# F-GRPO: Don't Let Your Policy Learn the Obvious and Forget the Rare

Anonymous ACL submission

## Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) is commonly based on group sampling to estimate advantages and stabilize policy updates. In practice, computational limits often rule out very large groups, so training proceeds with finite rollout sets that can reinforce only the correct behavior they expose. At practical group sizes, updates can miss rare-correct trajectories while still containing mixed rewards, concentrating probability on more common sampled solutions. We derive the probability of such prompt-local tail-miss events as a function of group size, showing non-monotonic behavior, and in the categorical abstraction characterize how unsampled-correct mass can shrink even as total correct mass grows. Motivated by this analysis, we propose a difficulty-aware scaling coefficient, inspired by Focal loss, that down-weights updates on high-success sampled groups. Empirically, categorical simulation illustrates the same effect in the categorical setting, Maze provides a single-solution test, and LLM experiments include a representative GRPO group-size sweep together with fixed-$N$ transfer across GRPO, DAPO, and CISPO. On Qwen2.5-7B at $N$=8, our method improves average math pass@256 from 64.1 $\rightarrow$ 70.3 (GRPO), 69.3 $\rightarrow$ 72.5 (DAPO), and 73.2 $\rightarrow$ 76.8 (CISPO); OOD pass@256 also improves in all three cases, 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 controls how much of a prompt's rollout space is exposed to each update, 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?*

Our contributions are as follows:

- We make explicit the tail-miss probability, characterizing when an active RLVR update omits a rare correct subset for a prompt. Its non-monotonic

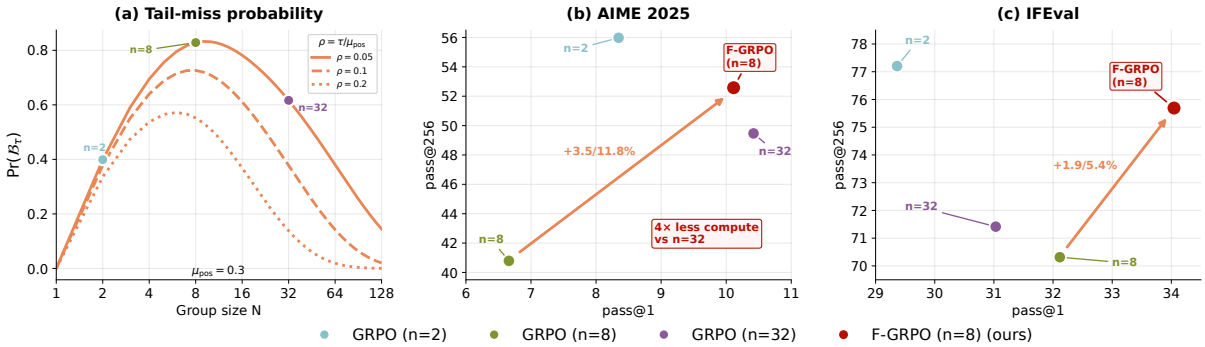


Figure 1: (a) Schematic theoretical tail-miss probability: an update is *active* (mixed rewards) but *misses* rare-correct solutions within a prompt. The curve is non-monotonic in group size $N$: small groups are often inactive, large groups cover rare modes, and intermediate groups can combine activity with incomplete 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× less compute than $N$=32.

dependence on group size helps reconcile prior conflicting findings: small groups can preserve high pass@$k$ through inactivity, large groups through coverage, while intermediate groups, common under compute constraints, may maximize active tail-miss events.

- Building on the categorical framework of Hu et al. (2025), we analyze redistribution within the correct set and show that unsampled-correct mass can decrease even when total correct mass increases.
- We use a single-solution Maze setup, showing that pass@$K$ can degrade even when each prompt has a unique correct sequence.
- We propose F-GRPO, a difficulty-aware advantage scaling applicable to group-relative objectives including GRPO, DAPO, and CISPO, and show at fixed $N$=8 that it transfers across methods and models, consistently improving math and OOD pass@256 while preserving or improving pass@1, without additional rollout cost.

## 2 Preliminaries

### 2.1 Reinforcement Learning with Verifiable Rewards

We consider RLVR for language model reasoning. Given a prompt $x$ from a distribution $D$, the policy $\pi_\theta$ 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

$$R_i = R_w + (R_c - R_w)\,\mathbb{I}[o_i \text{ is correct}] \quad (1)$$

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 $\Omega_x$ denote the space of complete rollouts and $\mathcal{C}(x) \subseteq \Omega_x$ the subset of correct rollouts. The per-prompt success probability is

$$\mu_{\text{pos}}(x) := \Pr_{o \sim \pi_\theta(\cdot|x)}[o \in \mathcal{C}(x)]. \quad (2)$$

In training, prompts are drawn from the empirical prompt distribution; the finite-sampling analysis below is stated conditionally on a fixed prompt $x$.

### 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

$$\widehat{A}_i^{\text{GRPO}} = \frac{R_i - \bar{R}}{\sigma_R + \epsilon}, \quad (3)$$

where $\bar{R} = \frac{1}{N}\sum_{j=1}^N R_j$ and $\sigma_R = \text{std}(\{R_j\}_{j=1}^N)$.

GRPO optimizes a clipped surrogate objective using token-level importance ratios; Appendix B records the GRPO, DAPO, and CISPO objective details used in our experiments.

A key property of group-relative advantages is that when all sampled rewards are identical ($\sigma_R = 0$), we have $\widehat{A}_i^{\text{GRPO}} = 0$ 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

$$Q_{\mathrm{pos}} := \sum_{i \in \mathcal{P}} p_i, \quad Q_{\mathrm{neg}} := 1 - Q_{\mathrm{pos}}. \tag{4}$$

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 $P_{\mathrm{pos}} := \sum_{i \in A} p_i$, $P_{\mathrm{neg}} := \sum_{i \in B} p_i$, $A_2 := \sum_{i \in A} p_i^2$, $B_2 := \sum_{i \in B} p_i^2$, $U_{\mathrm{pos},2} := \sum_{i \in U \cap \mathcal{P}} p_i^2$, and $U_{\mathrm{neg},2} := \sum_{i \in U \cap \mathcal{N}} p_i^2$. Assign rewards as in (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 and 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

$$\Delta z_i = \frac{\eta}{N} p_i (R_i - S_R), \tag{5}$$

where $\eta$ 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; Appendix C records the expression and term interpretation.

This categorical framework captures one-step local redistribution under sampled updates and provides the controlled abstraction used in Section 3.2. To avoid conflation, we keep its notation separate from trajectory-level quantities: $\mu_{\mathrm{pos}}(x)$ denotes per-prompt success probability (2), while $Q_{\mathrm{pos}}$ denotes positive mass in the categorical abstraction (4).

## 3 Finite-Sampling Bias in Group-Relative RLVR

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). These recommendations can be understood by analyzing RLVR at the level where group-relative sampling acts directly: a finite set of rollouts for a prompt. We study two complementary consequences of this finite sampling. First, for a fixed prompt, an update may be active while omitting a low-mass correct region. Second, given the sampled set, local redistribution can move probability away from unsampled correct outcomes. Together, these quantities describe how finite groups can improve accuracy while narrowing coverage.

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

We first isolate a single prompt $x$ and ask when a finite group can update the policy without ever sampling a low-mass correct region. Fix a target subset $E_x \subseteq \mathcal{C}(x)$ of correct rollouts. Let

$$\tau_E(x) := \Pr_{o \sim \pi_\theta(\cdot|x)}[o \in E_x], \tag{6}$$

and assume $0 < \tau_E(x) < \mu_{\mathrm{pos}}(x)$. We call $E_x$ rare under the current policy when $\tau_E(x)/\mu_{\mathrm{pos}}(x)$ is small.

Let $X_x$ denote the number of correct rollouts among the $N$ samples for prompt $x$. For group-relative methods such as GRPO, the learning signal vanishes when all sampled rewards are identical, i.e., $X_x \in \{0, N\}$. Define the *active event*

$$\mathcal{A}_N(x) := \{0 < X_x < N\}, \tag{7}$$

with probability $\Pr(\mathcal{A}_N(x) \mid x) = 1 - \mu_{\mathrm{pos}}(x)^N - (1 - \mu_{\mathrm{pos}}(x))^N$.

Let $Y_i^E = \mathbb{I}[o_i \in E_x]$, so $\Pr(Y_i^E = 1 \mid x) = \tau_E(x)$. We are interested in the event that the update is active yet the target subset $E_x$ receives no samples:

$$\mathcal{B}_{E,N}(x) := \mathcal{A}_N(x) \cap \Big\{ \sum_{i=1}^{N} Y_i^E = 0 \Big\}. \tag{8}$$

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

$$\Pr(\mathcal{B}_{E,N}(x) \mid x) = (1 - \tau)^N - (\mu_{\mathrm{pos}} - \tau)^N - (1 - \mu_{\mathrm{pos}})^N.$$

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

Lemma 3.1 reveals a non-monotonic dependence on $N$ after conditioning on $x$. Two competing effects determine $\Pr(\mathcal{B}_{E,N}(x) \mid x)$: the coverage factor $(1 - \tau_E(x))^N$ decreases with $N$, improving the chance of sampling $E_x$, while activity $\Pr(\mathcal{A}_N(x) \mid x)$ increases from near zero toward one. Their interaction induces the qualitative picture in Figure 1(a) and Appendix Figure 3: small $N$ can preserve higher pass@$k$ through inactivity (Wu et al., 2025b; Dang et al., 2025), large $N$ through

coverage (Hu et al., 2025), while intermediate $N$ can yield frequent active updates that miss low-mass correct regions, consistent with distribution sharpening observations (He et al., 2025). Because $\mu_{\text{pos}}(x)$ and $\tau_E(x)$ vary across prompts and evolve during optimization, the result is most naturally read as a regime descriptor: it characterizes how inactivity, undercoverage, and coverage trade off as $N$ changes. Section 5.4 later compares this qualitative picture with empirical group-size trends.

### 3.2 Unsampled-correct mass under finite sampling

The tail-miss analysis identifies *when* low-mass correct regions are vulnerable: active updates can occur while a target subset $E_x$ is absent from the sampled group. We now use the categorical framework (Section 2.3) to characterize the *mechanism* by which unsampled positive mass can decrease.

While (17) shows that total correct mass $Q_{\text{pos}}$ tends to increase with $N$ in the local categorical slice, it does not reveal redistribution within the correct set. Define the unsampled-correct mass

$$Q_{\text{u,pos}} := \sum_{i \in U \cap \mathcal{P}} p_i = Q_{\text{pos}} - P_{\text{pos}}. \tag{9}$$

This quantity measures the positive probability mass outside the distinct sampled positive set in the categorical slice.

**Proposition 3.2.** *Under the one-step surrogate update (5),*

$$\Delta Q_{\text{u,pos}} = \frac{\eta}{N}\Big[\underbrace{-S_R\, U_{\text{pos},2}}_{\textit{direct drift}} - Q_{\text{u,pos}} \underbrace{\big((R_c - S_R)A_2 + (R_w - S_R)B_2 - S_R U_2\big)}_{\textit{normalization coupling}}\Big]. \tag{10}$$

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

Equation (10) shows that $\Delta Q_{\text{u,pos}}$ can be negative even when $\Delta Q_{\text{pos}} > 0$: total positive mass can increase while unsampled positive mass decreases. 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 in the finite-action abstraction.

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 G) captures how probability gains by sampled-correct actions draw mass away from unsampled-correct ones through softmax normalization. In reward-positive batches, both terms can contribute negatively, so the update may reinforce the sampled correct subset while reducing probability on unsampled correct actions.

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 yields the intermediate-$N$ risk from Section 3.1: active updates that still miss low-probability correct regions and then pull mass toward the sampled correct subset. The categorical simulation in Section 5.1 measures this redistribution directly through retained positive mass.

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

The finite-sampling effects analyzed in Section 3 are prompt-local. For each prompt, rewards are assigned to sampled complete rollouts and group-relative coefficients are computed from the same finite group. Consequently, direct positive evidence enters the update only through correct rollouts that appear in the group, while unsampled correct behavior is affected only indirectly. F-GRPO follows this group granularity by multiplying the group-relative signal with a single group-dependent weight.

