F-GRPO: DON'T LET YOUR POLICY LEARN THE OBVIOUS AND FORGET THE RARE

Daniil Plyusov[1,2*] **Alexey Gorbatovski**[1*] **Boris Shaposhnikov**[1]
Viacheslav Sinii[1] **Alexey Malakhov**[1] **Daniil Gavrilov**[1]

[1]T-Tech
[2]Saint Petersburg Electrotechnical University "LETI"

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) is commonly based on group sampling to estimate advantages and stabilize policy updates. In practice, large group sizes are not feasible due to computational limits, which biases learning toward trajectories that are already likely. Smaller groups often miss rare-correct trajectories while still containing mixed rewards, concentrating probability on common solutions. We derive the probability that updates miss rare-correct modes as a function of group size, showing non-monotonic behavior, and characterize how updates redistribute mass within the correct set, revealing that unsampled-correct mass can shrink even as total correct mass grows. Motivated by this analysis, we propose a difficulty-aware advantage scaling coefficient, inspired by Focal loss, that down-weights updates on high-success prompts. The lightweight modification can be directly integrated into any group-relative RLVR algorithm such as GRPO, DAPO, and CISPO. On Qwen2.5-7B across in-domain and out-of-domain benchmarks, our method improves pass@256 from 64.1 $\rightarrow$ 70.3 (GRPO), 69.3 $\rightarrow$ 72.5 (DAPO), and 73.2 $\rightarrow$ 76.8 (CISPO), while preserving or improving pass@1, without increasing group size or computational cost.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has become a standard paradigm for post-training large language models (LLMs), enabling strong gains on reasoning-intensive tasks without reliance on human preference data (Zhang et al., 2025). By leveraging automatically checkable reward signals, RLVR has driven state-of-the-art performance in mathematical reasoning (Li et al., 2024), code generation (Jimenez et al., 2023), and general problem solving (Chollet et al., 2025), and is now widely adopted in large-scale post-training (Guo et al., 2025; Yang et al., 2025; Team et al., 2025; Shao et al., 2024).

Despite these successes, a growing body of work suggests that RLVR does not primarily introduce new knowledge, but instead sharpens the output distribution toward solutions already accessible to the base model (Yue et al., 2025; Ni et al., 2025; Wu et al., 2025a; Dang et al., 2025). Empirical evidence based on pass@k (Chen et al., 2021) indicates that RLVR-trained models may underperform their base counterparts at sufficiently large sampling budgets, consistent with a narrowing of solution diversity (Matsutani et al., 2025). At the same time, other studies argue that prolonged or carefully scaled RL can expand the effective reasoning boundary (Liu et al., 2025b; Yuan et al., 2025), leaving the role of RLVR an open question.

Most modern RLVR systems rely on group-relative methods such as GRPO (Shao et al., 2024) and its variants (Yu et al., 2025; Chen et al., 2025a; Liu et al., 2025c), which compute advantages from multiple rollouts per prompt. The group size thus becomes a critical design choice, yet existing work provides conflicting guidance: Wu et al. (2025b) show that two rollouts suffice and connect GRPO to DPO (Rafailov et al., 2023), while Hu et al. (2025) advocate scaling rollouts to broaden exploration. Since group size directly controls which trajectories receive learning signal,

*Equal contribution. Correspondence to: alexey.gorbatovski@gmail.com

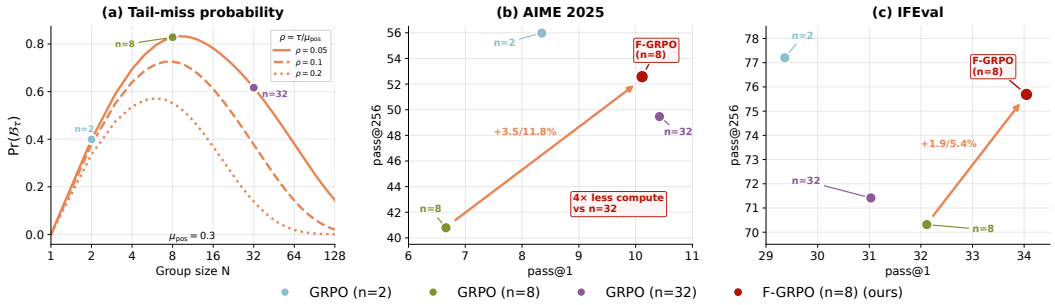

Figure 1: (a) Probability that a training update is *active* (mixed rewards in batch) yet *misses* rare-correct solutions, as a function of group size N. This probability peaks at intermediate N: small groups rarely produce learning signal, large groups cover rare modes, but moderate groups combine active updates with poor coverage. (b,c) Empirical consequences on AIME 2025 (math) and IFEval (OOD): GRPO at N=8 improves pass@1 over N=2 but degrades pass@256, consistent with the sharpening regime. F-GRPO at N=8 recovers pass@256 while maintaining pass@1, using $4\times$ less compute than N=32.

understanding its interaction with sharpening is essential. This raises a fundamental question: *how does group size affect the optimization dynamics of group-relative RLVR with binary rewards, and can we mitigate sharpening without scaling computational cost?*

In this paper, we analyze the sampling dynamics of group-relative RLVR and propose F-GRPO, a lightweight modification that addresses sharpening at practical group sizes. Our contributions are as follows:

- We derive a closed-form tail-miss probability for active RLVR updates missing rare-correct modes, revealing non-monotonic dependence on group size: small groups preserve diversity through inactivity, large groups through coverage, while intermediate groups maximize sharpening risk.
- Building on Hu et al. (2025), we show that unsampled-correct mass can decrease even when total correct mass increases.
- We propose F-GRPO, a difficulty-aware advantage scaling for any group-relative objective (GRPO, DAPO, CISPO), demonstrating consistent pass@256 improvements on both in-domain math and out-of-domain reasoning benchmarks while preserving pass@1 across three model families without additional cost.

Figure 1 illustrates the core finding: tail-miss probability peaks at intermediate group sizes, and F-GRPO at N=8 matches or exceeds GRPO at N=32, achieving higher pass@256 (52.6 vs. 49.5 on AIME 2025; 75.7 vs. 71.4 on IFEval) and improved OOD pass@1 (34.0 vs. 31.0), while using $4\times$ fewer rollouts.

2 Preliminaries

2.1 Reinforcement Learning with Verifiable Rewards

We consider reinforcement learning with verifiable rewards (RLVR) for language model reasoning. Given a prompt x, the policy π_θ generates complete responses (trajectories). We sample a group of N i.i.d. rollouts $\{o_i\}_{i=1}^N \sim \pi_\theta(\cdot \mid x)$ and assign binary outcome rewards $R_i = R_w + (R_c - R_w)\,\mathbb{I}[o_i \text{ is correct}]$, where $R_c > R_w$ (typically $R_c = 1$, $R_w \in \{0, -1\}$). We work with outcome-level rewards: the reward depends only on final correctness.

For each prompt x, let Ω_x denote the space of complete rollouts and $\mathcal{C}(x) \subseteq \Omega_x$ the subset of correct rollouts. Define the success probability $\mu_{\text{pos}}(x) := \Pr_{o\sim\pi_\theta(\cdot|x)}[o \in \mathcal{C}(x)]$. For analysis, we consider a designated subset $\mathcal{C}_{\text{rare}}(x) \subseteq \mathcal{C}(x)$ of correct rollouts with mass $\tau(x) := \Pr_{o\sim\pi_\theta(\cdot|x)}[o \in \mathcal{C}_{\text{rare}}(x)]$. By construction $0 \le \tau(x) \le \mu_{\text{pos}}(x)$. We call $\mathcal{C}_{\text{rare}}(x)$ "rare-correct" when $\rho(x) := \tau(x)/\mu_{\text{pos}}(x)$ is small; this ratio can change as π_θ evolves.

2.2 GROUP-RELATIVE POLICY OPTIMIZATION

Group Relative Policy Optimization (GRPO) (Shao et al., 2024) eliminates the learned value function by computing advantages relative to the sampled group. For a prompt x with N rollouts $\{o_i\}_{i=1}^N$ and rewards $\{R_i\}_{i=1}^N$, the group-relative advantage is $\widehat{A}_i^{\mathrm{GRPO}} = (R_i - \bar{R})/(\sigma_R + \epsilon)$, where $\bar{R} = \frac{1}{N}\sum_{j=1}^N R_j$ and $\sigma_R = \mathrm{std}(\{R_j\}_{j=1}^N)$.

GRPO optimizes a clipped surrogate objective. Let $o_i = (y_{i,1}, \ldots, y_{i,T_i})$ denote the token sequence for rollout i, with importance ratio $r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(y_{i,t}|x, y_{i,<t})}$. The GRPO objective is

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_x \left[\frac{1}{N} \sum_{i=1}^{N} \frac{1}{T_i} \sum_{t=1}^{T_i} L_{i,t}^{\mathrm{clip}} \; - \; \beta\, \mathbb{D}_{\mathrm{KL}}\left(\pi_\theta \,\|\, \pi_{\mathrm{ref}}\right) \right], \tag{1}$$

where $L_{i,t}^{\mathrm{clip}} = \min\left(r_{i,t}\widehat{A}_i,\ \mathrm{clip}(r_{i,t}, 1{-}\varepsilon, 1{+}\varepsilon)\widehat{A}_i\right)$. We set $\beta = 0$ following DAPO (Yu et al., 2025). DAPO modifies this with asymmetric clipping bounds $\mathrm{clip}(r_{i,t}, 1{-}\varepsilon_{\mathrm{low}}, 1{+}\varepsilon_{\mathrm{high}})$ where $\varepsilon_{\mathrm{high}} > \varepsilon_{\mathrm{low}}$, relaxing the upper bound for low-probability actions.

