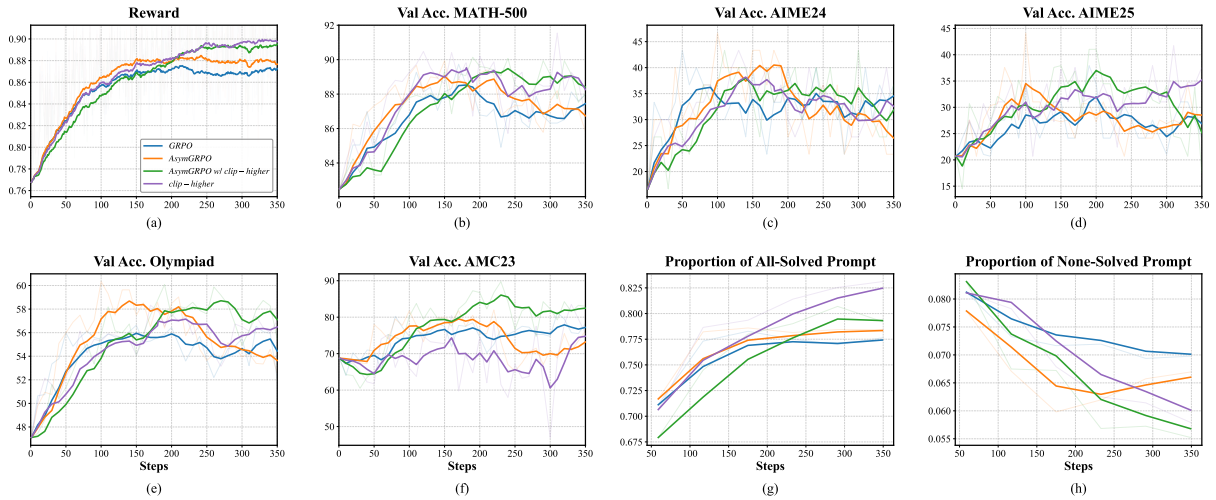

Figure 5: **Extended Training Dynamics and Performance Metrics.** (a) Evolution of the Training Reward. (b)–(f) Validation accuracy trajectories on individual mathematical reasoning benchmarks (MATH-500, AIME24, AIME25, AMC23, and Olympiad). (g) The proportion of prompts yielding exclusively correct responses. (h) The proportion of prompts yielding exclusively incorrect responses.

E Experimental Settings and Hyperparameter Choices for Main Results (Section 4)

The experimental configuration for our main results largely aligns with the setup described in Appendix C, utilizing the verl framework on a node equipped with 4 × NVIDIA H100 GPUs. In this section, we focus exclusively on training the Qwen3-4B model. MATH-500 serves as the validation set: for each run, we select the checkpoint achieving the highest validation accuracy and evaluate it on all mathematical reasoning benchmarks.

We compare our proposed method against a comprehensive set of baselines, including standard GRPO (Guo et al., 2025), GRPO with Entropy Regularization, GRPO with Clip-higher (Yu et al., 2025), GRPO with Entropy Advantage (Cheng et al., 2025), Dr.GRPO (Liu et al., 2025), and Pass@K Training (Chen et al., 2025). Regarding specific hyperparameters for the baseline variants, we set the coefficient for Entropy Regularization to 0.001, the upper clipping threshold for Clip-higher to $\epsilon_{\text{high}} = 0.28$, $K = 5$ for Pass@K Training, and the scaling factor for Entropy Advantage to $\kappa = 2, \alpha = 0.4$. For our proposed **AsymGRPO** configurations, we utilize $\beta_{\text{pos}} = 0.9$ and $\beta_{\text{neg}} = 0.4$ for the standard setting. In the symmetric ablation ($\beta_{\text{pos}} = \beta_{\text{neg}}$), both coefficients are set to 0.7. When integrating with Clip-higher, we decrease β_{neg} to 0.3 while maintaining $\beta_{\text{pos}} = 0.9$ and using $\epsilon_{\text{high}} = 0.28$.

F Supplementary Experimental Results

Figure 5 illustrates the additional training dynamics, providing a detailed view of the training reward, per-dataset validation accuracy, and the evolution of prompt response distributions (perfect vs. zero rates) throughout the training process.

G Information About Use of AI Assistants

The use of AI assistants in this work was limited to grammatical polishing and the correction of typographical errors. The original draft was entirely written by the authors, and all AI-suggested modifications were rigorously verified by the authors to ensure accuracy and intent.

H Licenses

Qwen3 (Yang et al., 2025a) and Qwen2.5-Math (Yang et al., 2024) are distributed under the Apache License 2.0. The MATH dataset (Hendrycks et al., 2021) and its subset MATH-500 (Lightman et al., 2023) are released under the MIT license. The OlympiadBench dataset (He et al., 2024) is released under the Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) license. The AIME and AMC datasets are utilized strictly for academic research and evaluation purposes. All resources are used in accordance with their respective licensing terms.