The rare correct subset $E_x$ and the unsampled-correct mass are latent during training. The available group-level statistic is the number of correct rollouts $X$, equivalently the empirical success rate $\widehat{\mu}_{\text{pos}}(x) = X/N$. We use this statistic as an observable summary of sampled positive exposure. A small nonzero value indicates that correct behavior has appeared but remains sparsely observed, whereas a large value indicates that the current update is supported by many positive rollouts. The resulting weighting attenuates high-success groups more strongly and attenuates sparse-positive mixed groups less.

### 4.1 Focal Weight

Define the empirical success rate for prompt $x$ as

$$\widehat{\mu}_{\text{pos}}(x) := \frac{\bar{R}(x) - R_w}{R_c - R_w} = \frac{X}{N} \in [0, 1], \tag{11}$$

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: $\mathbb{E}[\widehat{\mu}_{\text{pos}}(x)] = \mu_{\text{pos}}(x)$.

Under binary rewards, $\widehat{\mu}_{\text{pos}}(x) = X/N$ is a convenient observable summary of how much correct behavior the sampled group has already exposed for prompt $x$. It distinguishes groups where correct rollouts are still rare from groups where the sampled set already contains many correct rollouts. Appendix H further shows that the expected distinct sampled-correct mass $\mathbb{E}[P_{\text{pos}} \mid X = k]$ is non-decreasing in $k$, so larger $\widehat{\mu}_{\text{pos}}(x)$ corresponds in expectation to broader sampled-correct exposure; in the categorical picture, the expected $S_R$ moves in the same direction.

This makes sampled group success a convenient quantity to condition on. Inspired by Focal loss (Lin et al., 2017), we use a weight that decreases with empirical success to reduce the contribution of high-success groups. Define the Focal weight

$$g(x) := \big(1 - \widehat{\mu}_{\text{pos}}(x)\big)^{\gamma}, \quad \gamma \geq 0. \tag{12}$$

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 $\widehat{\mu}_{\text{pos}}(x) \to 1$.

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, attenuating high-success groups more strongly than mixed low-success ones. Appendix I visualizes this effect.

### 4.2 Integration with Group-Relative Methods

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

$$\widehat{A}_i^{\text{F-GRPO}} := g(x) \cdot \widehat{A}_i^{\text{GRPO}}. \tag{13}$$

The Focal weight is orthogonal to the clipping and importance-weighting choices in the base optimizer: it multiplies the group-level advantage without changing reward values, clipping bounds, or importance-weight construction. It can therefore be combined with objectives that differ in these details. We denote the corresponding 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; $\gamma$ is the only new hyperparameter.

## 5 Experiments & Results

We evaluate the paper's claims in four complementary settings. We first test the local redistribution mechanism in a closed categorical simulation where the relevant quantities are directly observable. We then use single-solution Maze to ask whether evaluation degradation can appear across prompts even when each prompt has only one correct trajectory, and compare the resulting picture with one representative GRPO group-size sweep in LLM training. Finally, we test whether the proposed fixed-$N$ reweighting transfers across GRPO-family methods and remains distinct from generic regularization or simple update-scale reduction.

### 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. This is the closed-world setting where the local quantities from Section 3.2 are directly observable, so it directly tests the categorical mechanism: whether total positive mass can grow while retained positive mass collapses. 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 $Q_{\text{pos}}$, we track $\mathcal{M}_{\text{ret}}(t)$, the *retained positive mass*, which measures the fraction of initial correct-action probability that remains at or above its starting value (Appendix J Eq. 23). Values near 1 indicate that initially positive actions have not lost probability mass; values near 0 indicate concentration onto a smaller subset of correct actions.

Figure 2 presents the results. Panel (a) confirms that $Q_{\text{pos}}$ increases for all group sizes, consistent with Hu et al. (2025). However, panel (b) shows that $\mathcal{M}_{\text{ret}}$ behaves non-monotonically: both small and large $N$ retain more positive mass, 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 $Q_{\text{pos}}$ grows slowly but positive mass is retained; **(II)** the concentration zone (shaded) where $Q_{\text{pos}}$ grows rapidly but $\mathcal{M}_{\text{ret}}$ collapses; and **(III)** large $N$ where both metrics are high. No-

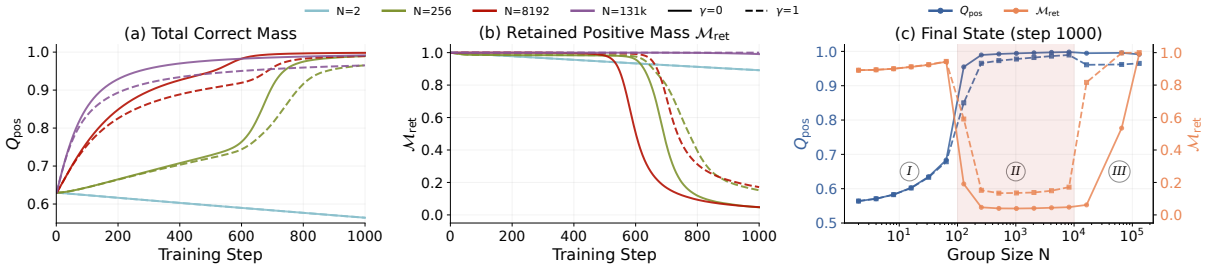


Figure 2: Categorical policy simulation following Hu et al. (2025) setup. **(a)** Total correct mass $Q_{\text{pos}}$ vs. training step. **(b)** Retained positive mass $\mathcal{M}_{\text{ret}}$ vs. step. **(c)** Final metrics vs. group size $N$, with three regimes: **I** slow $Q_{\text{pos}}$ growth, positive mass retained; **II** concentration zone (shaded), $Q_{\text{pos}}$ grows but $\mathcal{M}_{\text{ret}}$ collapses; **III** both metrics high. Solid: $\gamma$=0; dashed: $\gamma$=1. $N$=131k maintains $\mathcal{M}_{\text{ret}}\approx 1$ throughout, consistent with $\Pr(\mathcal{B}_{E,N}(x) \mid x) < 10^{-3}$ for a non-anchor correct action under the initial distribution (Appendix J).

tably, $N$=131,072 maintains $\mathcal{M}_{\text{ret}} \approx 1$ throughout training, consistent with Lemma 3.1, which predicts $\Pr(\mathcal{B}_{E,N}(x) \mid x) < 10^{-3}$ for a non-anchor correct action under the initial distribution (see Appendix J). Dashed lines ($\gamma$=1) show improved $\mathcal{M}_{\text{ret}}$ retention, particularly in the concentration zone.

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 yield the lowest retained positive mass.

### 5.2 Single-Solution Maze Experiments

Maze provides a single-solution setting from concurrent work by Tajwar et al. (2026), where each prompt has exactly one correct action sequence. We compare GRPO and F-GRPO with $\gamma \in \{0.5, 1.0\}$ across $N \in \{4, 8, 16, 32, 64, 128\}$ after SFT initialization; full details are in Appendix K.

Appendix Figure 6 shows that F-GRPO improves final pass@1 and pass@$K$, and reduces the early training-time drop in pass@$K$. At final evaluation, F-GRPO with $\gamma$=0.5 improves pass@1 for every $N$ (74.4–93.6 vs. 65.6–75.8 for GRPO) and pass@256 (75.9–96.5 vs. 67.3–76.9). Because each prompt has a unique target sequence, the observed large-$K$ degradation reflects reduced target-sequence coverage at the evaluation-distribution level rather than loss of alternative correct trajectories within a prompt. Thus Maze separates the pass@$K$ effect from within-prompt multi-solution diversity and shows that the same weighting also mitigates degradation in a single-solution setting.

### 5.3 LLM Experimental Setup

We evaluate on Qwen2.5-7B (Yang et al., 2024), Qwen3-4B-Base (Yang et al., 2025), and Llama-3.2-3B-Instruct (Grattafiori et al., 2024), trained on DeepScaleR (Luo et al., 2025) using verl (Sheng et al., 2024). We report pass@1 and pass@256 on in-domain math benchmarks (MATH500 (Hendrycks et al., 2021), AIME24/25 (Art of Problem Solving, 2024a), AMC23 (Art of Problem Solving, 2024b), Minerva Math (Lewkowycz et al., 2022), Olympiad Bench (He et al., 2024)) and OOD benchmarks (GPQA Diamond (Rein et al., 2023), IFEval (Zhou et al., 2023), SynLogic (Liu et al., 2025a)). Full training, evaluation, and $\gamma$-selection details are in Appendix L; additional pass@$k$ tables and training dynamics are in Appendices O and P.

### 5.4 Group Size Regimes and Focal Weighting

Having validated the local categorical mechanism and used Maze as a single-solution setting, we next examine empirical group-size trends in LLM training. Table 2 compares GRPO at $N \in \{2, 4, 8, 16, 32\}$ with F-GRPO at $N = 8$ on Qwen2.5-7B. These group sizes are chosen to probe small, intermediate, and larger rollout regimes while keeping rollout cost tractable; we do not aim to exhaustively map performance as a function of $N$.

GRPO exhibits non-monotonic pass@256 behavior across group sizes: $N$=2 yields high pass@256 but lowest pass@1, a pattern consistent with slow policy change under infrequent active updates. Moving to intermediate group sizes improves math pass@1 but reduces pass@256; the degradation is strongest around $N$=8 on in-domain benchmarks and spans $N$=8-16 on OOD benchmarks.

| 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** |
| ***Qwen3-4B-Base*** | | | | | | | | | | | |
| GRPO | 39.9/71.1 | 14.8/52.5 | 8.3/45.6 | 54.6/96.2 | 79.5/96.4 | **39.2**/61.7 | 43.1/74.2 | 22.2/67.9 | 36.9/91.2 | 10.7/57.1 | 19.0/55.4 |
| F-GRPO | **42.7**/**76.0** | **17.4**/**61.1** | **15.8**/**60.0** | **58.3**/**98.2** | **82.3**/**97.4** | 38.5/**63.6** | **43.7**/**75.5** | **24.3**/**73.6** | **41.2**/**93.6** | 10.7/**64.5** | **21.1**/**62.8** |
| DAPO | 45.1/76.1 | 18.6/**63.8** | 17.6/54.0 | 62.0/96.2 | 84.4/**97.8** | **40.0**/67.1 | 47.8/77.9 | 20.7/74.4 | 33.8/88.0 | **11.8**/71.3 | 16.4/63.9 |
| F-DAPO | **45.3**/**76.4** | **19.1**/59.7 | **17.7**/**58.3** | **63.0**/**96.8** | 84.4/97.7 | 39.5/**67.9** | **48.2**/**78.0** | **23.2**/**76.8** | **39.6**/**90.6** | 11.3/**72.4** | **18.7**/**67.5** |
| CISPO | 45.8/75.3 | 21.5/59.3 | 17.6/55.4 | **63.1**/97.4 | 84.2/97.9 | 40.2/65.1 | 48.4/76.6 | 22.6/76.5 | 32.6/85.9 | **11.6**/**71.1** | 23.6/72.4 |
| F-CISPO | **46.0**/**79.2** | 21.5/**69.5** | **18.6**/**62.7** | 60.7/**98.6** | 84.2/**98.2** | **42.1**/**67.2** | **48.9**/**78.8** | **23.9**/**80.0** | **35.9**/**90.3** | 11.1/68.0 | **24.6**/**81.6** |
| ***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 across the reported method-model pairs; pass@1 is generally comparable, with some configurations showing modest trade-offs. **Bold**: better within baseline/Focal pair; underline: statistically significant ($p$<0.05, see Appendix Q).

At $N$=32, pass@256 partially recovers while math pass@1 continues to improve. This pattern is qualitatively consistent with an intermediate concentration regime between small-$N$ inactivity and larger-$N$ coverage. The local theory does not predict the exact boundaries for LLM training, but it does motivate this qualitative comparison.

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× fewer rollouts. Pass@1 shows a modest trade-off on in-domain benchmarks but improves on OOD tasks, suggesting that Focal weighting can recover much of the pass@256 loss observed under GRPO at intermediate $N$ without increasing the rollout budget in this setting.

| Method | Avg. Math ↑ | Avg. OOD ↑ | $\Delta\text{NLL}_{\text{rare}}$ ↓ |
|---|---|---|---|
| GRPO $N$=2 | 36.2 / **75.0** | 18.0 / **67.3** | **0.19** |
| GRPO $N$=4 | 36.4 / 71.1 | 18.7 / 59.7 | 0.44 |
| GRPO $N$=8 | 37.3 / 64.1 | 17.1 / 55.9 | 0.68 |
| GRPO $N$=16 | 38.4 / 67.5 | 17.7 / 55.7 | 0.66 |
| 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. $\Delta\text{NLL}_{\text{rare}}$: diagnostic increase in negative log-likelihood on a fixed set of correct trajectories that were rare under the base model by NLL (lower = less deviation from base distribution; see Appendix M.2). **Bold**: best. Gray rows mark the same group size $N$=8 comparison. Full per-benchmark results in Appendix M.