CISPO (Chen et al., 2025a) clips the importance weights directly rather than the surrogate product. Define the clipped weight $\widehat{r}_{i,t} = \mathrm{clip}\big(r_{i,t}, 1{-}\varepsilon_{\mathrm{low}}^{\mathrm{IS}}, 1{+}\varepsilon_{\mathrm{high}}^{\mathrm{IS}}\big)$ and optimizes a REINFORCE-style objective

$$\mathcal{L}_{\mathrm{CISPO}}(\theta) = \mathbb{E}_{i,t}\left[\mathrm{sg}(\widehat{r}_{i,t})\, \widehat{A}_i^{\mathrm{GRPO}}\, \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right], \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient.

A key property of group-relative advantages is that when all sampled rewards are identical ($\sigma_R = 0$), we have $\widehat{A}_i^{\mathrm{GRPO}} = 0$ for all i, which yields zero learning signal. This occurs when all rollouts are correct or all are incorrect.

2.3 CATEGORICAL POLICY FRAMEWORK

To analyze how RLVR updates redistribute probability mass, we adopt the categorical policy framework of (Hu et al., 2025). Consider $p = \mathrm{softmax}(z)$ over a finite action space $\mathcal{A}$, partitioned into correct actions $\mathcal{P}$ and incorrect $\mathcal{N} = \mathcal{A} \setminus \mathcal{P}$. Define the total correct and incorrect masses

$$Q_{\mathrm{pos}} := \sum_{i \in \mathcal{P}} p_i, \quad Q_{\mathrm{neg}} := 1 - Q_{\mathrm{pos}}. \tag{3}$$

Draw N i.i.d. samples from p. Let $A \subseteq \mathcal{P}$ and $B \subseteq \mathcal{N}$ denote sampled correct and incorrect actions, $U = \mathcal{A} \setminus (A \cup B)$ the unsampled actions. Define the sampled masses and concentration measures $P_{\mathrm{pos}} := \sum_{i \in A} p_i$, $P_{\mathrm{neg}} := \sum_{i \in B} p_i$, $A_2 := \sum_{i \in A} p_i^2$, and $B_2 := \sum_{i \in B} p_i^2$.

For the unsampled set, define $U_{\mathrm{pos},2} := \sum_{i \in U \cap \mathcal{P}} p_i^2$ and $U_{\mathrm{neg},2} := \sum_{i \in U \cap \mathcal{N}} p_i^2$. Assign rewards as in Section 2.1 for sampled actions, with $R_i = 0$ for unsampled. The batch baseline is $S_R := R_c P_{\mathrm{pos}} + R_w P_{\mathrm{neg}}$.

We analyze TRPO-style linear surrogate updates and their unbiased Monte Carlo estimates. Under standard regularity conditions, expectation and differentiation may be interchanged (Asmussen & Glynn, 2007; Hu et al., 2025). Differentiating the sample surrogate with respect to the logits z_j (using $\partial p_i / \partial z_j = p_i(\delta_{ij} - p_j)$) yields the one-step logit update

$$\Delta z_i = \frac{\eta}{N} p_i (R_i - S_R), \tag{4}$$

where η is the learning rate. For unsampled actions ($i \in U$), this reduces to $\Delta z_i = -\frac{\eta}{N} S_R p_i$.

From this update rule, Hu et al. (2025) derive the one-step change in total correct mass:

$$\Delta Q_{\mathrm{pos}} = \frac{\eta}{N}\Big[(R_c - S_R) Q_{\mathrm{neg}} A_2 + (S_R - R_w) Q_{\mathrm{pos}} B_2 + S_R (Q_{\mathrm{pos}} U_{\mathrm{neg},2} - Q_{\mathrm{neg}} U_{\mathrm{pos},2})\Big]. \tag{5}$$

The first two terms are always non-negative: promoting sampled correct actions and demoting sampled incorrect actions both transfer mass to the correct pool. The third term, the unsampled coupling, can be positive or negative depending on S_R and the relative concentration of unsampled

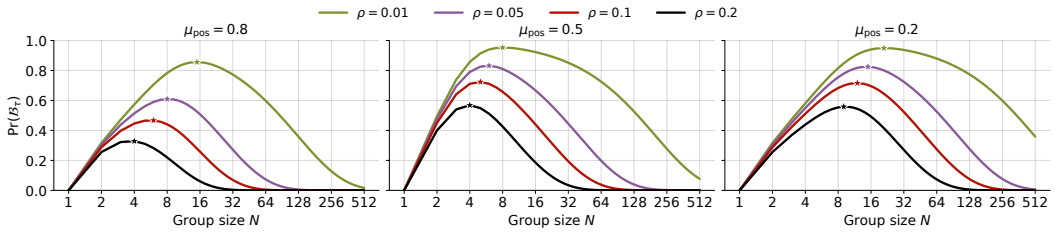

Figure 2: Tail-miss probability $\Pr(\mathcal{B}_\tau)$ from Lemma 3.1 versus group size N. Each panel fixes $\mu_{\text{pos}} \in \{0.8, 0.5, 0.2\}$; curves vary $\rho = \tau/\mu_{\text{pos}}$, the fraction of correct mass in the rare-correct region. Stars mark peaks. For all parameter combinations, $\Pr(\mathcal{B}_\tau)$ peaks at intermediate N: small N yields low activity, large N yields good coverage, but moderate N combines active groups with poor coverage of rare modes. Smaller ρ shifts the peak rightward and upward.

masses. As the unsampled second moments decay with N, increasing rollout size drives this coupling toward zero.

This categorical framework directly models token-level update dynamics. For trajectory-level RLVR, we maintain separate notation to avoid conflation: $\mu_{\text{pos}}(x)$ denotes the per-prompt success probability (Section 2.1, while Q_{pos} refers to correct mass in the categorical setting equation 3.

3 Theoretical Analysis

Recent work offers seemingly conflicting guidance on group size in RLVR: very small groups ($N = 2$) can match larger ones efficiently (Wu et al., 2025b), moderate sizes improve pass@1 while sharpening the distribution (He et al., 2025), and large groups stabilize learning (Hu et al., 2025). We develop a theoretical framework that reconciles these findings.

3.1 Tail-miss probability and the group size trade-off

We begin with a sampling analysis at the trajectory level. Consider N i.i.d. rollouts from $\pi_\theta(\cdot|x)$ with success probability $\mu_{\text{pos}}(x)$ and rare-correct mass $\tau(x)$ (Section 2.1), where $0 < \tau(x) < \mu_{\text{pos}}(x)$.

Let X denote the number of correct rollouts among the N samples. For group-relative methods such as GRPO, the learning signal vanishes when all sampled rewards are identical, i.e., $X \in \{0, N\}$. Define the *active event* $\mathcal{A}_N := \{0 < X < N\}$, with probability $\Pr(\mathcal{A}_N) = 1 - \mu_{\text{pos}}(x)^N - (1 - \mu_{\text{pos}}(x))^N$.

Let $Y_i = \mathbb{I}[\text{rollout } i \in \mathcal{C}_{\text{rare}}(x)]$, so $\Pr(Y_i = 1) = \tau(x)$. We are interested in the event that the update is active yet the rare-correct region receives no samples: $\mathcal{B}_\tau := \mathcal{A}_N \cap \left\{ \sum_{i=1}^N Y_i = 0 \right\}$.

Lemma 3.1. *For any $N \geq 1$, writing $\mu_{\text{pos}} = \mu_{\text{pos}}(x)$ and $\tau = \tau(x)$ for brevity,*

$$\Pr(\mathcal{B}_\tau) = (1 - \tau)^N - (\mu_{\text{pos}} - \tau)^N - (1 - \mu_{\text{pos}})^N. \tag{6}$$

The proof partitions rollouts into three disjoint regions and applies inclusion-exclusion (Appendix A).

Equation equation 6 reveals a non-monotonic dependence on N. Two competing effects determine $\Pr(\mathcal{B}_\tau)$: the coverage factor $(1 - \tau)^N$ decreases with N, improving the chance of sampling rare-correct modes, while activity $\Pr(\mathcal{A}_N)$ increases from near zero toward one. Their interaction produces three distinct regimes (Figures 1(a) and 2):

Small N (e.g., $N = 2$): Activity $\Pr(\mathcal{A}_N)$ is low, most groups are homogeneous, yielding zero learning signal. The policy changes slowly from the base model, preserving output diversity. This regime favors pass@k for large k but limits pass@1 improvement, consistent with the finding that minimal group sizes maintain diversity at the cost of sample efficiency (Wu et al., 2025b; Dang et al., 2025).

Intermediate N: $\Pr(\mathcal{B}_\tau)$ peaks; updates are frequently active yet often miss rare-correct modes. He et al. (2025) observe this regime at $N = 32$: pass@1 improves while pass@k for large k degrades, indicating distribution sharpening.

Large N: Coverage improves as $(1-\tau)^N \to 0$ and unsampled mass diminishes. This is the regime analyzed by (Hu et al., 2025), where scaling N stabilizes learning and can improve both metrics.

This framework reconciles the seemingly contradictory recommendations: small N preserves diversity through inactivity; large N through coverage; intermediate N, most common in practice due to computational constraints, is where sharpening is most likely. Figure 1(b,c) illustrates this empirically (in-domain: AIME 2025; OOD: IFEval). At $N = 8$, pass@1 improves relative to $N = 2$ but pass@256 degrades, reflecting the sharpening trade-off. Increasing N to 32 improves both pass@1 and pass@256 compared to $N = 8$, consistent with Hu et al. (2025); in our setup $N = 32$ falls in the large-N regime, whereas for He et al. (2025) it was intermediate. This shift in regime boundaries, determined by μ_{pos}, τ, and their evolution during training, also explains the smaller degradation on OOD IFEval.

3.2 Unsampled-correct mass under finite sampling

The tail-miss analysis identifies *when* rare-correct modes are vulnerable (intermediate N where $\Pr(\mathcal{B}_\tau)$ peaks). We now use the categorical framework (Section 2.3) to characterize the *mechanism* by which their mass decreases.

While equation 5 shows that total correct mass Q_{pos} tends to increase with N, it does not reveal redistribution within the correct set. Define the unsampled-correct mass

$$Q_{\text{u,pos}} := \sum_{i \in U \cap \mathcal{P}} p_i = Q_{\text{pos}} - P_{\text{pos}}. \tag{7}$$

This quantity measures how much correct probability is "left behind" by sampling.