**Deviation from Base-Model Rare Solutions.** We also report $\Delta\text{NLL}_{\text{rare}}$, computed on correct trajectories that were rare under the base model by NLL (Appendix M.2). Higher values indicate greater deviation from the base distribution on these trajectories. The ordering $\Delta\text{NLL}_{\text{rare}}(N{=}2) < \Delta\text{NLL}_{\text{rare}}(N{=}32) < \Delta\text{NLL}_{\text{rare}}(N{=}8)$ is qualitatively consistent with the pass@256 degradation pattern, with F-GRPO at $N$=8 achieving an intermediate value (0.46), indicating less deviation on these base-model-rare correct trajectories than its baseline.

### 5.5 Focal Weighting Across Methods

We now ask a more practical question: whether the same fixed-$N$ reweighting transfers across GRPO-family optimizers. Table 1 reports results for GRPO, DAPO, and CISPO at $N$=8, a commonly used group size (Shao et al., 2024; Zeng et al., 2025; Liu et al., 2025d), across three models at different scales. On Qwen2.5-7B, pass@256 gains are +6.2 (GRPO), +3.2 (DAPO), and +3.6 (CISPO); corresponding gains are +4.9/+0.3/+3.9 on Qwen3-4B-Base and +3.5/+8.1/+1.7 on Llama-3.2-3B-

Instruct. Across all nine method-model combinations, Focal weighting improves both math and OOD pass@256 (average +3.9 and +4.1). Math pass@1 is preserved or improved in 9/9 cases, with gains up to +2.8, and OOD pass@1 improves in 7/9 cases (average +1.5).

### 5.6 Comparison with Regularization and Update-Scale Controls

We compare F-GRPO against GRPO with entropy bonus (GRPO-$\mathcal{H}$), GRPO with KL penalty (GRPO-KL), GRPO with a lower learning rate (GRPO low-LR), and differential smoothing (DS-GRPO; Gai et al., 2025) using the Qwen2.5-7B setup. GRPO low-LR checks whether the average scale reduction alone is sufficient by matching the average advantage-magnitude scale induced by Focal weighting. For $\gamma = 0.5$, this gives $\mathrm{lr}_{\mathrm{low}} = 6.8\times10^{-7}$; coefficient tuning and the multiplier derivation are in Appendices L and N.

| Method | Avg. Math ↑ | Avg. OOD ↑ |
|---|---|---|
| GRPO low-LR | 37.8$^{\dagger}$/ 69.2 | 16.4 / 57.9 |
| GRPO-$\mathcal{H}$ | 37.8$^{\dagger}$/ 69.5 | 18.7 / 59.9 |
| GRPO-KL | 37.2 / 72.0$^{\dagger}$ | **19.4** / 60.0 |
| DS-GRPO | 37.7 / **73.8** | 17.9 / **68.3** |
| F-GRPO | **38.6** / 70.3 | 19.2$^{\dagger}$/ 63.3$^{\dagger}$ |

Table 3: Qwen2.5-7B comparison at $N$=8. Metrics are average pass@1 / pass@256. **Bold**: best; †: second best.

F-GRPO achieves the highest math pass@1 (38.6) and the second-highest OOD pass@1 (19.2, behind GRPO-KL at 19.4, which incurs reference-model overhead). DS-GRPO has higher pass@256 than F-GRPO (73.8 vs. 70.3 on math; 68.3 vs. 63.3 on OOD), but lower pass@1 (37.7 vs. 38.6 on math; 17.9 vs. 19.2 on OOD). GRPO low-LR improves pass@256 over default GRPO but remains below F-GRPO, suggesting that average scale reduction alone does not fully account for the gains. In this setup, the controls suggest that the gains are not explained solely by generic regularization or simple step-size reduction.

## 6 Related Work

**Distribution Sharpening and Group Size.** A growing body of work documents that RLVR improves pass@1 while degrading pass@$k$ for large $k$, indicating concentration onto fewer solutions (Dang et al., 2025; Yue et al., 2025; Wu et al., 2025a). The optimal rollout count remains debated: Wu et al. (2025b) show that $N$=2 is theoretically justified and compute-efficient, while Hu et al. (2025) advocate large groups for coverage. We provide a complementary perspective: finite-sampling effects in group-relative methods, including rare-mode undercoverage within prompts and success-based reweighting during training.

**Difficulty-Aware, Entropy, and Token-level Approaches.** Related approaches include difficulty-aware reweighting (Lin et al., 2017; Bengio et al., 2009; Parashar et al., 2025; Zhou et al., 2025; Yang et al., 2026), rare-trajectory rewards (He et al., 2025), differential smoothing (Gai et al., 2025), entropy interventions (Cui et al., 2025; Cheng et al., 2025; Agarwal et al., 2025), and token-level concentration methods (Hao et al., 2026; Peng et al., 2025; Wang et al., 2025). Due to page-limit constraints, these approaches are discussed in Appendix A. Our focus here is on finite-group sampling in group-relative RLVR, its dependence on rollout count $N$, and success-based weighting of whole sampled groups.

## 7 Conclusion

This work studies finite-group RLVR at the prompt-local level where group-relative updates are actually formed. Our theoretical analysis derives a closed-form tail-miss probability exhibiting non-monotonic dependence on $N$: small groups limit active misses through inactivity, large groups through coverage, but intermediate $N$ maximize the probability of active misses. In the categorical abstraction, we further characterize redistribution within the correct set, showing that unsampled-correct mass can shrink even as total correct mass grows. Empirically, categorical simulation illustrates this mechanism in the categorical setting; a single-solution Maze setting shows that evaluation pass@$K$ can degrade across prompts even when each prompt has one correct trajectory; and a representative GRPO LLM sweep is qualitatively consistent with the same regime picture. Motivated by this analysis, Focal weighting provides a lightweight fixed-$N$ mitigation that improves pass@256 across GRPO, DAPO, and CISPO while preserving or improving pass@1 in most reported settings, with especially consistent OOD gains and no extra rollout cost.

## 8 Limitations

Our theoretical analysis isolates one finite-sampling mechanism in group-relative RLVR with binary outcome rewards, namely that active groups can miss low-mass correct regions and then reallocate probability toward the sampled correct subset. The resulting claim is local, and the categorical simulation tests that local redistribution mechanism directly in a controlled setting. In this framing, the Maze experiments ask whether pass@$K$ degradation can arise even when each prompt has a single correct trajectory, and the LLM group-size sweep asks only whether qualitatively similar trends appear in practical training. F-GRPO is evaluated on that basis as a fixed-$N$ mitigation motivated by this mechanism, not as a general rule for RLVR concentration or group-size selection.

## References


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

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

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

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48. ACM.

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

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

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, and 39 others. 2021. Evaluating large language models trained on code.

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*.

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

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. 2025. The entropy mechanism of reinforcement learning for reasoning language models. *Preprint*, arXiv:2505.22617.

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

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

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

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

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. 2026. Rethinking entropy interventions in rlvr: An entropy change perspective. *arXiv preprint arXiv:2510.10150v2*, arXiv:2510.10150.

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

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

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

Hugging Face. 2026. Math-verify: A robust mathematical expression evaluation system. https://github.com/huggingface/Math-Verify. 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. 2023. Swe-bench: Can language models resolve real-world github issues? *arXiv preprint arXiv:2310.06770*.

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

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

Jia Li, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Huang, Kashif Rasul, Longhui Yu, Albert Q Jiang, Ziju Shen, and 1 others. 2024. Numinamath: The largest public dataset in ai4maths with 860k pairs of competition math problems and solutions. *Hugging Face repository*, 13(9):9.

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

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. 2025a. Synlogic: Synthesizing verifiable reasoning data at scale for learning logical reasoning and beyond. *arXiv preprint arXiv: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. 2025b. Prorl: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*.

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

Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, and 1 others. 2025d. Part i: Tricks or traps? a deep dive into rl for llm reasoning. *arXiv preprint arXiv:2508.08221*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*. Poster.

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. 2025. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion Blog.

Kohsei Matsutani, Shota Takashiro, Gouki Minegishi, Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. 2025. Rl squeezes, sft expands: A comparative study of reasoning llms. *Preprint*, arXiv:2509.21128.

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

Shubham Parashar, Shurui Gui, Xiner Li, Hongyi Ling, Sushil Vemuri, Blake Olson, Eric Li, Yu Zhang, James Caverlee, Dileep Kalathil, and 1 others. 2025. Curriculum reinforcement learning from easy to hard tasks improves llm reasoning. *arXiv preprint arXiv:2506.06632*.

Ruotian Peng, Yi Ren, Zhouliang Yu, Weiyang Liu, and Yandong Wen. 2025. Simko: Simple pass@k policy optimization. *Preprint*, arXiv:2510.14807. ArXiv:2510.14807v2, last revised 21 Oct 2025.

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

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

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*. 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, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

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

Fahim Tajwar, Guanning Zeng, Yueer Zhou, Yuda Song, Daman Arora, Yiding Jiang, Jeff Schneider, Ruslan Salakhutdinov, Haiwen Feng, and Andrea Zanette. 2026. Maximum likelihood reinforcement learning. *arXiv preprint arXiv:2602.02710*.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

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

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

Yihong Wu, Liheng Ma, Lei Ding, Muzhi Li, Xinyu Wang, Kejia Chen, Zhan Su, Zhanguang Zhang, Chenyang Huang, Yingxue Zhang, and 1 others. 2025b. It takes two: Your grpo is secretly dpo. *arXiv preprint arXiv:2510.00977*.

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

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, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, Jiajun Chai, Guojun Yin, Wei Lin, Shuai Ma, Fuzhen Zhuang, Deqing Wang, and 1 others. 2026. Your group-relative advantage is biased. *arXiv preprint arXiv:2601.08521*.

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*.

Lifan Yuan, Weize Chen, Yuchen Zhang, Ganqu Cui, Hanbin Wang, Ziming You, Ning Ding, Zhiyuan Liu, Maosong Sun, and Hao Peng. 2025. 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*.

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*.

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

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*.

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. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *Proceedings of the VLDB Endowment*, 16(12):3848–3860.

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. 2023. Sglang: Efficient execution of structured language model programs. *arXiv preprint arXiv: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. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*. Submitted on 14 Nov 2023.

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

## A Additional Related Work

**Distribution Sharpening in RLVR.** A growing body of work documents that RLVR improves pass@1 while degrading pass@$k$ for large $k$, indicating concentration onto fewer solutions (Dang et al., 2025; Yue et al., 2025; Wu et al., 2025a). Chen et al. (2025b) attribute this to overconfidence induced by cross-entropy training and propose confidence limiting. We provide a complementary perspective: finite-sampling effects in group-relative methods, including rare-mode undercoverage within prompts and success-based reweighting during training.

**Group Size and Sampling Dynamics.** The optimal rollout count remains debated. Wu et al. (2025b) show that $N{=}2$ is theoretically justified and compute-efficient, while Hu et al. (2025) advocate large groups for coverage, showing that scaling $N$ ensures non-negative change in total correct mass. We connect these views by analyzing finite-group sampling as a local bias: small and large $N$ preserve coverage through inactivity and coverage respectively, while intermediate $N$, common in practice, can maximize active updates that miss rare-correct regions and concentrate mass on sampled positives. Concurrently (Tajwar et al., 2026) derives likelihood-based estimators for correctness-based RL; our Focal weight is a plug-in group-level scaling for group-relative estimators.

**Difficulty-Aware Training.** Reweighting by difficulty has established 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) dynamically rebalances loss contributions across difficulty groups to equalize loss scale. He et al. (2025) identify rank bias in GRPO and propose unlikeliness reward to up-weight rare correct trajectories. Concurrently, Yang et al. (2026) analyze bias in group-relative advantages across prompt difficulty and propose history-aware adaptive difficulty weighting. Gai et al. (2025) analyze selection and reinforcement bias and propose differential smoothing that modifies reward for correct and incorrect trajectories. Our focus here is on finite-group sampling in group-relative RLVR, its dependence on rollout count $N$, and success-based weighting of whole sampled groups.