Proposition 3.2. *Under the one-step surrogate update equation 4,*

$$\Delta Q_{\text{u,pos}} = \frac{\eta}{N}\Big[\underbrace{-S_R\, U_{\text{pos},2}}_{\textit{direct drift}} - Q_{\text{u,pos}} \underbrace{\big((R_c - S_R)A_2 + (R_w - S_R)B_2 - S_R U_2\big)}_{\textit{normalization coupling}}\Big]. \tag{8}$$

The proof applies the subset-mass identity from Appendix C with $\mathcal{S} = U \cap \mathcal{P}$; see Appendix D for details.

Equation equation 8 shows that $\Delta Q_{\text{u,pos}}$ can be negative even when $\Delta Q_{\text{pos}} > 0$: RLVR can increase total correct mass while concentrating it onto sampled-correct actions at the expense of unsampled-correct ones. This complements Hu et al. (2025), who showed that reward-positive batches ($S_R > 0$) push unsampled logits downward. Our formula makes explicit how this affects redistribution *within* the correct set.

The mechanism operates through two terms. The *direct drift* $-S_R U_{\text{pos},2}$ pushes unsampled-correct mass downward when $S_R > 0$, with magnitude scaling with the concentration $U_{\text{pos},2}$. The *normalization coupling* (analyzed in detail in Appendix E) captures how probability gains by sampled-correct actions draw mass away from unsampled-correct ones through softmax normalization. In reward-positive batches, both terms contribute negatively.

As Hu et al. (2025) observe, scaling N suppresses $U_{\text{pos},2}$ and ensures $\Delta Q_{\text{pos}} \geq 0$ with the *direct drift* term tending to zero. However, practical constraints limit how far N can be scaled: computational cost grows linearly with N, and improving pass@1 requires active groups (ruling out very small N where most groups are homogeneous). This places typical RLVR training in the intermediate-N regime identified in Section 3.1, where $\Pr(\mathcal{B}_\tau)$ peaks.

4 F-GRPO: Focal weighting for Group-Relative Policy Optimization

The categorical analysis in Section 3.2 identifies $S_R > 0$ as the condition driving concentration of correct mass. To operationalize this insight at the trajectory level, we need an observable per-prompt statistic that tracks the magnitude of the S_R-driven drift.

4.1 FOCAL WEIGHT

Define the empirical success rate for prompt x as

$$\widehat{\mu}_{\text{pos}}(x) := \frac{\bar{R}(x) - R_w}{R_c - R_w} = \frac{X}{N} \in [0, 1], \tag{9}$$

where X is the number of correct rollouts and $\bar{R}(x) = \frac{1}{N}\sum_{i=1}^{N} R_i$ is the group mean reward. This is an unbiased estimator of the true success probability: $\mathbb{E}[\widehat{\mu}_{\text{pos}}(x)] = \mu_{\text{pos}}(x)$.

The token-level categorical analysis in Section 3.2 identifies $S_R > 0$ as the regime driving concentration, but S_R depends on the (unobserved) policy mass of *distinct* sampled rollouts. At the trajectory level, we therefore use $\widehat{\mu}_{\text{pos}}(x) = X/N$ as an *observable proxy* for this regime: under i.i.d. rollout sampling, conditioning on $X = k$ implies the distinct sampled-correct mass $\mathbb{E}[P_{\text{pos}} \mid X = k]$ is non-decreasing in k and $\mathbb{E}[P_{\text{neg}} \mid X = k]$ is non-increasing in k, so for standard RLVR rewards $R_w \leq 0$, $\mathbb{E}[S_R \mid X = k]$ is non-decreasing in k. See Appendix B for a formal proof of this sampling monotonicity. In the categorical update, unsampled actions receive zero reward but are still affected by the baseline subtraction, yielding $\Delta z_i \propto -S_R p_i$ for $i \in U$ (Eq. equation 4); as a result, the downward drift on unsampled-correct mass is strongest in reward-positive batches where $S_R > 0$ (Section 3.2). We thus aim to reduce updates on prompts that are likely to fall into this regime, using $\widehat{\mu}_{\text{pos}}(x)$ as a per-prompt proxy signal.

Because $\mathbb{E}[S_R \mid X = k]$ is non-decreasing in k, higher $\widehat{\mu}_{\text{pos}}(x)$ marks prompts where concentration pressure is strongest.

Since high $\widehat{\mu}_{\text{pos}}(x)$ indicates easy/high-success prompts, we adopt a functional form inspired by Focal loss (Lin et al., 2017) to down-weight their updates, where the drift mechanism is most pronounced. Define the difficulty weight

$$g(x) := \big(1 - \widehat{\mu}_{\text{pos}}(x)\big)^{\gamma}, \quad \gamma \geq 0. \tag{10}$$

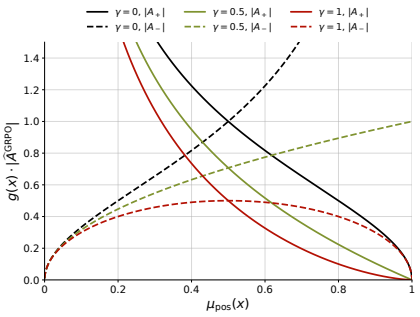

Figure 3: Scaled advantage magnitude $g(x) \cdot |\widehat{A}^{\text{GRPO}}|$ versus success probability $\mu_{\text{pos}}(x)$ for binary rewards. Solid lines: correct rollouts; dashed lines: incorrect rollouts.

When $\gamma = 0$, $g(x) = 1$ for all prompts, recovering standard GRPO. For $\gamma > 0$, prompts with high empirical success rate receive reduced weight: $g(x) \to 0$ as $\widehat{\mu}_{\text{pos}}(x) \to 1$.

4.2 INTEGRATION WITH GROUP-RELATIVE METHODS

We incorporate the difficulty weight by scaling the group-relative advantage:

$$\widehat{A}_i^{\text{F-GRPO}} := g(x) \cdot \widehat{A}_i^{\text{GRPO}}. \tag{11}$$

This modification applies to any method using group-relative advantages. While DAPO (Yu et al., 2025) and CISPO (Chen et al., 2025a) modify the clipping mechanism and importance weighting respectively, the concentration phenomenon we address arises from the sampling dynamics of group-relative advantage estimation, not from these algorithmic choices. The Focal weight $g(x)$ is thus orthogonal and can be applied independently. We denote the Focal-weighted variants as F-DAPO and F-CISPO. The modification is minimal: a single scalar $g(x) \in [0, 1]$ applied uniformly to all rollouts from the same prompt. No additional networks are required; γ is the only new hyperparameter.

Figure 3 visualizes the effect of Focal weighting. With binary rewards, the GRPO advantage magnitudes vary with $\mu_{\text{pos}}(x)$. The Focal weight $g(x) = (1 - \widehat{\mu}_{\text{pos}}(x))^{\gamma}$ scales these magnitudes, suppressing updates on high-success prompts. Analogous to Focal loss suppressing well-classified examples, this reduces gradient contribution from prompts where the concentration mechanism of Section 3.2 is most active.

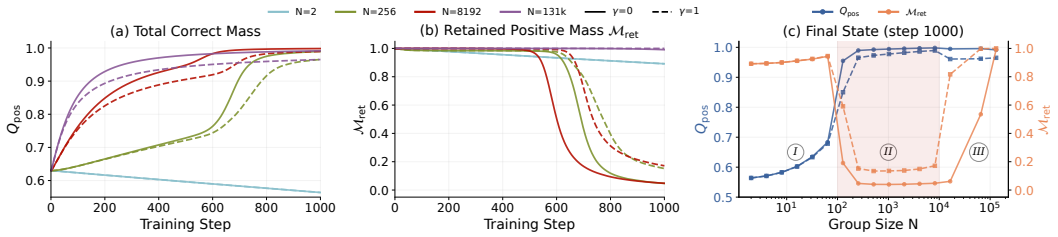

Figure 4: Categorical policy simulation following Hu et al. (2025) setup. **(a)** Total correct mass Q_{pos} vs. training step. **(b)** Retained positive mass $\mathcal{M}_{\text{ret}}$ vs. step. **(c)** Final metrics vs. group size N, with three regimes: **I** slow Q_{pos} growth, diversity preserved; **II** concentration zone (shaded), Q_{pos} grows but $\mathcal{M}_{\text{ret}}$ collapses; **III** both metrics high. Solid: γ=0; dashed: γ=1. N=131k maintains $\mathcal{M}_{\text{ret}}\approx 1$ throughout, consistent with $\Pr(\mathcal{B}_\tau) < 10^{-3}$ (Appendix J).

5 Experiments & Results

5.1 Empirical Validation via Categorical Simulation

To complement the simulation analysis of Hu et al. (2025), we conduct experiments under the same categorical policy framework (Section 2.3) with an additional focus on *which* correct actions retain probability mass. Following Hu et al. (2025), we simulate a softmax policy over 128,000 actions (10,000 correct) trained with group-relative updates; see Appendix J for details.

Beyond tracking total correct mass Q_{pos}, we track $\mathcal{M}_{\text{ret}}(t)$, the *retained positive mass*, measuring the fraction of initial correct-action probability that remains at or above its starting value (Appendix J Eq. 21). Values near 1 indicate diversity preservation; values near 0 indicate concentration onto a subset of solutions.