**Entropy and Token-level Approaches.** The role of entropy in RLVR remains debated, with some advocating maximization for exploration (Cui et al., 2025; Cheng et al., 2025) and others reporting benefits from minimization (Agarwal et al., 2025). Several methods address token-level concentration by reweighting tokens based on entropy dynamics or probability structure (Hao et al., 2026; Peng et al., 2025; Wang et al., 2025). These approaches regulate how probability mass is distributed *within* trajectories; our Focal weighting instead regulates *which prompts* contribute most strongly to group-relative updates.

## B Group-Relative Objective Details

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 $r_{i,t}(\theta) = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$. The GRPO objective is

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_x\Big[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i} L_{i,t}^{\text{clip}} - \beta\,\mathbb{D}_{\text{KL}}\left(\pi_\theta \,\|\, \pi_{\text{ref}}\right)\Big] \tag{14}$$

where $L_{i,t}^{\text{clip}} = \min\big(r_{i,t}\widehat{A}_i,\ \text{clip}(r_{i,t}, 1{-}\varepsilon, 1{+}\varepsilon)\widehat{A}_i\big)$. We set $\beta = 0$ following DAPO (Yu et al., 2025). DAPO modifies this with asymmetric clipping bounds $\text{clip}(r_{i,t}, 1{-}\varepsilon_{\text{low}}, 1{+}\varepsilon_{\text{high}})$ where $\varepsilon_{\text{high}} > \varepsilon_{\text{low}}$, 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

$$\widehat{r}_{i,t} = \text{clip}\big(r_{i,t}, 1{-}\varepsilon_{\text{low}}^{\text{IS}}, 1{+}\varepsilon_{\text{high}}^{\text{IS}}\big), \tag{15}$$

and optimizes a REINFORCE-style objective

$$\mathcal{L}_{\text{CISPO}}(\theta) = \mathbb{E}_{i,t}\Big[\text{sg}(\widehat{r}_{i,t})\,\widehat{A}_i^{\text{GRPO}} \cdot \log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\Big] \tag{16}$$

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

## C Categorical Policy Total-Mass Expression

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

$$\Delta Q_{\text{pos}} = \frac{\eta}{N}\Big[(R_c - S_R)Q_{\text{neg}}A_2 + (S_R - R_w)Q_{\text{pos}}B_2 + S_R\big(Q_{\text{pos}}U_{\text{neg},2} - Q_{\text{neg}}U_{\text{pos},2}\big)\Big]. \tag{17}$$

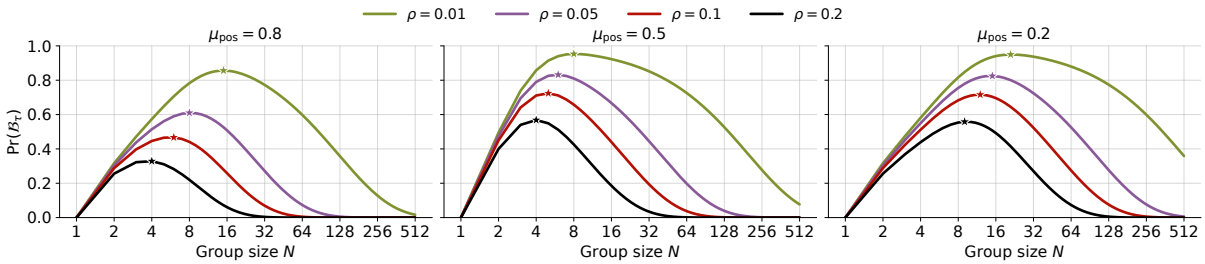


Figure 3: Conditional tail-miss probability $\Pr(\mathcal{B}_{E,N}(x) \mid x)$ from Lemma 3.1 versus group size $N$ for rare-mode undercoverage. Each panel fixes $\mu_{\text{pos}}(x) \in \{0.8, 0.5, 0.2\}$; curves vary $\rho = \tau_E(x)/\mu_{\text{pos}}(x)$, the fraction of correct mass in the rare-correct region. Stars mark peaks. For all parameter combinations, the conditional probability 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 $\rho$ shifts the peak rightward and upward.

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

## D Proof of Lemma 3.1

*Proof.* Fix a prompt $x$ and a target subset $E_x \subseteq \mathcal{C}(x)$. Omit $(x)$ for readability and write $\mathcal{C}$ for $\mathcal{C}(x)$ and $E$ for $E_x$. Each rollout falls into one of three disjoint regions: the target correct subset $E$ with probability $\tau$, the remaining correct region $\mathcal{C} \setminus E$ 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 $E$ is $(1-\tau)^N$. Conditioned on this event, all rollouts lie in $(\mathcal{C} \setminus E) \cup (\Omega \setminus \mathcal{C})$. The group is inactive (hence $\mathcal{B}_{E,N}(x)$ does not occur) in two disjoint cases: all rollouts are correct but not in $E$, with probability $(\mu_{\text{pos}} - \tau)^N$; or all rollouts are incorrect, with probability $(1 - \mu_{\text{pos}})^N$. Thus

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

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

$$\Delta p_i = \sum_{j \in \mathcal{A}} \frac{\partial p_i}{\partial z_j} \Delta z_j = p_i \Big( \Delta z_i - \sum_{j \in \mathcal{A}} p_j \Delta z_j \Big). \tag{18}$$

For any subset $\mathcal{S} \subseteq \mathcal{A}$, define its probability mass $Q_{\mathcal{S}} := \sum_{i \in \mathcal{S}} p_i$. Summing (18) over $i \in \mathcal{S}$ yields the *subset-mass identity*:

$$\Delta Q_{\mathcal{S}} := \sum_{i \in \mathcal{S}} \Delta p_i = \sum_{i \in \mathcal{S}} p_i \Delta z_i - Q_{\mathcal{S}} \sum_{j \in \mathcal{A}} p_j \Delta z_j. \tag{19}$$

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 $\Delta Q_{\text{pos}}$.** Setting $\mathcal{S} = \mathcal{P}$ and using the one-step update (5) recovers the mass balance equation (17) 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 (5).

## F Proof of Proposition 3.2

*Proof.* Apply the subset-mass identity (Appendix E, Eq. (19)) with $\mathcal{S} = U \cap \mathcal{P}$:

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

For $i \in U \cap \mathcal{P}$, we have $R_i = 0$, so by (5), $\Delta z_i = -\frac{\eta}{N} S_R p_i$. Thus the first sum becomes

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

For the normalization term, partitioning by reward value:

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

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 (10). □

## G Detailed Term Analysis for Proposition 3.2

We analyze each term in (10) to understand when unsampled-correct mass decreases.

**Direct drift term.** The term $-S_R\, U_{\text{pos},2}$ 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 $U_{\text{pos},2}$, the concentration of unsampled-correct probability.

**Normalization coupling.** The second term couples $Q_{\text{u,pos}}$ 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 $\Delta Q_{\text{u,pos}} < 0$ while $\Delta Q_{\text{pos}} > 0$?** Consider a reward-positive batch ($S_R > 0$) on a prompt with high success probability. In this regime:

- The direct drift $-S_R U_{\text{pos},2} < 0$ 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, $\Delta Q_{\text{pos}}$ from (17) 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.

## H Monotonicity of Sampled Distinct Mass Conditioned on $X$

This appendix formalizes the monotonicity claim used in Section 4 (Focal Weight): as the observed correct count $X$ increases, the expected distinct sampled-correct mass is non-decreasing. As a corollary, the corresponding categorical baseline $S_R$ is also monotone in $X$.

**Setup.** Fix a prompt $x$ and write $\pi(o) := \pi_\theta(o \mid x)$ for brevity. Let $\Omega_x$ 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

$$S_R := R_c\, P_{\text{pos}} + R_w\, P_{\text{neg}}. \tag{20}$$

**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 $q_{\text{pos}}$ over $\mathcal{C}$, and the $N - k$ incorrect rollouts are i.i.d. from $q_{\text{neg}}$ over $\Omega_x \setminus \mathcal{C}$.

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

$$\mathbb{E}[P_{\text{pos}} \mid X = k] = \sum_{o \in \mathcal{C}} \pi(o) \times \Big(1 - (1 - q_{\text{pos}}(o))^k\Big), \tag{21}$$

$$\mathbb{E}[P_{\text{neg}} \mid X = k] = \sum_{o \notin \mathcal{C}} \pi(o) \times \Big(1 - (1 - q_{\text{neg}}(o))^{N-k}\Big). \tag{22}$$

*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 $P_{\text{pos}}$; the argument for $P_{\text{neg}}$ 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 $q_{\text{pos}}$. 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 (21).

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 $P_{\text{neg}}$, 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 (22). 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 H.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 (20) 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 H.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 $k$. □

## I Focal Weight Visualization

Figure 4 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, attenuating high-success groups more strongly than mixed low-success ones.

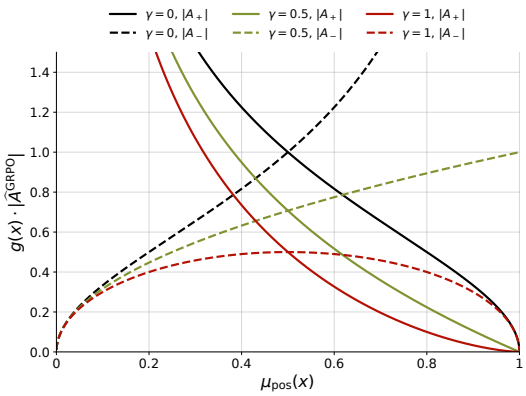


Figure 4: Scaled advantage magnitude $g(x) \cdot |\widehat{A}^{\text{GRPO}}|$ versus success probability $\mu_{\text{pos}}(x)$ for binary rewards. Solid lines: correct rollouts; dashed lines: incorrect rollouts. Higher $\gamma$ attenuates high empirical-success groups more strongly, thereby changing the relative weighting of active updates.

## 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 temperature

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 conditional tail-miss probability from Lemma 3.1 for a typical non-anchor correct action with $\tau_E = \tau_{\text{leaf}} \approx 6.3 \times 10^{-5}$. Figure 5 shows $\Pr(\mathcal{B}_{E,N}(x) \mid x)$ 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 $N = 2^{17} = 131{,}072$, $\Pr(\mathcal{B}_{E,N}(x) \mid x) < 10^{-3}$, predicting that such a group size should preserve probability mass on non-anchor correct actions. This prediction aligns with the simulation results in Figure 2: $N$=131,072 is the only configuration that maintains $\mathcal{M}_{\text{ret}} \approx 1$ throughout training.

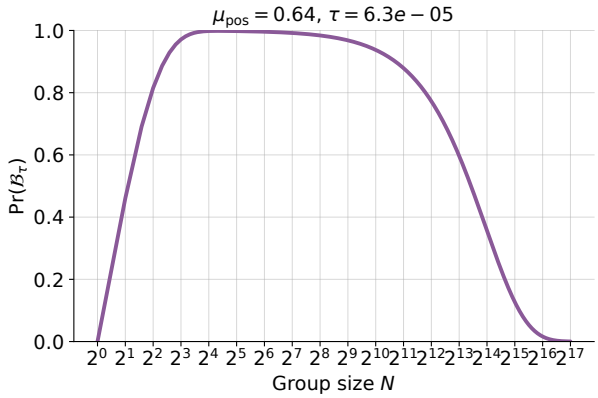


Figure 5: Pointwise tail-miss probability $\Pr(\mathcal{B}_{E,N}(x) \mid x)$ versus group size $N$ for $\mu_{\text{pos}} = 0.64$ and $\tau_E = 6.3 \times 10^{-5}$, corresponding to a non-anchor correct action in the simulation. The non-monotonic shape explains the concentration zone: intermediate $N$ maximizes the probability that a correct action is unsampled while the batch contains mixed rewards.