Figure 4 presents the results. Panel (a) confirms that Q_{pos} increases for all group sizes, consistent with Hu et al. (2025). However, panel (b) support that $\mathcal{M}_{\text{ret}}$ behaves non-monotonically: both small and large N preserve diversity, while intermediate values suffer severe concentration. This demonstrates that $\Delta Q_{\text{pos}} > 0$ does not guarantee preservation of unsampled correct actions. Panel (c) summarizes the final state across all group sizes, with three regimes labeled: **(I)** small N where Q_{pos} grows slowly but diversity is preserved; **(II)** the concentration zone (shaded) where Q_{pos} grows rapidly but $\mathcal{M}_{\text{ret}}$ collapses; and **(III)** large N where both metrics are high. Notably, N=131,072 maintains $\mathcal{M}_{\text{ret}} \approx 1$ throughout training, consistent with Lemma 3.1, which predicts $\Pr(\mathcal{B}_\tau) < 10^{-3}$ at this group size (see Appendix J). Dashed lines (γ=1) show improved $\mathcal{M}_{\text{ret}}$ retention, particularly in the concentration zone. In this single-tree setting, Focal weighting suppresses updates on high-success batches where concentration pressure peaks; with multiple prompts, it additionally reallocates gradient toward harder examples.

The specific boundaries of the concentration zone depend on the initial distribution and should not be interpreted as quantitative predictions for LLM training. The key insight is the qualitative pattern: intermediate group sizes can exhibit worse diversity than either extreme.

5.2 LLM Experimental Setup

Models & Datasets. We evaluate on Qwen2.5-7B (Yang et al., 2024a), Qwen2.5-1.5B-Math (Yang et al., 2024b), and Llama-3.2-3B-Instruct (Grattafiori et al., 2024), covering different model families and scales. All models are trained on DeepScaleR (Luo et al., 2025), a challenging dataset of competition-level mathematics problems.

Training. We implement our method using the verl framework (Sheng et al., 2024). Key hyperparameters: global batch size 256, mini-batch size 64, learning rate 1×10^{-6}, and 10 training epochs. The γ is selected from $\{0.5, 1.0\}$ based on average math pass@1 on the best checkpoint. Full training details are in Appendix H.

Evaluation. We report pass@1 and pass@256 to measure single-attempt accuracy and solution diversity. For in-domain evaluation, we use standard mathematical reasoning benchmarks: MATH500 (Hendrycks et al., 2021), AIME24/25 (Art of Problem Solving, 2024a), AMC23 (Art

Method	In-domain							Out-of-domain			
	Avg.	AIME24	AIME25	AMC	MATH500	Minerva	Olympiad	Avg. OOD	IFEval	SynLogic	GPQA
Qwen2.5-7B											
GRPO	37.3/64.1	15.0/37.7	6.7/40.8	52.9/87.3	75.8/92.8	**36.0**/60.2	**37.8**/65.8	17.1/55.9	32.1/70.3	7.9/51.3	11.3/46.2
F-GRPO	**38.6**/**70.3**	**15.9**/**46.2**	**10.1**/**52.6**	**56.2**/**96.3**	**76.2**/**95.1**	35.7/**60.3**	37.5/**71.6**	**19.2**/**63.3**	**34.0**/**75.7**	**8.7**/**57.0**	**15.0**/**57.3**
DAPO	39.4/69.3	16.8/49.8	**12.0**/45.6	53.3/91.9	78.6/95.2	**35.5**/61.2	40.5/71.8	15.7/58.4	24.1/67.1	7.5/53.3	**15.4**/54.9
F-DAPO	**40.5**/**72.5**	**20.9**/**53.4**	11.5/**52.9**	**55.9**/**93.7**	**79.1**/**96.6**	35.0/**62.9**	**40.9**/**75.6**	**17.9**/**63.6**	**30.8**/**71.1**	**7.9**/**62.4**	15.0/**57.4**
CISPO	39.5/73.2	14.6/45.9	9.7/59.8	**57.8**/96.1	78.7/97.0	**34.7**/63.3	41.5/76.9	14.9/59.0	24.2/67.9	8.0/53.6	12.6/55.5
F-CISPO	39.5/**76.8**	**14.8**/**59.7**	**13.0**/**64.6**	53.3/**97.1**	**79.0**/**97.8**	34.6/**64.3**	**42.4**/**77.5**	**18.1**/**65.9**	**30.7**/**70.6**	**8.2**/**60.0**	**15.4**/**67.1**
Qwen2.5-1.5B-Math											
GRPO	**36.7**/74.4	**13.8**/**61.1**	9.9/**58.0**	**53.1**/**96.2**	**75.4**/95.6	**31.9**/**61.1**	36.3/74.3	7.9/43.1	**12.2**/52.0	**4.9**/27.4	6.6/50.1
F-GRPO	36.3/**74.5**	13.0/60.7	**10.5**/57.9	51.6/95.9	74.7/**96.1**	31.0/61.0	**37.0**/**75.5**	**8.3**/**46.5**	11.4/**55.4**	4.8/**27.5**	**8.8**/**56.5**
DAPO	37.7/74.3	**16.5**/58.4	9.8/59.2	**54.5**/95.2	76.5/96.2	**32.6**/63.5	36.4/73.2	8.7/45.4	12.7/50.0	**5.0**/**26.9**	8.6/59.4
F-DAPO	**37.8**/**76.0**	16.1/**61.1**	**10.3**/**61.0**	54.4/**97.0**	**76.8**/**97.0**	32.2/**63.8**	**37.2**/**76.2**	**9.1**/**46.3**	**13.2**/**51.8**	4.9/26.4	**9.2**/**60.7**
CISPO	**38.9**/72.9	**16.8**/60.8	10.5/53.8	**58.6**/95.7	**77.3**/95.5	**32.6**/59.8	**37.6**/71.9	8.6/41.0	13.2/48.2	**5.2**/**26.3**	7.4/48.4
F-CISPO	37.4/**76.1**	14.5/**64.2**	**11.2**/**59.7**	53.8/**99.1**	76.8/**96.5**	31.7/**63.2**	36.4/**74.0**	**10.1**/**47.7**	**13.4**/**52.9**	4.9/26.1	**12.0**/**64.2**
Llama-3.2-3B-Instruct											
GRPO	23.0/59.9	10.7/40.7	0.7/21.5	**30.5**/88.2	**55.0**/90.6	**21.8**/59.0	19.4/59.3	**25.5**/56.5	54.1/78.0	**4.7**/**36.4**	**17.5**/55.1
F-GRPO	23.0/**63.4**	**12.1**/**46.1**	**1.0**/**29.5**	29.8/**90.6**	54.1/**92.9**	21.0/**60.1**	**20.1**/**61.3**	25.4/**57.6**	**56.4**/**79.6**	4.6/35.5	15.2/**57.6**
DAPO	24.3/54.2	**12.8**/40.8	1.0/18.5	**33.1**/79.5	55.9/83.8	**22.4**/54.1	21.0/48.4	23.9/51.3	51.2/77.8	**4.8**/28.9	15.7/47.0
F-DAPO	**24.8**/**62.3**	11.1/**44.4**	**1.7**/**28.7**	31.9/**88.3**	**58.6**/**92.0**	22.3/**59.3**	**23.2**/**61.3**	**24.8**/**55.4**	**53.0**/**79.5**	4.3/**33.0**	**17.0**/**53.7**
CISPO	24.1/58.0	9.7/39.4	1.0/**25.4**	32.9/79.1	**56.9**/89.1	21.8/**59.5**	**22.5**/55.4	**25.7**/52.5	**54.6**/**78.4**	4.3/29.4	**18.2**/**49.7**
F-CISPO	**24.5**/**59.7**	**10.6**/**42.8**	**2.0**/24.5	**34.1**/**82.6**	56.5/**91.0**	**22.1**/58.8	21.5/**58.7**	25.0/**53.0**	52.6/77.3	**5.4**/**33.9**	17.0/47.7

Table 1: Pass@1 / pass@256 across three models and six methods at N=8. Focal weighting (F-GRPO, F-DAPO, F-CISPO) consistently improves pass@256 with stable or improved pass@1. **Bold**: better within baseline/Focal pair; underline: statistically significant ($p<0.05$, see Appendix I).

of Problem Solving, 2024b), Minerva Math (Lewkowycz et al., 2022), and Olympiad Bench (He et al., 2024). To assess whether diversity benefits transfer beyond the training distribution, we include out-of-domain (OOD) benchmarks spanning distinct reasoning types: GPQA Diamond (Rein et al., 2023) (graduate-level science QA), IFEval (Zhou et al., 2023) (instruction following), and SynLogic (Liu et al., 2025a) (synthetic logical reasoning). Evaluation details are in Appendix H.

5.3 Group Size Regimes and Focal Weighting

Having observed the three-regime pattern in categorical simulation (Section 5.1), we examine whether analogous behavior arises in LLM training. Table 2 compares GRPO at $N \in \{2, 8, 32\}$ with F-GRPO at $N = 8$ on Qwen2.5-7B. These values are chosen to span different operating regimes while keeping rollout cost tractable; we do not aim to exhaustively map performance as a function of N.

Method	Avg. Math ↑	Avg. OOD ↑	$\Delta\text{NLL}_{\text{rare}}$ ↓
GRPO N=2	36.2 / **75.0**	18.0†/ **67.3**	**0.19**
GRPO N=8	37.3 / 64.1	17.1 / 55.9	0.68
GRPO N=32	**39.2** / 70.1	17.7 / 61.7	0.52
F-GRPO N=8	38.6†/ 70.3†	**19.2** / 63.3†	0.46†

Table 2: Comparison of GRPO at varying group sizes versus F-GRPO at fixed N=8 on Qwen2.5-7B. Metrics: average pass@1 / pass@256 on in-domain math and OOD benchmarks. $\Delta\text{NLL}_{\text{rare}}$: increase in negative log-likelihood on trajectories that were correct but low-probability under the base model (lower = less deviation from base distribution; see Appendix F.2). **Bold**: best; †: second best. Full per-benchmark results in Appendix F.

GRPO exhibits non-monotonic behavior: N=2 yields highest pass@256 but lowest pass@1, a pattern consistent with diversity preservation through infrequent active updates. At N=8, pass@1 improves but pass@256 drops to its lowest values across both in-domain and OOD benchmarks, suggesting distribution sharpening, consistent with prior observations (Yue et al., 2025; Dang et al., 2025). At N=32, pass@256 partially recovers while pass@1 continues to improve. This pattern aligns qualitatively with the three-regime framework of Section 3.1.

At N=8, F-GRPO matches GRPO at N=32 on pass@256 (70.3 vs. 70.1 on math; 63.3 vs. 61.7 on OOD) using $4\times$ fewer rollouts. Pass@1 shows a modest trade-off on in-domain benchmarks but improves on OOD tasks, suggesting that Focal weighting can mitigate concentration relative to GRPO at higher rollout budgets (e.g., N=32) without increasing the rollout budget in this setting.

Deviation from Base-Model Rare Solutions. We report $\Delta\text{NLL}_{\text{rare}}$, an empirical proxy for redistribution of probability mass away from solutions that were correct but low-probability under the base model (details in Appendix F.2). Higher values indicate greater deviation from the base distribution on

these trajectories. The ordering $\Delta\text{NLL}_{\text{rare}}(N{=}2) < \Delta\text{NLL}_{\text{rare}}(N{=}32) < \Delta\text{NLL}_{\text{rare}}(N{=}8)$ mirrors the pass@256 ordering, with F-GRPO at $N{=}8$ achieving an intermediate value (0.46) that reflects reduced concentration relative to its baseline.

5.4 Focal Weighting Across Methods

We evaluate Focal weighting on GRPO, DAPO, and CISPO at $N{=}8$, a commonly used group size (Shao et al., 2024; Zeng et al., 2025; Liu et al., 2025d). Table 1 reports results across three model families and scales.

Across all nine method–model combinations, Focal weighting improves both math and OOD pass@256 (average $+3.5$ and $+3.8$) while pass@1 remains stable or improves. The largest gains appear on Qwen2.5-7B (up to $+6.2$ math pass@256, $+7.4$ OOD pass@256). Notably, OOD pass@1 improves in 7/9 cases (average $+1.1$), suggesting that preserving solution diversity benefits generalization. Per-model breakdowns are in Table 1.

5.5 Comparison with Entropy and KL Regularization

We compare F-GRPO against common diversity-preserving regularizers: GRPO with entropy bonus (GRPO-$\mathcal{H}$) and GRPO with KL penalty (GRPO-KL) using Qwen2.5-7B setup. We tune coefficients following Appendix H; full results in Appendix G.

F-GRPO achieves the highest math pass@1 (38.6 vs. 37.8/37.2) and OOD pass@256 (63.3 vs. 59.9/60.0). GRPO-KL obtains higher math pass@256 (72.0 vs. 70.3), but requires maintaining a reference model in memory, increasing computational overhead. F-GRPO provides a simpler alternative with stronger pass@1 and OOD transfer.

6 Related Work

Distribution Sharpening in RLVR. RLVR improves pass@1 while degrading pass@k for large k (Dang et al., 2025; Yue et al., 2025; Wu et al., 2025a); Chen et al. (2025b) attribute this to overconfidence and propose confidence limiting. We identify a complementary mechanism: finite-sampling failure in group-relative methods that systematically misses rare-correct modes.

Group Size and Sampling Dynamics. Wu et al. (2025b) show $N{=}2$ suffices, while Hu et al. (2025) advocate large groups for coverage. Our tail-miss probability reconciles both: small and large N preserve diversity through inactivity and coverage respectively, while intermediate N maximizes sharpening risk.

Difficulty-Aware Training. Difficulty reweighting has roots in Focal loss (Lin et al., 2017) and curriculum learning (Bengio et al., 2009; Parashar et al., 2025). In RLVR, Zhou et al. (2025) rebalance loss across difficulty groups, He et al. (2025) up-weight rare correct trajectories, and Gai et al. (2025) propose differential smoothing of rewards. Unlike these trajectory-level modifications, we scale the entire prompt-level gradient contribution, directly targeting the $S_R > 0$ regime.

Entropy and Token-level Approaches. Entropy regularization remains debated (Cui et al., 2025; Cheng et al., 2025; Agarwal et al., 2025). Token-level reweighting methods (Hao et al., 2026; Peng et al., 2025; Wang et al., 2025) regulate mass distribution *within* trajectories; our Focal weighting is orthogonal, regulating *which prompts* contribute to learning.

7 Conclusion

This work identifies finite group size N as a critical factor driving distribution sharpening in group-relative RLVR with binary rewards, where intermediate rollout counts, most common in practice due to computational constraints, systematically suppress rare-correct trajectories while concentrating mass onto common solutions. Our theoretical analysis derives a closed-form tail-miss probability exhibiting non-monotonic dependence on N: small groups preserve diversity through inactivity, large groups through coverage, but intermediate N maximizes active updates that miss rare-correct modes. We further characterize redistribution within the correct set, proving that unsampled-correct mass can

shrink even as total correct mass grows. Motivated by this analysis, we propose Focal weighting, a lightweight difficulty, aware advantage scaling applicable to any group-relative objective including GRPO, DAPO, and CISPO. Empirically, we validate the three-regime behavior across different N values and demonstrate consistent pass@256 improvements while preserving or improving pass@1 across three model families, at no extra computational cost.

REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

Art of Problem Solving. Aime problems and solutions. https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions, 2024a. Accessed: 2025-04-20.

Art of Problem Solving. Amc problems and solutions. https://artofproblemsolving.com/wiki/index.php?title=AMC_Problems_and_Solutions, 2024b. Accessed: 2025-04-20.

Søren Asmussen and Peter W Glynn. *Stochastic simulation: algorithms and analysis*, volume 57. Springer, 2007.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://ronan.collobert.com/pub/2009_curriculum_icml.pdf.

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

Feng Chen, Allan Raventos, Nan Cheng, Surya Ganguli, and Shaul Druckmann. Rethinking fine-tuning when scaling test-time compute: Limiting confidence improves mathematical reasoning. *arXiv preprint arXiv:2502.07154*, 2025b.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. 2021.

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

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models, 2025. URL https://arxiv.org/abs/2505.22617.

Xingyu Dang, Christina Baek, Kaiyue Wen, Zico Kolter, and Aditi Raghunathan. Weight ensembling improves reasoning in language models. *arXiv preprint arXiv:2504.10478*, 2025.

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

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

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

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in rlvr: An entropy change perspective, 2026. URL https://arxiv.org/abs/2510.10150v2.

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

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, and et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

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

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

Hugging Face. Math-verify: A robust mathematical expression evaluation system. https://github.com/huggingface/Math-Verify, 2026. GitHub repository, commit ba3d3aa (latest at time of access), accessed 2026-01-25.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*, 2023.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for llms. *arXiv preprint arXiv:2510.13786*, 2025. doi: 10.48550/arXiv.2510.13786. URL https://arxiv.org/abs/2510.13786.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, and et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, et al. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9, 2024.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Junteng Liu, Yuanxiang Fan, Zhuo Jiang, Han Ding, Yongyi Hu, Chi Zhang, Yiqi Shi, Shitong Weng, Aili Chen, Shiqi Chen, Yunan Huang, Mozhi Zhang, Pengyu Zhao, Junjie Yan, and Junxian He. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *arXiv preprint arXiv:2505.19641*, 2025a. doi: 10.48550/arXiv.2505.19641. URL https://arxiv.org/abs/2505.19641. Version v4, last revised 4 Jun 2025.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025b.

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

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*, 2025d.

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

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Erran Li, Raluca Ada Popa, and Ion Stoica. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl, 2025. Notion Blog.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Rl squeezes, sft expands: A comparative study of reasoning llms, 2025. URL https://arxiv.org/abs/2509.21128.

Kangqi Ni, Zhen Tan, Zijie Liu, Pingzhi Li, and Tianlong Chen. Can grpo help llms transcend their pretraining origin? *arXiv preprint arXiv:2510.15990*, 2025.

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, et al. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*, 2025.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. Simko: Simple pass@k policy optimization, 2025. URL https://arxiv.org/abs/2510.14807. arXiv:2510.14807v2, last revised 21 Oct 2025.

Dimitris N. Politis, Joseph P. Romano, and Michael Wolf. *subsampling*. Springer Series in Statistics. Springer, New York, 1999.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022. Submitted on 20 Nov 2023.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv:2409.19256*, 2024. URL https://arxiv.org/abs/2409.19256. Submitted: 28 Sep 2024; PDF available.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

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

Fang Wu, Weihao Xuan, Ximing Lu, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may not escape its origin. *arXiv preprint arXiv:2507.14843*, 2025a.

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, et al. It takes two: Your grpo is secretly dpo. *arXiv preprint arXiv:2510.00977*, 2025b.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024a.

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

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

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

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. From $f(x)$ and $g(x)$ to $f(g(x))$: Llms learn new skills in rl by composing old ones. *arXiv preprint arXiv:2509.25123*, 2025.

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

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*, 2025.

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

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860, 2023. doi: 10.14778/3611540.3611569. URL https://doi.org/10.14778/3611540.3611569.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. Sglang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*, 2023. doi: 10.48550/arXiv.2312.07104. URL https://arxiv.org/abs/2312.07104. Submitted 12 Dec 2023; revised (v2) 6 Jun 2024.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023. doi: 10.48550/arXiv.2311.07911. URL https://arxiv.org/abs/2311.07911. Submitted on 14 Nov 2023.

Jingyu Zhou, Lu Ma, Hao Liang, Chengyu Shen, Bin Cui, and Wentao Zhang. Daro: Difficulty-aware reweighting policy optimization, 2025. URL https://arxiv.org/abs/2510.09001.

A Proof of Lemma 3.1

Proof. Fix a prompt x and omit (x) for readability. Each rollout falls into one of three disjoint regions: the rare-correct region $\mathcal{C}_{\text{rare}}$ with probability τ, the remaining correct region $\mathcal{C} \setminus \mathcal{C}_{\text{rare}}$ with probability $\mu_{\text{pos}} - \tau$, or the incorrect region $\Omega \setminus \mathcal{C}$ with probability $1 - \mu_{\text{pos}}$.