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, 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 $Q_{\text{pos}}(t) = \sum_{a \in \mathcal{A}^+} \pi_t(a)$ and retained positive mass:

$$\mathcal{M}_{\text{ret}}(t) = 1 - \frac{\sum_{a \in \mathcal{A}^+} \max\big(0,\, \pi_0(a) - \pi_t(a)\big)}{\sum_{a \in \mathcal{A}^+} \pi_0(a)}. \tag{23}$$

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

## K Maze Experiment Details

We follow the maze-navigation setup from concurrent work (Tajwar et al., 2026). The environment uses 1M procedurally generated 17×17 training mazes and 256 held-out mazes. Mazes are represented with symbolic grid tokens, and the model autoregressively emits navigation actions ending with a termination token. We use the reported lightweight decoder-only Transformer configuration, with approximately 3M parameters.

Training follows the concurrent group-size sweep. We first run SFT from scratch on provided target trajectories for 1,500 steps with batch size 32, AdamW (Loshchilov and Hutter, 2019), and learning rate $5\times10^{-4}$, which initializes the output format and yields nonzero pass rate. We then run fully on-policy RL with one parameter update per RL step, data batch size 256, learning rate $1\times10^{-4}$, and rollout number $N \in \{4, 8, 16, 32, 64, 128\}$. Final Maze results are reported after 20K RL steps. Each prompt is evaluated against a single target action sequence, so pass@$K$ estimates the probability that at least one of $K$ samples reaches that sequence.

Figure 6 shows that F-GRPO improves final pass@1 and pass@$K$, and reduces the early training-time drop in pass@$K$. At final evaluation, F-GRPO with $\gamma$=0.5 improves pass@1 for every $N$ (74.4–93.6 vs. 65.6–75.8 for GRPO) and reduces $-\log(\text{pass@}K)$ across larger $K$; exact final pass@1/128/256 values are in Table 4, and the training-dynamics figure additionally reports Test pass@2048. Training dynamics in Figure 7 show that GRPO often exhibits an early evaluation pass@$K$ drop followed by recovery, while F-GRPO reduces this degradation. Because each prompt has a unique target sequence, that early large-$K$ drop cannot be attributed to redistribution among multiple correct trajectories within the same prompt. The Maze results therefore indicate that

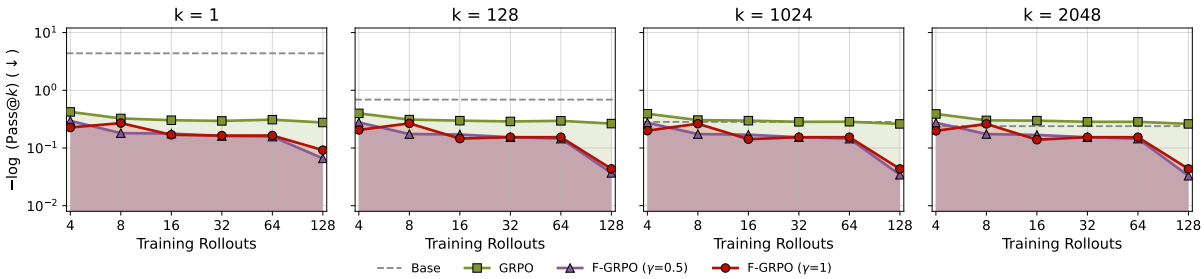


Figure 6: Single-solution Maze evaluation across group sizes. Each maze prompt has one target action sequence. The figure reports $-\log(\text{pass@}K)$ for $K \in \{1, 128, 1024, 2048\}$; lower is better. F-GRPO improves target-sequence coverage across group sizes.

| $N$ | pass@1 | | | pass@128 | | | pass@256 | | |
|---|---|---|---|---|---|---|---|---|---|
| | **GRPO** | $\gamma$=0.5 | $\gamma$=1.0 | **GRPO** | $\gamma$=0.5 | $\gamma$=1.0 | **GRPO** | $\gamma$=0.5 | $\gamma$=1.0 |
| 4 | 65.6 | 74.4 | **79.7** | 67.1 | 75.8 | **81.4** | 67.3 | 75.9 | **81.5** |
| 8 | 72.2 | **83.4** | 76.3 | 73.3 | **84.0** | 76.6 | 73.4 | **84.0** | 76.6 |
| 16 | 73.9 | 83.7 | **84.4** | 74.2 | 84.3 | **86.4** | 74.2 | 84.3 | **86.5** |
| 32 | 74.5 | **85.1** | 84.9 | 74.9 | **85.8** | **85.8** | 75.1 | **85.8** | **85.8** |
| 64 | 73.4 | **85.4** | 84.8 | 74.4 | **86.5** | 85.8 | 74.7 | **86.5** | 85.8 |
| 128 | 75.8 | **93.6** | 91.2 | 76.7 | **96.4** | 95.7 | 76.9 | **96.5** | 95.7 |

Table 4: Final Maze evaluation metrics after the 20K-step run. Values are percentages on the held-out maze set. $\gamma$ columns denote F-GRPO. **Bold**: best among GRPO, F-GRPO $\gamma$=0.5, and F-GRPO $\gamma$=1.0 for the same $N$ and pass@$K$.

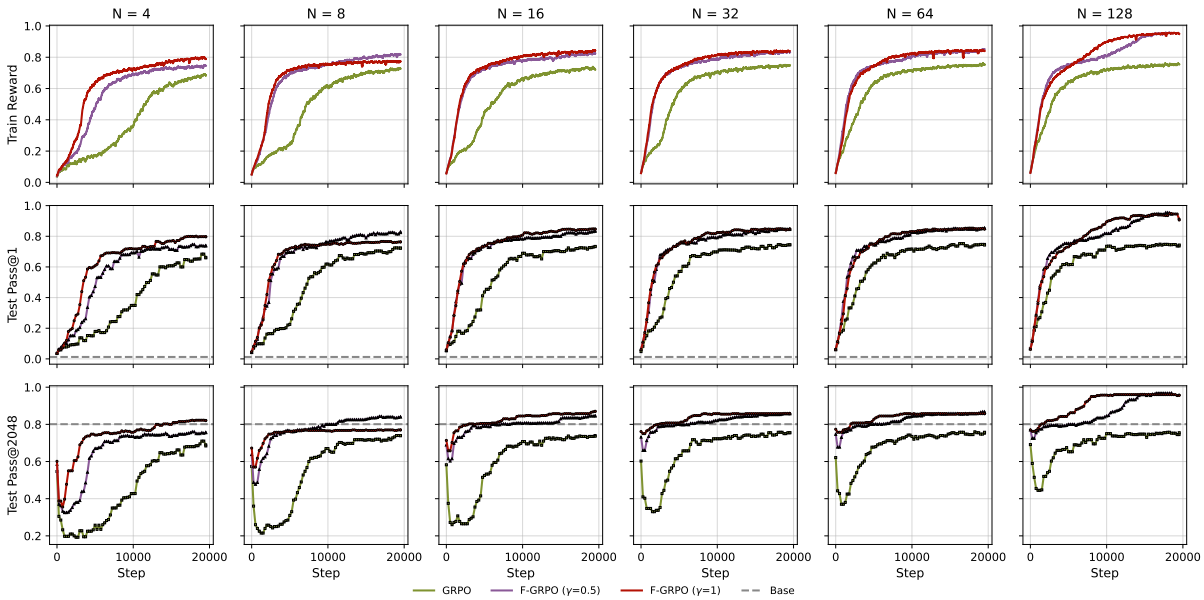


Figure 7: Training dynamics on single-solution Maze. GRPO often shows an early drop in evaluation pass@$K$, indicating reduced target-sequence coverage, followed by later recovery. F-GRPO reduces the early degradation across group sizes while improving convergence. Large-$K$ pass@$K$ is used here as a diagnostic for target-sequence coverage.

the observed degradation is not solely a within-prompt multi-solution phenomenon.

Figure 7 reports training reward, pass@1, and diagnostic large-$K$ pass@$K$ across group sizes. The transient evaluation drop under GRPO is consistent with the possibility that finite-group updates selected from training prompts can reduce target-sequence coverage. This dynamics-based evidence is complementary to the final pass@$K$ versus group-size summary in Figure 6.

# L Experimental Details

## L.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. For Qwen3-4B-Base, we additionally enable the model's thinking mode during training.

## L.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 4096 tokens for Qwen3-4B-Base and 8192 tokens for the other models. Following (Yu et al., 2025), we drop the KL-divergence regularization term and use token-mean loss aggregation. Unless otherwise stated, all experiments use learning rate $1 \times 10^{-6}$ 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.

| Parameter | Value |
|---|---|
| Optimizer | AdamW |
| $(\beta_1, \beta_2)$ | (0.9, 0.999) |
| Weight decay | 0.01 |
| Gradient norm clipping | 1.0 |
| Learning rate | $1 \times 10^{-6}$ |
| 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.

**Entropy and KL Regularization:** for the regularization controls in Section 5.6, 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.

## L.3 Evaluation Protocol

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

$$\text{pass@}k := \mathbb{E}_{\text{Problems}}\left[1 - \frac{\binom{n-c}{k}}{\binom{n}{k}}\right], \qquad (24)$$

where $n$ is the total number of samples and $c$ is the number of correct samples.

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 evaluation prompt settings are in Tables 6 and 7.

| Parameter | Qwen 2.5-7B | Qwen3 4B-Base | 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 | 4096 | 8192 |

Table 6: Evaluation configurations.

## L.4 Focal Weight Hyperparameter $\gamma$

Following the model-selection protocol above, we sweep the Focal exponent $\gamma \in \{0.5, 1.0, 2.0\}$ for each Focal-weighted method (F-GRPO, F-DAPO, F-CISPO). For reproducibility, the selected $\gamma$ values for the setups reported in Table 1 are summarized in Table 8. Table 9 therefore makes the resulting pass@1 / pass@256 trade-off explicit, with $\gamma = 0$ denoting the corresponding unweighted baseline. Across model-method combinations, the selected values in Table 8 are always $\gamma = 0.5$ or $\gamma = 1.0$. Larger $\gamma = 2.0$ sometimes further improves pass@256 but more often introduces pass@1 trade-offs, suggesting that the practical search range is small.

# M Group Size Comparison: Full Results

## M.1 Per-Benchmark Results

Table 10 provides full per-benchmark results for the group size comparison discussed in Section 5.4.

| Benchmark | Qwen2.5-7B | Qwen3-4B-Base | Llama-3.2-3B-Instruct |
|---|---|---|---|
| Mathematical reasoning | Please reason step by step, and put your final answer within \boxed{}. | System: Please reason step by step, and put your final answer within \boxed{}. Thinking: on. | 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{}. | System: Please reason step by step, and put your final answer within \boxed{}. Thinking: on. | 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. | No system prompt. Thinking: off. | Cutting Knowledge Date: December 2023\nToday Date: [date] |
| SynLogic | You are a helpful assistant. | No system prompt. Thinking: on. | Cutting Knowledge Date: December 2023\nToday Date: [date] |

Table 7: Evaluation prompt settings.

| Model | F-GRPO $\gamma$ | F-DAPO $\gamma$ | F-CISPO $\gamma$ |
|---|---|---|---|
| Qwen2.5-7B | 0.5 | 0.5 | 1.0 |
| Qwen3-4B-Base | 0.5 | 0.5 | 0.5 |
| Llama-3.2-3B-Instruct | 0.5 | 1.0 | 0.5 |