The probability that no rollout lies in the rare-correct region is $(1-\tau)^N$. Conditioned on this event, all rollouts lie in $(\mathcal{C} \setminus \mathcal{C}_{\text{rare}}) \cup (\Omega \setminus \mathcal{C})$. The group is inactive (hence $\mathcal{B}_\tau$ does not occur) in two disjoint cases: all rollouts are correct but not rare-correct, with probability $(\mu_{\text{pos}} - \tau)^N$; or all rollouts are incorrect, with probability $(1-\mu_{\text{pos}})^N$. Thus

$$\Pr(\mathcal{B}_\tau) = (1-\tau)^N - (\mu_{\text{pos}} - \tau)^N - (1-\mu_{\text{pos}})^N. \qquad \square$$

B Monotonicity of Sampled Distinct Mass Conditioned on X

This appendix formalizes the monotonicity claim used in Section 4 (Focal Weight): although the categorical baseline S_R depends on the probability mass of *distinct* sampled rollouts, its conditional expectation is monotone in the observed correct count X.

Setup. Fix a prompt x and write $\pi(o) := \pi_\theta(o \mid x)$ for brevity. Let Ω_x be the rollout space and $\mathcal{C} := \mathcal{C}(x) \subseteq \Omega_x$ the set of correct rollouts (Section 2.1). Sample N i.i.d. rollouts $o_1, \ldots, o_N \sim \pi(\cdot)$, and let $X := \sum_{i=1}^N \mathbb{I}[o_i \in \mathcal{C}]$ be the number of correct rollouts.

Define the *distinct sampled sets*

$$A := \{\, o_i : o_i \in \mathcal{C} \,\}, \qquad B := \{\, o_i : o_i \notin \mathcal{C} \,\},$$

where braces denote a set (duplicates removed). Define the corresponding sampled masses

$$P_{\text{pos}} := \sum_{o \in A} \pi(o), \qquad P_{\text{neg}} := \sum_{o \in B} \pi(o).$$

These are the trajectory-level analogues of the categorical quantities in Section 2.3. As in that section, define

$$S_R := R_c\, P_{\text{pos}} + R_w\, P_{\text{neg}}. \tag{12}$$

Conditional Distributions. Let $\mu_{\text{pos}} := \Pr_{o\sim\pi}[o \in \mathcal{C}]$. For $o \in \mathcal{C}$, define the conditional (restricted) distribution

$$q_{\text{pos}}(o) := \Pr[o_i = o \mid o_i \in \mathcal{C}] = \frac{\pi(o)}{\mu_{\text{pos}}}.$$

Similarly, for $o \notin \mathcal{C}$, define

$$q_{\text{neg}}(o) := \Pr[o_i = o \mid o_i \notin \mathcal{C}] = \frac{\pi(o)}{1-\mu_{\text{pos}}}.$$

By exchangeability of i.i.d. sampling, conditioning on $X = k$ implies that the k correct rollouts are i.i.d. from q_{pos} over $\mathcal{C}$, and the $N-k$ incorrect rollouts are i.i.d. from q_{neg} over $\Omega_x \setminus \mathcal{C}$.

Lemma B.1. *For all integers $k \in \{0, 1, \ldots, N\}$,*

$$\mathbb{E}[P_{\text{pos}} \mid X = k] = \sum_{o \in \mathcal{C}} \pi(o) \left(1 - (1 - q_{\text{pos}}(o))^k\right), \tag{13}$$

$$\mathbb{E}[P_{\text{neg}} \mid X = k] = \sum_{o \notin \mathcal{C}} \pi(o) \left(1 - (1 - q_{\text{neg}}(o))^{N-k}\right). \tag{14}$$

Moreover, $\mathbb{E}[P_{\text{pos}} \mid X = k]$ is non-decreasing in k, and $\mathbb{E}[P_{\text{neg}} \mid X = k]$ is non-increasing in k.

Proof. We prove the statement for P_{pos}; the argument for P_{neg} is identical with $N-k$ in place of k.

Condition on $X = k$. For any fixed $o \in \mathcal{C}$, the event $\{o \in A\}$ is exactly the event that o appears at least once among the k correct i.i.d. draws from q_{pos}. Thus

$$\Pr(o \in A \mid X = k) = 1 - (1 - q_{\text{pos}}(o))^k.$$

Using linearity of expectation and the definition $P_{\text{pos}} = \sum_{o \in \mathcal{C}} \pi(o)\, \mathbb{I}\{o \in A\}$,

$$\mathbb{E}[P_{\text{pos}} \mid X = k] = \sum_{o \in \mathcal{C}} \pi(o)\ \Pr(o \in A \mid X = k) = \sum_{o \in \mathcal{C}} \pi(o)\, \Big(1 - (1 - q_{\text{pos}}(o))^k\Big),$$

which is equation 13.

To show monotonicity, compute the discrete difference:

$$\mathbb{E}[P_{\text{pos}} \mid X = k{+}1] - \mathbb{E}[P_{\text{pos}} \mid X = k] = \sum_{o \in \mathcal{C}} \pi(o)\, (1 - q_{\text{pos}}(o))^k\, q_{\text{pos}}(o)\ \geq\ 0,$$

so $\mathbb{E}[P_{\text{pos}} \mid X = k]$ is non-decreasing in k.

For P_{neg}, conditioned on $X = k$, each $o \notin \mathcal{C}$ is included in B with probability $1 - (1 - q_{\text{neg}}(o))^{N-k}$. This yields equation 14. Since $N{-}k$ decreases as k increases and $m \mapsto 1{-}(1{-}q)^m$ is non-decreasing in m, it follows that $\mathbb{E}[P_{\text{neg}} \mid X = k]$ is non-increasing in k. □

Corollary B.2. *Assume standard RLVR rewards $R_c > R_w$ and $R_w \leq 0$ (Section 2.1). Then $\mathbb{E}[S_R \mid X = k]$ is non-decreasing in k.*

Proof. By definition equation 12 and linearity of expectation,

$$\mathbb{E}[S_R \mid X = k] = R_c\, \mathbb{E}[P_{\text{pos}} \mid X = k] + R_w\, \mathbb{E}[P_{\text{neg}} \mid X = k].$$

By Lemma B.1, the first term is non-decreasing in k because $R_c > 0$, and the second term is also non-decreasing in k because $R_w \leq 0$ and $\mathbb{E}[P_{\text{neg}} \mid X = k]$ is non-increasing in k. Hence their sum is non-decreasing in k. □

C First-order Softmax Expansion and Subset-mass Identity

This appendix records standard first-order identities for the softmax map that underlie the analysis in Section 3.

Let $p = \mathrm{softmax}(z)$ over $\mathcal{A}$ and consider a small logit perturbation Δz. The softmax Jacobian $\frac{\partial p_i}{\partial z_j} = p_i(\mathbf{1}\{i = j\} - p_j)$ implies the first-order probability change

$$\Delta p_i = \sum_{j \in \mathcal{A}} \frac{\partial p_i}{\partial z_j} \Delta z_j = p_i \Big(\Delta z_i - \sum_{j \in \mathcal{A}} p_j \Delta z_j\Big). \tag{15}$$

For any subset $\mathcal{S} \subseteq \mathcal{A}$, define its probability mass $Q_{\mathcal{S}} := \sum_{i \in \mathcal{S}} p_i$. Summing equation 15 over $i \in \mathcal{S}$ yields the *subset-mass identity*:

$$\Delta Q_{\mathcal{S}} := \sum_{i \in \mathcal{S}} \Delta p_i = \sum_{i \in \mathcal{S}} p_i \Delta z_i - Q_{\mathcal{S}} \sum_{j \in \mathcal{A}} p_j \Delta z_j. \tag{16}$$

The first term captures the direct effect of logit changes on actions in $\mathcal{S}$, while the second term captures the indirect effect through softmax normalization: when probability mass moves elsewhere, $Q_{\mathcal{S}}$ changes even if the logits of actions in $\mathcal{S}$ are unchanged.

Application to ΔQ_{pos}. Setting $\mathcal{S} = \mathcal{P}$ and using the one-step update equation 4 recovers the mass balance equation equation 5 of (Hu et al., 2025).

Application to $\Delta Q_{\text{u,pos}}$. Setting $\mathcal{S} = U \cap \mathcal{P}$ (unsampled correct actions) yields Proposition 3.2. The key observation is that for $i \in U$, we have $R_i = 0$, so $\Delta z_i = -\frac{\eta}{N} S_R p_i$ from equation 4.

D Proof of Proposition 3.2

Proof. Apply the subset-mass identity (Appendix C, Eq. equation 16) with $\mathcal{S} = U \cap \mathcal{P}$:

$$\Delta Q_{\text{u,pos}} = \sum_{i \in U \cap \mathcal{P}} p_i \Delta z_i - Q_{\text{u,pos}} \sum_{j \in \mathcal{A}} p_j \Delta z_j. \tag{17}$$

For $i \in U \cap \mathcal{P}$, we have $R_i = 0$, so by equation 4, $\Delta z_i = -\frac{\eta}{N} S_R p_i$. Thus the first sum becomes

$$\sum_{i \in U \cap \mathcal{P}} p_i \Delta z_i = -\frac{\eta}{N} S_R \sum_{i \in U \cap \mathcal{P}} p_i^2 = -\frac{\eta}{N} S_R \, U_{\mathrm{pos},2}. \tag{18}$$

For the normalization term, partitioning by reward value:

$$\begin{aligned} \sum_{j \in \mathcal{A}} p_j \Delta z_j &= \frac{\eta}{N} \sum_{j \in \mathcal{A}} p_j^2 (R_j - S_R) \\ &= \frac{\eta}{N} \Big[(R_c - S_R) A_2 + (R_w - S_R) B_2 - S_R U_2 \Big], \end{aligned} \tag{19}$$

where we used $R_j = R_c$ for $j \in A$, $R_j = R_w$ for $j \in B$, and $R_j = 0$ for $j \in U$. Substituting both expressions yields equation 8. □

E Detailed Term Analysis for Proposition 3.2

We analyze each term in equation 8 to understand when unsampled-correct mass decreases.