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

| Method | $\gamma$ | Qwen2.5-7B | | Qwen3-4B-Base | | Llama-3.2-3B-Instruct | |
|---|---|---|---|---|---|---|---|
| | | Avg. Math | Avg. OOD | Avg. Math | Avg. OOD | Avg. Math | Avg. OOD |
| GRPO | 0 | 37.3$^{\dagger}$/ 64.1 | 17.1 / 55.9 | 39.9 / 71.1 | 22.2 / 67.9 | **23.0** / 59.9 | **25.5** / 56.5$^{\dagger}$ |
| GRPO | 0.5 | **38.6** / 70.3 | **19.2** / **63.3** | **42.7** / 76.0$^{\dagger}$ | **24.3** / 73.6 | **23.0** / 63.4$^{\dagger}$ | 25.4$^{\dagger}$/ **57.6** |
| GRPO | 1.0 | 37.2 / 70.4$^{\dagger}$ | 17.9 / 60.6 | 41.2 / 73.5 | 22.5 / **75.9** | 22.5$^{\dagger}$/ 63.2 | 23.7 / 54.3 |
| GRPO | 2.0 | 36.1 / **70.7** | 18.9$^{\dagger}$/ 62.3$^{\dagger}$ | 41.3$^{\dagger}$/ **77.8** | 23.2$^{\dagger}$/ 73.7$^{\dagger}$ | 21.8 / **64.0** | 25.0 / 54.6 |
| DAPO | 0 | 39.4 / 69.3 | 15.7 / 58.4 | 45.1$^{\dagger}$/ 76.1 | 20.7$^{\dagger}$/ 74.4 | 24.3$^{\dagger}$/ 54.2 | 23.9 / 51.3 |
| DAPO | 0.5 | **40.5** / 72.5 | 17.9$^{\dagger}$/ 63.6 | **45.3** / 76.4 | **23.2** / 76.8$^{\dagger}$ | 21.6 / 55.3 | 24.1 / 52.3 |
| DAPO | 1.0 | 39.8$^{\dagger}$/ 74.9$^{\dagger}$ | **18.0** / **65.7** | 43.5 / 79.0$^{\dagger}$ | 18.7 / 76.7 | **24.8** / **62.3** | **24.8** / 55.4$^{\dagger}$ |
| DAPO | 2.0 | 39.5 / **76.8** | 17.6 / 64.0$^{\dagger}$ | 42.9 / **80.0** | 18.2 / **81.3** | 23.9 / 61.1$^{\dagger}$ | 24.2$^{\dagger}$/ **55.5** |
| CISPO | 0 | **39.5** / 73.2 | 14.9 / 59.0 | 45.8$^{\dagger}$/ 75.3 | 22.6 / 76.5 | 24.1$^{\dagger}$/ 58.0 | **25.7** / 52.5 |
| CISPO | 0.5 | 39.0$^{\dagger}$/ 72.1 | 17.2 / 59.4 | **46.0** / 79.2$^{\dagger}$ | **23.9** / 80.0$^{\dagger}$ | **24.5** / 59.7 | 25.0$^{\dagger}$/ 53.0$^{\dagger}$ |
| CISPO | 1.0 | **39.5** / **76.8** | **18.1** / **65.9** | 43.6 / 79.1 | 21.8 / 76.5 | 23.9 / 59.8$^{\dagger}$ | 24.3 / 52.9 |
| CISPO | 2.0 | 38.0 / 75.1$^{\dagger}$ | 17.9$^{\dagger}$/ 64.4$^{\dagger}$ | 44.2 / **79.5** | 22.9$^{\dagger}$/ **80.7** | 22.0 / **63.8** | 23.5 / **57.0** |

Table 9: Focal-weight sensitivity at $N{=}8$ across models and base methods. Metrics are average pass@1 / pass@256. $\gamma = 0$ denotes the unweighted baseline for each base method. **Bold**: best; †: second best within each base method and metric column.

| 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 $N{=}4$ | 36.4 / 71.1 | 9.6 / 48.9 | 5.4 / 52.6 | 53.3 / 95.9 | 76.3 / 94.7 | 35.2 / 61.3 | 38.5 / 73.0 | 18.7 / 59.7 | 32.1 / 72.2 | **9.3** / 57.0 | 14.7 / 49.8 |
| 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 $N{=}16$ | 38.4 / 67.5 | 13.0 / 43.9 | 7.1 / 48.0 | 57.7 / 93.4 | 76.7 / 93.2 | 35.7 / 58.0 | **40.1** / 68.4 | 17.7 / 55.7 | 31.6 / 68.9 | 8.6 / 49.6 | 12.9 / 48.6 |
| 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 10: 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. Gray rows mark the same group size $N{=}8$ comparison.

### M.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

| Method | In-domain | | | | | | | Out-of-domain | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Avg. | AIME24 | AIME25 | AMC | MATH500 | Minerva | Olympiad | Avg. OOD | IFEval | SynLogic | GPQA |
| GRPO low-LR | 37.8$^{\dagger}$/ 69.2 | 15.0$^{\dagger}$/ 51.0 | 8.7 / 44.5 | 53.3 / 95.3 | 75.9 / 94.4 | 35.0$^{\dagger}$/ 59.7 | **38.7** / 70.4 | 16.4 / 57.9 | 30.0 / 69.3 | 7.8 / 55.4 | 11.5 / 49.1 |
| GRPO ($\mathcal{H}$) | 37.8$^{\dagger}$/ 69.5 | 14.9 / 48.9 | 7.3 / 52.2 | 55.8$^{\dagger}$/ 90.8 | 75.6 / 94.6 | 34.9 / 61.3 | 38.2$^{\dagger}$/ 69.2 | 18.7 / 59.9 | 32.1 / 71.9 | **9.8** / **59.9** | 14.3 / 47.8 |
| GRPO (KL) | 37.2 / 72.0$^{\dagger}$ | 13.2 / 53.4$^{\dagger}$ | 8.7 / **53.7** | 52.1 / 95.9$^{\dagger}$ | **76.7** / 95.2$^{\dagger}$ | 34.7 / 61.5$^{\dagger}$ | 38.0 / 72.3$^{\dagger}$ | **19.4** / 60.0 | 32.4 / 70.8 | 8.8$^{\dagger}$/ 51.7 | **17.1** / 57.5$^{\dagger}$ |
| DS-GRPO | 37.7 / **73.8** | 14.3 / **57.7** | **12.2** / 53.6$^{\dagger}$ | 51.6 / 95.5 | 76.6$^{\dagger}$/ **96.8** | 33.5 / **62.8** | 37.8 / **76.7** | 17.9 / **68.3** | 33.4$^{\dagger}$/ **85.9** | 7.5 / 57.7$^{\dagger}$ | 12.7 / **61.3** |
| F-GRPO | **38.6** / 70.3 | **15.9** / 46.2 | 10.1$^{\dagger}$/ 52.6 | **56.2** / **96.3** | 76.2 / 95.1 | **35.7** / 60.3 | 37.5 / 71.6 | 19.2$^{\dagger}$/ 63.3$^{\dagger}$ | **34.0** / 75.7$^{\dagger}$ | 8.7 / 57.0 | 15.0$^{\dagger}$/ 57.3 |

Table 11: F-GRPO vs. lower-LR, regularized, and DS-GRPO baselines on Qwen2.5-7B at $N$=8. GRPO low-LR uses $\mathrm{lr}_{\mathrm{low}} = 6.8\times10^{-7}$; GRPO-$\mathcal{H}$ and GRPO-KL use coefficient 0.001. Pass@1 / pass@256. **Bold**: best; †: second best.

initially low-probability correct solutions.

## N Regularization and Update-Scale Controls: Full Results

**Learning-rate multiplier.** For GRPO low-LR, we match the average advantage-magnitude scale induced by Focal weighting. Write $\widehat{\mu} = \widehat{\mu}_{\mathrm{pos}}(x)$. With binary rewards, the mean absolute normalized GRPO advantage at sampled success rate $\widehat{\mu}$ is proportional to $\sqrt{\widehat{\mu}(1-\widehat{\mu})}$, while F-GRPO multiplies it by $(1-\widehat{\mu})^{\gamma}$. Averaging this multiplicative reduction over the success-rate axis $\widehat{\mu} \in [0,1]$ gives

$$m_{\gamma} = \frac{\int_0^1 (1-\widehat{\mu})^{\gamma}\sqrt{\widehat{\mu}(1-\widehat{\mu})}\,d\widehat{\mu}}{\int_0^1 \sqrt{\widehat{\mu}(1-\widehat{\mu})}\,d\widehat{\mu}}.$$

The numerator is $B(\frac{3}{2}, \gamma+\frac{3}{2})$ and the denominator is $B(\frac{3}{2},\frac{3}{2})$, where $B(a,b) := \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$ is the beta function. Therefore,

$$m_{\gamma} = \frac{B(\frac{3}{2}, \gamma+\frac{3}{2})}{B(\frac{3}{2},\frac{3}{2})}.$$

Using $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ together with $\Gamma(\frac{3}{2}) = \sqrt{\pi}/2$ and $\Gamma(3) = 2$, we obtain

$$m_{\gamma} = \frac{4}{\sqrt{\pi}}\frac{\Gamma(\gamma+\frac{3}{2})}{\Gamma(\gamma+3)}.$$

## O LLM Pass@$k$ Tables

Tables 12-14 report benchmark-level pass@$k$ values up to $k$=256 for the same selected checkpoints used in Table 1. For each model, we follow the same evaluation protocol as in the main results. Each cell shows baseline/focal pass@$k$, with the better value in bold.

## P LLM Training Dynamics

We include training-dynamics plots (Figures 8 - 10) for the Qwen3-4B-Base setup, for which denser periodic evaluation was practical because generation length was capped at 4096 tokens. These curves are intended as diagnostics rather than replacements for the main evaluation protocol. We score AIME25 every 25 optimization steps using 64 generations per problem, so the plotted pass@1 is the corresponding estimate from 64 samples rather than the final 1024-sample evaluation reported in the main tables. In the reward panel, the bold curve shows an exponential moving average (EMA) with smoothing coefficient 0.08, overlaid on the raw per-step reward trace.

## Q 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.

## R Notation

Table 15 summarizes the main notation used throughout the paper.