Direct drift term. The term $-S_R \, U_{\mathrm{pos},2}$ arises because unsampled actions receive zero reward but are still affected by the baseline subtraction. When $S_R > 0$ (reward-positive batch), this term is negative and pushes unsampled-correct mass downward. The magnitude scales with $U_{\mathrm{pos},2}$, the concentration of unsampled-correct probability.

Normalization coupling. The second term couples $Q_{\mathrm{u,pos}}$ to the mass changes elsewhere. The factor in parentheses has three components:

- $(R_c - S_R)A_2 \geq 0$: sampled-correct actions gain probability, which through normalization draws mass away from unsampled-correct actions.
- $(R_w - S_R)B_2 \leq 0$: sampled-incorrect actions lose probability, which through normalization donates mass to all other actions including unsampled-correct ones.
- $-S_R U_2$: when $S_R > 0$, unsampled actions (both correct and incorrect) lose probability through baseline subtraction.

When does $\Delta Q_{\mathrm{u,pos}} < 0$ while $\Delta Q_{\mathrm{pos}} > 0$? Consider a reward-positive batch ($S_R > 0$) on a prompt with high success probability. In this regime:

- The direct drift $-S_R U_{\mathrm{pos},2} < 0$ actively pushes unsampled-correct mass down.
- The normalization coupling is dominated by $(R_c - S_R)A_2 > 0$ when sampled-correct mass is concentrated, further draining unsampled-correct mass.
- Meanwhile, ΔQ_{pos} from equation 5 remains positive because its first two terms (mass transfer from incorrect to correct pool) outweigh the unsampled coupling.

Thus RLVR can increase total correct mass while concentrating it onto the sampled-correct subset, shrinking the probability of correct actions that happen not to be sampled.

F Group Size Comparison: Full Results

F.1 Per-Benchmark Results

Table 3 provide full per-benchmark results for the group size comparison discussed in Section 5.3.

F.2 NLL on Rare-Correct Trajectories

To construct a proxy for rare-correct modes, we sample 256 prompts from the training set and generate 800 rollouts per prompt from the base model, retaining only correct trajectories. For each retained trajectory, we compute its length-normalized NLL under the base model. We define the "rare-correct" subset as the top 1% by base-model NLL among these correct trajectories, yielding 1,263 trajectories in total. We then compute the NLL of this fixed subset under each trained model; larger values indicate reduced probability assigned to these initially low-probability correct solutions.

Method	In-domain							Out-of-domain			
	Avg.	AIME24	AIME25	AMC	MATH500	Minerva	Olympiad	Avg. OOD	IFEval	SynLogic	GPQA
GRPO $N{=}2$	36.2 / **75.0**	12.7 / **59.1**	8.3 / **56.0**	51.9 / **97.0**	74.5 / **96.7**	33.2 / **65.6**	36.7 / **75.7**	18.0†/ **67.3**	29.4 / **77.2**	6.7 / 54.3	**17.8** / **70.3**
GRPO $N{=}8$	37.3 / 64.1	15.0†/ 37.7	6.7 / 40.8	52.9 / 87.3	75.8 / 92.8	**36.0** / 60.2	37.8†/ 65.8	17.1 / 55.9	32.1†/ 70.3	7.9 / 51.3	11.3 / 46.2
GRPO $N{=}32$	**39.2** / 70.1	13.0 / 50.2†	**10.4** / 49.5	**60.9** / 95.5	**77.3** / 94.3	34.9 / 59.9	**38.9** / 71.3	17.7 / 61.7	31.0 / 71.4	**8.9** / **61.6**	13.4 / 51.9
F-GRPO $N{=}8$	38.6†/ 70.3†	**15.9** / 46.2	10.1†/ 52.6†	56.2†/ 96.3†	76.2†/ 95.1†	35.7†/ 60.3†	37.5 / 71.6†	**19.2** / 63.3†	**34.0** / 75.7†	8.7†/ 57.0†	15.0†/ 57.3†

Table 3: GRPO with different N and F-GRPO on both in-domain math and out-of-domain benchmarks (Qwen2.5-7B). Pass@1 / Pass@256. **Bold**: best; †: second best.

G Entropy and KL Regularization: Full Results

Method	In-domain							Out-of-domain			
	Avg.	AIME24	AIME25	AMC	MATH500	Minerva	Olympiad	Avg. OOD	IFEval	SynLogic	GPQA
F-GRPO	**38.6** / 70.3†	**15.9** / 46.2	**10.1** / 52.6†	**56.2** / **96.3**	76.2†/ 95.1†	**35.7** / 60.3	37.5 / 71.6†	19.2†/ **63.3**	**34.0** / **75.7**	8.7 / 57.0†	15.0†/ 57.3†
GRPO ($\mathcal{H}$)	37.8†/ 69.5	14.9†/ 48.9†	7.3 / 52.2	55.8†/ 90.8	75.6 / 94.6	34.9†/ 61.3†	**38.2** / 69.2	18.7 / 59.9	32.1 / 71.9†	**9.8** / **59.9**	14.3 / 47.8
GRPO (KL)	37.2 / **72.0**	13.2 / **53.4**	8.7†/ **53.7**	52.1 / 95.9†	**76.7** / **95.2**	34.7 / **61.5**	38.0†/ **72.3**	**19.4** / 60.0†	32.4†/ 70.8	8.8†/ 51.7	**17.1** / **57.5**

Table 4: F-GRPO vs. GRPO with entropy bonus (GRPO-$\mathcal{H}$, coefficient 0.001) and KL penalty (GRPO-KL, coefficient 0.001) on Qwen2.5-7B at $N{=}8$. Pass@1 / pass@256. **Bold**: best; †: second best.

H Experimental Details

H.1 Dataset Preprocessing

All models are trained on the DeepScaleR math dataset (Luo et al., 2025). We filter samples longer than 1024 tokens and remove duplicates with conflicting answers, retaining 39,202 samples. The system prompt `"Please reason step by step, and put your final answer within \boxed{}."` is prepended to all training inputs.

H.2 Training Configuration

Training uses the verl pipeline (Sheng et al., 2024) with sglang (Zheng et al., 2023) for rollout generation, on 16 NVIDIA H100 GPUs with FSDP2 (Zhao et al., 2023). Maximum response lengths are 3072 tokens for Qwen2.5-1.5B-Math and 8192 tokens for other models. Following (Yu et al., 2025), we drop the KL-divergence regularization term and use token-mean loss aggregation. In all our experiments we use learning rate 1×10^{-6} according to (Shao et al., 2024; Yu et al., 2025).

Clipping parameters: $\epsilon_{\text{low}}{=}0.2$, $\epsilon_{\text{high}}{=}0.2$ for GRPO; $\epsilon_{\text{low}}{=}0.2$, $\epsilon_{\text{high}}{=}0.28$ for DAPO; $\epsilon_{\text{low}}{=}1.0$, $\epsilon_{\text{high}}{=}5.0$ for CISPO, following (Khatri et al., 2025). Rewards are assigned via math-verify (Hugging Face, 2026): 1.0 for correct, 0.0 for incorrect. Complete hyperparameters are in Table 5.

Entropy and KL Regularization: for the comparison in Section 5.5, we tune the entropy bonus coefficient over $\{0.0001, 0.001\}$ and the KL penalty coefficient over $\{0.001, 0.01\}$. We select the best checkpoint for each configuration based on average math pass@1. The best-performing coefficients are 0.001 for both entropy bonus and KL penalty.

H.3 Focal Weight Hyperparameter γ

We sweep the Focal exponent $\gamma \in \{0.5, 1.0, 2.0\}$ for each Focal-weighted method (F-GRPO, F-DAPO, F-CISPO) and select the best value by average in-domain math pass@1 at the best checkpoint. For reproducibility, the selected γ values for the setups reported in Table 1 are summarized in Table 6. Overall, the method is robust to the choice of γ: across setups, the best results are attained at both $\gamma = 0.5$ and $\gamma = 1.0$.

Parameter	Value
Optimizer	AdamW (Loshchilov & Hutter, 2017)
(β_1, β_2)	(0.9, 0.999)
Weight decay	0.01
Gradient norm clipping	1.0
Learning rate	1×10^{-6}
LR scheduler	Constant
Warmup steps	15
Global batch size	256
Mini-batch size	64
Num training epochs	10
PPO epochs	1
Sampling temperature	1.0
(top-p, top-k)	(1.0, -1)

Table 5: Training hyperparameters.

Model	F-GRPO γ	F-DAPO γ	F-CISPO γ
Qwen2.5-7B	0.5	0.5	1.0
Qwen2.5-1.5B-Math	0.5	0.5	1.0
Llama-3.2-3B-Instruct	0.5	1.0	0.5

Table 6: Selected Focal weight γ for each method-model setup at N=8 (Table 1). The sweep range is $\{0.5, 1.0, 2.0\}$.

H.4 Evaluation Protocol

We report unbiased pass@k estimator Chen et al. (2021), the probability that at least one of k samples is correct:

$$\text{pass@}k := \mathbb{E}_{\text{Problems}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right], \tag{20}$$

where n is the total number of samples and c is the number of correct samples. We use $n = 256$ samples per problem and report pass@1 and pass@256.

For checkpoint selection, we save a checkpoint at the end of each epoch. We choose the best baseline checkpoint by average math pass@1, then compare to the best F-GRPO checkpoint obtained with equal or less compute. Evaluation uses sglang (Zheng et al., 2023) and math-verify (Hugging Face, 2026). Configurations and system prompts are in Tables 7 and 8.

I Statistical Significance

To assess the statistical significance of performance differences between the baseline and F-GRPO models, we employ a paired m-out-of-n subsampling test following (Politis et al., 1999). For each benchmark, we generate $n = 1024$ solutions per problem and use $m = 256$ generations (i.e., subsample size m) to estimate pass@1 and pass@256 metrics. Specifically, for each subsampling iteration we randomly sample $m = 256$ generations without replacement for each problem, compute the pass@k metric using the analytical formula $1 - \binom{n-c}{k}/\binom{n}{k}$ where n is the number of sampled generations and c is the number of correct solutions among them, and average across all problems to obtain a single pass@k estimate for both baseline and F-GRPO models. We perform 50,000 subsampling iterations to obtain the distribution of paired differences in pass@k between the two models.

We conduct a two-sided statistical test with significance level $\alpha = 0.05$. A difference is considered statistically significant if the two-sided p-value is less than 0.05, which is equivalent to the 95% confidence interval of the subsampling distribution not containing zero.

Parameter	Qwen2.5-7B	Qwen2.5-1.5B-Math	Llama3.2-3B
Temperature	1.0	1.0	1.0
top-p	1.0	1.0	1.0
top-k	-1	-1	-1
Max length	8192	3072	8192

Table 7: Evaluation configurations.

Benchmark	Qwen	Llama
Mathematical reasoning	`Please reason step by step, and put your final answer within \boxed{}.`	`Cutting Knowledge Date: December 2023\nToday Date: [date]\nPlease reason step by step, and put your final answer within \boxed{}.`
GPQA Diamond	`Please reason step by step, and put your final answer within \boxed{}.`	`Cutting Knowledge Date: December 2023\nToday Date: [date]\nPlease reason step by step, and put your final answer within \boxed{}.`
IFEval	`You are a helpful assistant.`	`Cutting Knowledge Date: December 2023\nToday Date: [date]`
SynLogic	`You are a helpful assistant.`	`Cutting Knowledge Date: December 2023\nToday Date: [date]`

Table 8: System prompts for evaluation.

J Categorical Simulation Details

We validate the theoretical framework using a categorical policy simulation. To enable direct comparison with prior work, we adopt the setup of Hu et al. (2025) with one modification to the learning rate, as described below.

The policy is a softmax distribution over $|\mathcal{A}| = 128{,}000$ actions. A subset $\mathcal{A}^+$ of 10,000 actions is designated as correct with reward $R = +1$; the remaining 118,000 actions receive $R = -1$. Following Hu et al. (2025), logits are initialized as: one "anchor" correct action receives $z_{\text{anchor}} = 5.0$; all other correct actions receive $z = 3.0$; incorrect actions receive $z = 0.0$. Under softmax with temperature $\tau = 1$, this yields initial total correct mass $Q_{\text{pos}} \approx 0.63$, anchor probability $p_{\text{anchor}} \approx 4.7 \times 10^{-4}$, and probability $\tau_{\text{leaf}} \approx 6.3 \times 10^{-5}$ for each non-anchor correct action.

Given this initial distribution, we can compute the tail-miss probability $\Pr(\mathcal{B}_\tau)$ from Lemma 3.1 for a typical non-anchor correct action with $\tau = \tau_{\text{leaf}} \approx 6.3 \times 10^{-5}$. Figure 5 shows $\Pr(\mathcal{B}_\tau)$ as a function of group size N. The probability rises steeply for small N, plateaus near 1 for intermediate values, and only declines toward zero for $N \gtrsim 2^{15}$. At $N = 2^{17} = 131{,}072$, $\Pr(\mathcal{B}_\tau) < 10^{-3}$, predicting that such a group size should preserve probability mass on non-anchor correct actions. This prediction aligns with the simulation results in Figure 4: N=131,072 is the only configuration that maintains $\mathcal{M}_{\text{ret}} \approx 1$ throughout training.

At each training step, we sample N actions i.i.d. from the current policy, compute group-relative advantages $\tilde{r}_j = R_j - \frac{1}{N}\sum_k R_k$, and update logits via gradient ascent on $\mathcal{L} = \frac{1}{N}\sum_j \tilde{r}_j p_j$. When Focal weighting is applied, objective is scaled by $g = (1 - \widehat{\mu}_{\text{pos}})^\gamma$. We use learning rate $\eta = 10^{-2}$, which differs from $\eta = 10^{-3}$ in Hu et al. (2025). At the lower learning rate,

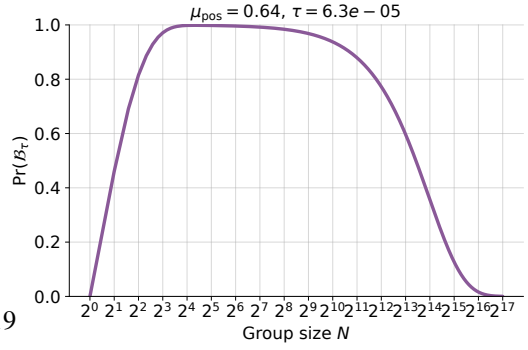

Figure 5: Tail-miss probability $\Pr(\mathcal{B}_\tau)$ versus

policy entropy after $1{,}000$ steps remained above 4 even for $N{=}65{,}536$, whereas LLM generation entropy during RLVR training is typically below 1. The higher learning rate produces dynamics that better reflect the concentration regimes observed in practice.

We sweep $N \in \{2, 4, \ldots, 131{,}072\}$ and $\gamma \in \{0, 1\}$, running $T{=}1{,}000$ steps per configuration. Results are averaged over 4 random seeds.

Metrics. We track total correct mass $Q_{\mathrm{pos}}(t) = \sum_{a \in \mathcal{A}^+} \pi_t(a)$ and retained positive mass:

$$\mathcal{M}_{\mathrm{ret}}(t) = 1 - \frac{\sum_{a \in \mathcal{A}^+} \max\big(0,\, \pi_0(a) - \pi_t(a)\big)}{\sum_{a \in \mathcal{A}^+} \pi_0(a)}. \tag{21}$$

$\mathcal{M}_{\mathrm{ret}}{=}1$ indicates no correct action has lost mass; $\mathcal{M}_{\mathrm{ret}}{\approx}0$ indicates concentration onto a smaller subset.

K NOTATION

Table 9 summarizes the main notation used throughout the paper.

Category	Symbol	Meaning
Trajectory-level variables	π_θ	The policy parameterized by θ
	x	Given prompt
	o, y	A complete response (trajectory) generated by π_θ when given x
	y_t	The t-th token of response y
	N	Group size: number of rollouts sampled per prompt
	R_i	Binary reward for rollout i (R_c if correct, R_w if incorrect)
	R_c, R_w	Reward values for correct and incorrect rollouts ($R_c > R_w$)
	$\mu_{\text{pos}}(x)$	Success probability: $\Pr_{o\sim\pi_\theta(\cdot\|x)}[o \in \mathcal{C}(x)]$
	$\tau(x)$	Rare-correct mass: $\Pr_{o\sim\pi_\theta(\cdot\|x)}[o \in \mathcal{C}_{\text{rare}}(x)]$
	$\rho(x)$	Ratio of rare-correct to total correct mass: $\tau(x)/\mu_{\text{pos}}(x)$
	$\widehat{\mu}_{\text{pos}}(x)$	Empirical success rate: fraction of correct rollouts in the sampled group
Categorical framework variables	$p = \text{softmax}(z)$	Policy over finite action space $\mathcal{A}$
	z_i	Logit for action i
	$\mathcal{P}, \mathcal{N}$	Sets of correct and incorrect actions
	A, B, U	Sampled correct actions, sampled incorrect actions, and unsampled actions
	$Q_{\text{pos}}, Q_{\text{neg}}$	Total correct and incorrect probability masses
	$P_{\text{pos}}, P_{\text{neg}}$	Sampled correct and incorrect probability masses
	$Q_{\text{u,pos}}$	Unsampled-correct probability mass
	A_2, B_2	Second moments: $\sum_{i\in A} p_i^2$, $\sum_{i\in B} p_i^2$
	U_2	Unsampled second moment: $\sum_{i\in U} p_i^2$
	$U_{\text{pos},2}, U_{\text{neg},2}$	Unsampled second moments for correct and incorrect actions
Expressions and operators	$\pi_\theta(\cdot \mid x, y_{<t})$	Conditional probability of generating token $\cdot$ given prompt x and previous tokens $y_{<t}$
	$\bar{R}$	Group mean reward: $\frac{1}{N}\sum_{j=1}^{N} R_j$
	σ_R	Standard deviation of rewards in the group
	$\widehat{A}_i^{\text{GRPO}}$	Group-relative advantage: $(R_i - \bar{R})/(\sigma_R + \epsilon)$
	$\widehat{A}_i^{\text{F-GRPO}}$	Focal-weighted advantage: $g(x) \cdot \widehat{A}_i^{\text{GRPO}}$
	$r_{i,t}(\theta)$	Importance ratio: $\pi_\theta(y_{i,t} \mid x, y_{i,<t})/\pi_{\theta_{\text{old}}}(y_{i,t} \mid x, y_{i,<t})$
	S_R	Batch baseline: $R_c P_{\text{pos}} + R_w P_{\text{neg}}$
	Δz_i	One-step logit update: $\frac{\eta}{N} p_i (R_i - S_R)$
	ΔQ_{pos}	One-step change in total correct mass
	$\Delta Q_{\text{u,pos}}$	One-step change in unsampled-correct mass
	$g(x)$	Difficulty weight: $(1 - \widehat{\mu}_{\text{pos}}(x))^\gamma$
	γ	Focal loss parameter controlling difficulty weighting strength
	η	Learning rate
	$\mathcal{M}_{\text{ret}}(t)$	Retained positive mass: fraction of initial correct probability that has not decreased at step t
Events and probabilities	$\mathcal{A}_N$	Active event: $\{0 < X < N\}$ where X is the number of correct rollouts
	$\mathcal{B}_\tau$	Tail-miss event: active update that misses rare-correct region
	$\Pr(\mathcal{B}_\tau)$	Probability of tail-miss event
Sets	Ω_x	Space of complete rollouts for prompt x
	$\mathcal{C}(x)$	Subset of correct rollouts for prompt x
	$\mathcal{C}_{\text{rare}}(x)$	Subset of rare-correct rollouts for prompt x
	$\mathcal{A}$	Finite action space in the categorical framework
	$\mathcal{A}^+$	Subset of correct actions in the categorical simulation

Table 9: Notation used in the paper.