| Method | $k$ | AIME24 | AIME25 | AMC23 | MATH500 | Minerva | Olympiad | IFEval | SynLogic | GPQA |
|---|---|---|---|---|---|---|---|---|---|---|
| GRPO/ F-GRPO | 1 | 15.0/**15.9** | 6.7/**10.1** | 52.9/**56.2** | 75.8/**76.2** | **36.0**/35.7 | **37.8**/37.5 | 32.1/**34.0** | 7.9/**8.7** | 11.3/**15.0** |
| | 2 | 17.4/**20.1** | 9.4/**14.2** | 59.9/**65.9** | 80.0/**81.2** | **41.1**/40.9 | 43.5/**43.9** | 39.9/**42.7** | 11.0/**12.1** | 16.8/**21.8** |
| | 4 | 19.8/**23.2** | 13.4/**19.1** | 65.1/**73.5** | 83.2/**84.8** | 45.4/45.4 | 48.6/**49.6** | 46.5/**49.9** | 15.1/**16.8** | 22.5/**29.1** |
| | 8 | 22.6/**26.2** | 18.3/**23.9** | 69.2/**79.3** | 85.9/**87.7** | 49.0/49.0 | 52.9/**54.4** | 52.0/**55.9** | 20.2/**22.9** | 28.0/**36.0** |
| | 16 | 25.6/**30.0** | 23.1/**28.3** | 73.1/**83.5** | 88.0/**90.0** | 51.9/51.9 | 56.6/**58.4** | 56.7/**60.9** | 26.2/**30.1** | 32.7/**41.8** |
| | 32 | 29.0/**34.6** | 27.1/**32.8** | 77.2/**87.2** | 89.5/**91.8** | **54.5**/54.3 | 59.6/**62.0** | 60.9/**65.5** | 32.7/**37.6** | 36.8/**46.7** |
| | 64 | 32.2/**39.2** | 30.9/**38.2** | 81.0/**91.2** | 90.7/**93.1** | **56.8**/56.4 | 62.0/**65.5** | 64.5/**69.7** | 39.4/**44.8** | 40.4/**50.8** |
| | 128 | 35.0/**42.9** | 35.2/**44.8** | 84.5/**94.5** | 91.8/**94.2** | **58.7**/58.4 | 63.9/**68.8** | 67.7/**73.1** | 45.7/**51.3** | 43.5/**54.2** |
| | 256 | 37.7/**46.2** | 40.8/**52.6** | 87.3/**96.3** | 92.8/**95.1** | 60.2/**60.3** | 65.8/**71.6** | 70.3/**75.7** | 51.3/**57.0** | 46.2/**57.3** |
| DAPO/ F-DAPO | 1 | 16.8/**20.9** | **12.0**/11.5 | 53.3/**55.9** | 78.6/**79.1** | **35.5**/35.0 | 40.5/**40.9** | 24.1/**30.8** | 7.5/**7.9** | **15.4**/15.0 |
| | 2 | 20.6/**25.3** | 15.6/**15.7** | 61.6/**64.2** | 82.9/**84.4** | 41.3/41.3 | 46.6/**48.2** | 33.1/**39.1** | 10.1/**11.4** | **22.2**/21.9 |
| | 4 | 23.9/**28.9** | 19.8/**20.1** | 68.5/**70.9** | 86.1/**88.0** | 45.8/**46.0** | 51.6/**54.3** | 41.0/**46.1** | 13.8/**16.6** | 28.8/**28.9** |
| | 8 | 27.4/**32.6** | 24.4/**25.3** | 74.3/**77.3** | 88.6/**90.8** | 49.6/**50.0** | 56.1/**59.2** | 47.5/**51.8** | 19.0/**23.6** | 34.3/**34.6** |
| | 16 | 31.1/**36.6** | 29.1/**31.1** | 79.4/**83.0** | 90.6/**92.9** | 52.8/**53.4** | 60.0/**63.3** | 52.8/**56.6** | 25.4/**31.9** | 38.7/**39.7** |
| | 32 | 35.2/**40.7** | 34.2/**37.2** | 83.9/**87.8** | 92.1/**94.4** | 55.5/**56.1** | 63.3/**67.0** | 57.3/**60.9** | 32.4/**40.3** | 42.8/**44.7** |
| | 64 | 39.5/**44.7** | 39.0/**43.1** | 87.6/**91.0** | 93.4/**95.4** | 57.8/**58.6** | 66.3/**70.3** | 61.1/**64.7** | 39.3/**48.2** | 46.9/**49.8** |
| | 128 | 44.3/**49.0** | 42.6/**48.2** | 90.4/**92.8** | 94.4/**96.1** | 59.7/**60.9** | 69.0/**73.1** | 64.4/**68.0** | 46.1/**55.5** | 51.0/**54.0** |
| | 256 | 49.8/**53.4** | 45.6/**52.9** | 91.9/**93.7** | 95.2/**96.6** | 61.2/**62.9** | 71.8/**75.6** | 67.1/**71.1** | 53.3/**62.4** | 54.9/**57.4** |
| CISPO/ F-CISPO | 1 | 14.6/**14.8** | 9.7/**13.0** | **57.8**/53.3 | 78.7/**79.0** | **34.7**/34.6 | 41.5/**42.4** | 24.2/**30.7** | 8.0/**8.2** | 12.6/**15.4** |
| | 2 | 18.7/**20.2** | 14.2/**18.0** | **67.4**/64.2 | 84.4/**85.0** | 40.5/**41.7** | 48.9/**50.1** | 32.2/**38.7** | 11.3/**11.7** | 19.2/**23.5** |
| | 4 | 22.9/**25.3** | 19.6/**22.8** | 74.9/**75.0** | 88.4/**88.9** | 45.8/**47.4** | 55.3/**56.3** | 39.1/**45.3** | 15.8/**16.7** | 26.6/**32.1** |
| | 8 | 27.8/**30.8** | 26.0/**28.0** | 82.0/**83.0** | 91.2/**91.7** | 50.5/**51.8** | 60.6/**61.3** | 45.3/**51.0** | 21.6/**23.1** | 33.7/**40.3** |
| | 16 | 32.9/**36.8** | 33.0/**34.4** | 87.4/**89.5** | 93.2/**93.9** | 54.2/**55.1** | 65.1/**65.6** | 50.9/**55.9** | 28.0/**30.5** | 39.7/**47.7** |
| | 32 | 37.3/**43.4** | 40.3/**42.1** | 91.5/**93.2** | 94.7/**95.5** | 57.0/**57.7** | 68.8/**69.2** | 55.9/**60.1** | 34.5/**38.2** | 44.8/**54.2** |
| | 64 | 40.3/**50.0** | 47.3/**50.5** | 94.2/**95.1** | 95.7/**96.6** | 59.3/**60.1** | 71.9/**72.1** | 60.4/**64.0** | 40.8/**45.8** | 49.1/**59.6** |
| | 128 | 42.7/**55.4** | 54.0/**58.4** | 95.5/**96.2** | 96.5/**97.3** | 61.4/**62.3** | 74.6/**74.9** | 64.3/**67.4** | 47.2/**53.1** | 52.6/**63.9** |
| | 256 | 45.9/**59.7** | 59.8/**64.6** | 96.1/**97.1** | 97.0/**97.8** | 63.3/**64.3** | 76.9/**77.5** | 67.9/**70.6** | 53.6/**60.0** | 55.5/**67.1** |

Table 12: Qwen2.5-7B pass@$k$ up to 256 for the selected checkpoints from Table 1. Cells report baseline/focal values.

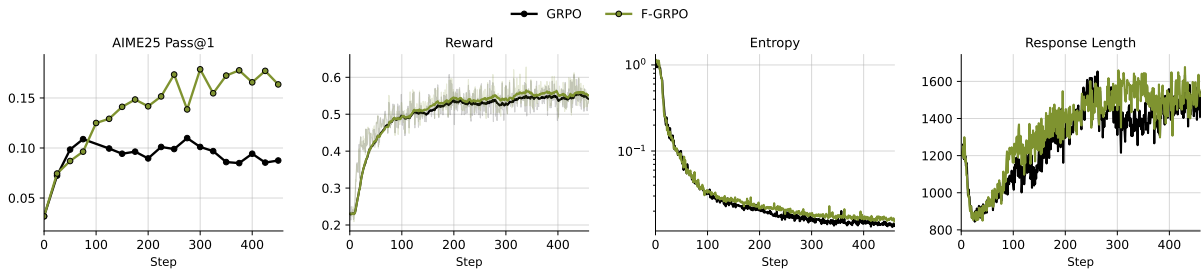


Figure 8: Training dynamics on Qwen3-4B-Base for GRPO and F-GRPO ($\gamma = 0.5$).

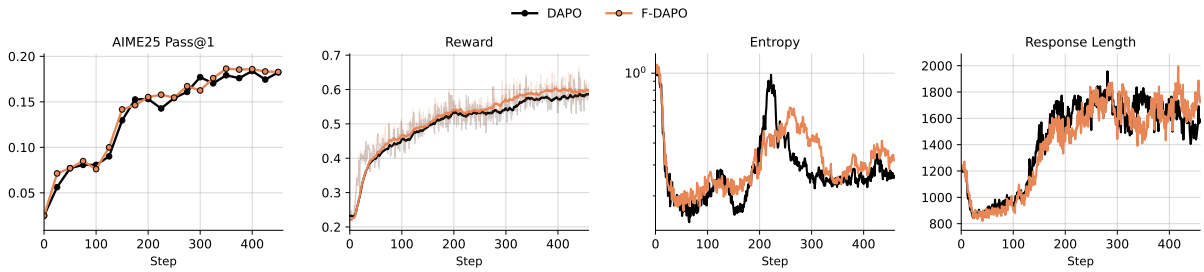


Figure 9: Training dynamics on Qwen3-4B-Base for DAPO and F-DAPO ($\gamma = 0.5$).

| Method | $k$ | AIME24 | AIME25 | AMC23 | MATH500 | Minerva | Olympiad | IFEval | SynLogic | GPQA |
|---|---|---|---|---|---|---|---|---|---|---|
| GRPO/ F-GRPO | 1 | 14.8/**17.4** | 8.3/**15.8** | 54.6/**58.3** | 79.5/**82.3** | **39.2**/38.5 | 43.1/**43.7** | 36.9/**41.2** | 10.7/10.7 | 19.0/**21.1** |
| | 2 | 18.4/**23.0** | 12.5/**20.9** | 62.7/**68.4** | 84.3/**87.0** | 43.7/**44.2** | 49.3/**49.9** | 47.8/**53.1** | 14.6/**15.3** | 26.0/**28.3** |
| | 4 | 22.1/**29.2** | 17.6/**25.7** | 70.0/**77.2** | 87.6/**90.1** | 47.6/**48.6** | 54.6/**55.1** | 57.8/**63.6** | 19.3/**21.1** | 32.6/**35.1** |
| | 8 | 26.4/**35.8** | 23.2/**30.7** | 76.3/**84.0** | 90.0/**92.4** | 51.0/**52.1** | 59.1/**59.6** | 66.6/**72.1** | 24.9/**28.0** | 38.0/**41.2** |
| | 16 | 31.8/**42.1** | 28.5/**35.8** | 81.8/**88.3** | 91.9/**94.0** | 53.9/**55.2** | 62.8/**63.6** | 73.8/**78.5** | 31.3/**35.8** | 42.3/**46.6** |
| | 32 | 37.7/**47.1** | 33.3/**40.8** | 86.7/**91.1** | 93.4/**95.3** | 56.4/**58.0** | 66.2/**67.1** | 79.4/**83.3** | 38.2/**43.8** | 45.7/**51.5** |
| | 64 | 43.4/**51.4** | 37.6/**46.2** | 91.0/**93.6** | 94.5/**96.2** | 58.5/**60.3** | 69.3/**70.2** | 83.8/**87.3** | 45.0/**51.4** | 49.0/**55.7** |
| | 128 | 48.4/**55.9** | 41.4/**52.7** | 94.3/**96.0** | 95.5/**96.8** | 60.1/**62.1** | 72.0/**73.0** | 87.7/**90.8** | 51.3/**58.3** | 52.3/**59.5** |
| | 256 | 52.5/**61.1** | 45.6/**60.0** | 96.2/**98.2** | 96.4/**97.4** | 61.7/**63.6** | 74.2/**75.5** | 91.2/**93.6** | 57.1/**64.5** | 55.4/**62.8** |
| DAPO/ F-DAPO | 1 | 18.6/**19.1** | 17.6/**17.7** | 62.0/**63.0** | 84.4/84.4 | **40.0**/39.5 | 47.8/**48.2** | 33.8/**39.6** | **11.8**/11.3 | 16.4/**18.7** |
| | 2 | 23.2/**23.7** | 21.8/**21.9** | 71.6/**72.8** | 88.7/**88.9** | 46.1/**46.2** | 54.9/**55.3** | 45.8/**52.4** | 17.3/**17.5** | 23.8/**27.0** |
| | 4 | 28.0/28.0 | 25.7/**26.3** | 78.7/**79.9** | 91.4/91.4 | 51.1/**51.5** | 60.3/**60.7** | 56.5/**63.2** | 24.2/**25.1** | 31.7/**35.3** |
| | 8 | **33.6**/32.8 | 30.2/**31.8** | 83.5/**84.6** | **93.2**/93.0 | 55.3/**55.6** | 64.5/**64.9** | 65.3/**71.4** | 32.0/**33.3** | 38.7/**42.6** |
| | 16 | **39.7**/38.3 | 35.6/**38.0** | 86.7/**88.3** | **94.7**/94.3 | 58.3/**59.0** | 67.8/**68.3** | 72.4/**77.3** | 40.1/**42.0** | 44.7/**49.0** |
| | 32 | **45.9**/43.6 | 41.5/**44.1** | 89.3/**91.6** | **95.9**/95.3 | 60.7/**61.8** | 70.6/**71.0** | 77.8/**81.6** | 48.6/**50.8** | 50.3/**54.8** |
| | 64 | **52.4**/48.1 | 46.7/**49.2** | 92.0/**94.1** | **96.9**/96.3 | 62.9/**64.1** | 73.2/**73.5** | 81.8/**85.2** | 56.8/**59.1** | 55.3/**60.1** |
| | 128 | **58.5**/53.1 | 50.4/**53.5** | 94.5/**95.6** | **97.5**/97.1 | 65.1/**66.0** | 75.6/**75.8** | 85.2/**88.1** | 64.3/**66.2** | 59.9/**64.4** |
| | 256 | **63.8**/59.7 | 54.0/**58.3** | 96.2/**96.8** | **97.8**/97.7 | 67.1/**67.9** | 77.9/**78.0** | 88.0/**90.6** | 71.3/**72.4** | 63.9/**67.5** |
| CISPO/ F-CISPO | 1 | 21.5/21.5 | 17.6/**18.6** | **63.1**/60.7 | 84.2/84.2 | 40.2/**42.1** | 48.4/**48.9** | 32.6/**35.9** | **11.6**/11.1 | 23.6/**24.6** |
| | 2 | **27.5**/27.2 | 20.9/**22.4** | **72.0**/71.3 | 88.3/**88.7** | 46.0/**48.2** | 55.2/**55.8** | 42.6/**47.3** | **17.1**/16.1 | 32.5/**34.3** |
| | 4 | 32.5/**32.8** | 24.7/**26.8** | **80.7**/80.5 | 91.0/**91.5** | 50.7/**52.7** | 60.5/**61.2** | 51.8/**57.5** | **24.0**/22.7 | 41.0/**43.7** |
| | 8 | 37.4/**38.5** | 30.0/**32.5** | 85.9/**86.0** | 92.9/**93.6** | 54.4/**56.2** | 64.5/**65.3** | 59.9/**66.4** | **32.4**/30.9 | 47.9/**51.7** |
| | 16 | 42.6/**44.2** | 36.4/**39.2** | 89.4/**89.7** | 94.4/**95.0** | 57.3/**59.0** | 67.7/**68.6** | 67.1/**73.5** | **41.4**/39.8 | 53.4/**58.4** |
| | 32 | 47.6/**50.4** | 42.4/**45.6** | 92.0/**92.3** | 95.6/**96.0** | 59.5/**61.5** | 70.4/**71.6** | 73.2/**78.9** | **50.1**/48.2 | 58.3/**64.6** |
| | 64 | 51.8/**56.6** | 46.8/**51.2** | 94.0/**94.7** | 96.6/**96.9** | 61.5/**63.7** | 72.7/**74.3** | 78.2/**83.2** | **57.8**/55.7 | 63.2/**70.8** |
| | 128 | 55.5/**63.0** | 50.7/**56.5** | 96.5/**96.6** | 97.4/**97.6** | 63.4/**65.5** | 74.7/**76.7** | 82.4/**87.0** | **64.8**/62.3 | 67.9/**76.6** |
| | 256 | 59.3/**69.5** | 55.4/**62.7** | 97.4/**98.6** | 97.9/**98.2** | 65.1/**67.2** | 76.6/**78.8** | 85.9/**90.3** | **71.1**/68.0 | 72.4/**81.6** |

Table 13: Qwen3-4B-Base pass@$k$ up to 256 for the selected checkpoints from Table 1. Cells report baseline/focal values.

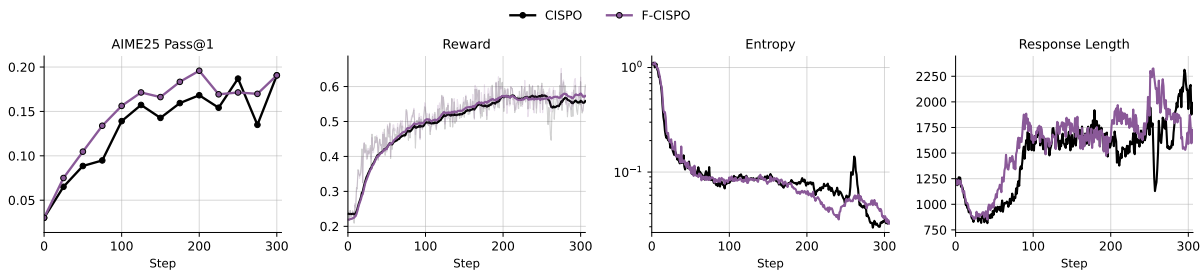


Figure 10: Training dynamics on Qwen3-4B-Base for CISPO and F-CISPO ($\gamma = 0.5$).

| Method | $k$ | AIME24 | AIME25 | AMC23 | MATH500 | Minerva | Olympiad | IFEval | SynLogic | GPQA |
|---|---|---|---|---|---|---|---|---|---|---|
| GRPO/ F-GRPO | 1 | 10.7/**12.1** | 0.7/**1.0** | **30.5**/29.8 | **55.0**/54.1 | **21.8**/21.0 | 19.4/**20.1** | 54.1/**56.4** | **4.7**/4.6 | **17.5**/15.2 |
| | 2 | 15.8/**17.5** | 1.4/**1.8** | 41.0/**41.4** | **63.4**/63.1 | 28.1/28.1 | 25.5/**26.7** | 61.7/**64.0** | **7.0**/6.8 | **24.4**/22.4 |
| | 4 | 20.6/**22.9** | 2.6/**3.2** | 50.6/**52.4** | 69.7/**70.1** | 34.0/**34.7** | 31.1/**32.8** | 66.7/**68.9** | **9.8**/9.4 | 29.5/**30.7** |
| | 8 | 24.3/**27.5** | 4.6/**5.1** | 58.7/**61.9** | 74.9/**75.7** | 39.2/**40.3** | 36.4/**38.4** | 70.0/**72.2** | **13.3**/12.7 | 35.9/**36.3** |
| | 16 | 26.9/**30.7** | 7.4/**7.6** | 65.7/**70.1** | 79.3/**80.4** | 43.6/**44.9** | 41.4/**43.7** | 72.1/**74.5** | **17.7**/17.0 | 41.2/**41.7** |
| | 32 | 29.2/**33.0** | 10.3/**10.6** | 72.4/**77.2** | 82.9/**84.7** | 47.6/**48.9** | 46.1/**48.6** | 73.8/**76.3** | **22.8**/21.9 | 45.6/**47.0** |
| | 64 | 32.0/**35.8** | 13.3/**14.3** | 78.7/**83.1** | 86.0/**88.3** | 51.4/**52.8** | 50.7/**53.2** | 75.3/**77.7** | **27.8**/26.9 | 49.6/**51.5** |
| | 128 | 35.8/**40.0** | 16.5/**20.5** | 83.9/**87.6** | 88.6/**91.1** | 55.2/**56.5** | 55.2/**57.4** | 76.7/**78.7** | **32.3**/31.5 | 52.9/**55.1** |
| | 256 | 40.7/**46.1** | 21.5/**29.5** | 88.2/**90.6** | 90.6/**92.9** | 59.0/**60.1** | 59.3/**61.3** | 78.0/**79.6** | **36.4**/35.5 | 55.1/**57.6** |
| DAPO/ F-DAPO | 1 | **12.8**/11.1 | 1.0/**1.7** | **33.1**/31.9 | 55.9/**58.6** | **22.4**/22.3 | 21.0/**23.2** | 51.2/**53.0** | **4.8**/4.3 | 15.7/**17.0** |
| | 2 | **17.2**/16.1 | 1.8/**3.0** | **43.3**/42.2 | 61.6/**66.5** | 27.9/**29.0** | 25.7/**29.3** | 59.7/**61.6** | **6.7**/6.3 | 21.2/**23.6** |
| | 4 | 20.4/**21.4** | 2.9/**4.8** | 52.6/**53.0** | 66.2/**72.5** | 32.9/**35.1** | 29.8/**34.8** | 65.6/**67.0** | 8.6/**8.8** | 26.1/**29.8** |
| | 8 | 24.8/**25.0** | 4.4/**6.8** | 60.4/**62.0** | 70.0/**77.5** | 37.3/**40.1** | 33.4/**39.7** | 69.2/**70.5** | 11.0/**12.2** | 30.6/**35.5** |
| | 16 | 28.2/**28.6** | 6.0/**8.9** | 65.8/**69.7** | 73.1/**81.7** | 41.1/**44.5** | 36.6/**44.3** | 71.7/**73.0** | 14.1/**16.4** | 34.9/**40.7** |
| | 32 | 31.3/**32.3** | 8.1/**11.5** | 70.2/**75.4** | 76.0/**85.2** | 44.6/**48.6** | 39.6/**48.7** | 73.7/**75.1** | 17.7/**21.1** | 38.9/**44.9** |
| | 64 | 34.5/**36.2** | 10.8/**15.5** | 74.5/**79.9** | 78.9/**88.1** | 47.8/**52.5** | 42.5/**53.1** | 75.4/**77.0** | 21.6/**25.6** | 42.2/**48.4** |
| | 128 | 37.5/**40.2** | 14.1/**21.5** | 77.7/**84.2** | 81.5/**90.4** | 50.9/**56.1** | 45.4/**57.3** | 76.7/**78.5** | 25.4/**29.4** | 44.7/**51.3** |
| | 256 | 40.8/**44.4** | 18.5/**28.7** | 79.5/**88.3** | 83.8/**92.0** | 54.1/**59.3** | 48.4/**61.3** | 77.8/**79.5** | 28.9/**33.0** | 47.0/**53.7** |
| CISPO/ F-CISPO | 1 | 9.7/**10.6** | 1.0/**2.0** | 32.9/**34.1** | **56.9**/56.5 | 21.8/**22.1** | **22.5**/21.5 | **54.6**/52.6 | 4.3/**5.4** | **18.2**/17.0 |
| | 2 | 15.0/**15.9** | 1.8/**3.4** | 43.2/**44.9** | 64.1/**64.5** | 28.1/28.1 | **27.9**/27.5 | **62.5**/60.6 | 6.2/**7.4** | **24.1**/22.8 |
| | 4 | 21.0/**21.2** | 3.2/**5.1** | 53.2/**53.3** | 69.8/**70.6** | **34.0**/33.7 | 32.8/**33.0** | **67.6**/65.7 | 8.6/**9.8** | **29.5**/27.8 |
| | 8 | 25.4/**26.8** | 5.1/**7.0** | 59.4/**61.3** | 74.3/**75.6** | **39.3**/38.8 | 37.3/**38.1** | **70.8**/69.0 | 11.7/**12.8** | **34.1**/32.3 |
| | 16 | 28.7/**31.3** | 7.5/**8.9** | 64.1/**66.9** | 77.9/**79.8** | **44.0**/43.3 | 41.4/**42.8** | **73.2**/71.3 | 15.1/**16.6** | **38.3**/36.2 |
| | 32 | 31.6/**34.1** | 10.8/**11.2** | 68.8/**70.6** | 81.0/**83.4** | **48.4**/47.5 | 45.1/**47.2** | **75.0**/73.1 | 18.7/**20.9** | **42.2**/39.5 |
| | 64 | 35.1/**36.1** | **15.1**/14.3 | 73.4/**74.1** | 83.9/**86.4** | **52.4**/51.5 | 48.7/**51.1** | **76.2**/74.7 | 22.1/**25.4** | **45.3**/42.4 |
| | 128 | 37.9/**39.0** | **20.0**/18.6 | 76.1/**78.9** | 86.7/**88.9** | **56.1**/55.3 | 52.1/**54.8** | **77.3**/76.2 | 25.7/**29.7** | **47.7**/45.1 |
| | 256 | 39.4/**42.8** | **25.4**/24.5 | 79.1/**82.6** | 89.1/**91.0** | **59.5**/58.8 | 55.4/**58.7** | **78.4**/77.3 | 29.4/**33.9** | **49.7**/47.7 |

Table 14: Llama-3.2-3B-Instruct pass@$k$ up to 256 for the selected checkpoints from Table 1. Cells report baseline/focal values.

| Category | Symbol | Meaning |
|---|---|---|
| **Trajectory-level variables** | $\pi_\theta$ | The policy parameterized by $\theta$ |
| | $D$ | Prompt distribution |
| | $x$ | Given prompt |
| | $o, y$ | A complete response (trajectory) generated by $\pi_\theta$ 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: $\Pr_{o\sim\pi_\theta(\cdot\|x)}[o \in \mathcal{C}(x)]$ |
| | $X_x$ | Number of correct rollouts for prompt $x$ in a sampled group |
| | $E_x$ | Target subset of correct rollouts for prompt $x$, $E_x \subseteq \mathcal{C}(x)$ |
| | $\tau_E(x)$ | Conditional mass of $E_x$: $\Pr_{o\sim\pi_\theta(\cdot\|x)}[o \in E_x]$ |
| | $\widehat{\mu}_{\text{pos}}(x)$ | 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: $\frac{1}{N}\sum_{j=1}^{N} R_j$ |
| | $\sigma_R$ | 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: $R_c P_{\text{pos}} + R_w P_{\text{neg}}$ |
| | $\Delta z_i$ | One-step logit update: $\frac{\eta}{N} p_i (R_i - S_R)$ |
| | $\Delta Q_{\text{pos}}$ | One-step change in total correct mass |
| | $\Delta Q_{\text{u,pos}}$ | One-step change in unsampled-correct mass |
| | $g(x)$ | Focal weight: $(1 - \widehat{\mu}_{\text{pos}}(x))^\gamma$ |
| | $\gamma$ | Focal-weight exponent controlling reweighting strength |
| | $\eta$ | 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** | $\mathcal{A}_N(x)$ | Active event for prompt $x$: $\{0 < X_x < N\}$ |
| | $\mathcal{B}_{E,N}(x)$ | Tail-miss event: active update for prompt $x$ that samples no rollout from $E_x$ |
| | $\Pr(\mathcal{B}_{E,N}(x) \mid x)$ | Pointwise probability of the tail-miss event |
| **Sets** | $\Omega_x$ | Space of complete rollouts for prompt $x$ |
| | $\mathcal{C}(x)$ | Subset of 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 15: Notation used in the paper.