UNIFORM-CORRECT POLICY OPTIMIZATION: BREAKING RLVR'S INDIFFERENCE TO DIVERSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has achieved substantial gains in single-attempt accuracy (Pass@1) on reasoning tasks, yet often suffers from reduced multi-sample coverage (Pass@K), indicating diversity collapse. We identify a structural cause for this degradation: common RLVR objectives, such as GRPO, are indifferent to how probability mass is distributed among correct solutions. Combined with stochastic training dynamics, this indifference induces a self-reinforcing collapse, in which probability mass concentrates on a narrow subset of correct outputs while alternative valid solutions are suppressed. We formalize this collapse mechanism and further characterize the *optimal policy structure* under two complementary criteria: conditional-entropy optimality and robustness, which identify the Uniform-Correct Policy as uniquely optimal. Motivated by this analysis, we propose *Uniform-Correct Policy Optimization* (UCPO), a modification to GRPO that adds a conditional uniformity penalty on the policy's distribution over correct solutions. The penalty redistributes gradient signal toward underrepresented correct responses, encouraging uniform allocation of probability mass within the correct set. Across three models (1.5B–7B parameters) and five mathematical reasoning benchmarks, **UCPO achieves a stronger correctness-coverage trade-off:** it obtains the highest average Pass@64 across all three model families while preserving Pass@1, avoiding the single-sample accuracy degradation of competing diversity-oriented methods. UCPO achieves gains of up to +10 points on AIME24 at Pass@64 and up to 45% higher equation-level diversity within the correct set.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a powerful paradigm for improving large language models (LLMs) on reasoning tasks such as mathematics and code generation, yielding substantial gains in single-attempt accuracy (Pass@1) across challenging reasoning benchmarks (DeepSeek-AI et al., 2025; Shao et al., 2024; Karan & Du, 2025; Yue et al., 2025). Despite these successes, RLVR often degrades Pass@K even as it improves Pass@1 (He et al., 2025; Cobbe et al., 2021; Song et al., 2025). This suggests that RL fine-tuning contracts the model's solution space (Wu et al., 2026; Karan & Du, 2025), concentrating probability mass on a small subset of correct solutions while suppressing alternative valid reasoning paths present in the base model.

Most prior efforts address this issue (Wang et al., 2025b; He et al., 2025; Jiang et al., 2025; Wang et al., 2025a; Cui et al., 2025; Chen et al., 2025) through entropy bonuses, exploration incentives, or token-level constraints. While these methods can alleviate diversity loss, they share a common limitation: none directly specifies how probability mass should be distributed among correct solutions. Lacking a principled target, they rely on proxies that may satisfy local heuristics (e.g., token-level entropy) while the policy still concentrates mass on a few correct solutions.

In this work, we analyze RLVR from the perspective of *optimal policy structure*. We ask three questions: (i) what policies are optimal under standard RLVR objectives, and what policies are actually attained in practice; (ii) what policy should be considered optimal when multiple correct solutions exist; and (iii) how such a policy can be learned via RL. These questions guide our analysis and the design of a new RL objective.

We first show that the standard RLVR objective is fundamentally underspecified: it is indifferent to how probability mass is distributed among correct solutions. As a result, a fully collapsed policy and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a maximally diverse policy will be considered equally optimal. Moreover, the stochastic dynamics of on-policy sampling induce a *self-reinforcing collapse* in which frequently sampled solutions are amplified by gradient updates and suppress alternatives. Thus, diversity collapse arises jointly from objective indifference and training dynamics.

We then study what the optimal policy should be when multiple correct solutions exist. We characterize this through two criteria: conditional-entropy optimality, which resolves RLVR's indifference among correctness-optimal policies, and robustness optimality, which maximizes retained correct probability under distribution shift. Both identify the Uniform-Correct Policy, which places zero mass on incorrect outputs while distributing probability uniformly across correct solutions.

Motivated by this analysis, we propose *Uniform-Correct Policy Optimization* (UCPO), a simple modification of GRPO that admits the Uniform-Correct Policy as its unique optimum. UCPO adds a uniformity penalty over correct solutions, which redistributes gradient signal to amplify underrepresented solutions while tempering dominant ones. Our contributions are as follows:

- We identify a fundamental limitation of RLVR: the objective is indifferent to how probability mass is distributed among correct solutions, allowing stochastic training dynamics to drive irreversible mode collapse. We validate this mechanism both theoretically (Section 3.2) and empirically through controlled experiments (Section 3.3).
- We resolve RLVR's indifference among correct solutions by characterizing the Uniform-Correct Policy: it uniquely maximizes conditional entropy among RLVR-optimal policies and is minimax-optimal under bounded distribution shift (Section 4).
- We derive Uniform-Correct Policy Optimization (UCPO) to achieve this optimal structure, and prove that the Uniform-Correct Policy is its unique optimum (Theorem 5.1). UCPO preserves the total advantage mass of GRPO but reallocates it toward underrepresented correct responses.
- Across three models (1.5B–7B parameters) and five mathematical benchmarks, UCPO improves the correctness–coverage trade-off, increasing Pass@K and equation-level diversity while better preserving Pass@1 than competing diversity-oriented methods. It achieves up to a +10 percentage-point gain on AIME24 at Pass@64 and up to 45% higher equation-level diversity, with negligible training overhead.

2 Preliminaries

2.1 RLVR Objective

Let x denote a prompt and y a model completion sampled from a policy $\pi_\theta(y \mid x)$. In RLVR, correctness is determined by a verifier $R(x, y) \in \{0, 1\}$, and the set of verifier-accepted (correct) solutions is $\mathcal{Y}^+(x) = \{y : R(x, y) = 1\}$. The RLVR objective maximizes the probability of producing any correct solution:

$$J(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[R(x, y)] = \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in\mathcal{Y}^+(x)} \pi_\theta(y|x)\right]. \tag{1}$$

We denote the total probability mass the policy places on correct outputs as $Z_\theta(x) = \sum_{y\in\mathcal{Y}^+(x)} \pi_\theta(y \mid x)$, so that $J(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}}[Z_\theta(x)]$. Maximizing the RLVR objective is equivalent to maximizing $Z_\theta(x)$, or equivalently $\log Z_\theta(x)$. Taking the gradient of Eq. 1 yields the standard policy gradient form:

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D},\ y\sim\pi_\theta(\cdot|x)}\left[R(x, y)\nabla_\theta \log \pi_\theta(y \mid x)\right]. \tag{2}$$

The gradient signal depends only on whether y is correct, not on how probability mass is distributed among different correct outputs.

2.2 Group Relative Policy Optimization

GRPO (Shao et al., 2024) estimates RLVR policy gradients using group-relative advantages. For each prompt x, it samples K rollouts $\{y_i\}_{i=1}^K$ from a rollout policy, obtains rewards $R_i = R(x, y_i)$,

and normalizes rewards within the group:

$$\bar{R} = \frac{1}{K}\sum_{i=1}^{K} R_i, \qquad A_i = \frac{R_i - \bar{R}}{\sigma_R + \varepsilon},$$

where σ_R is the within-group reward standard deviation. All correct responses in the group share positive advantage A^+, while incorrect responses share A^-. Thus, GRPO inherits RLVR's indifference to how probability mass is allocated within the correct set. Full GRPO details in Appendix B.1.

3 Structural Causes of Diversity Collapse in RLVR

We identify a structural limitation of RLVR objectives: they are indifferent to how probability mass is distributed among correct solutions. Combined with optimization dynamics, this indifference leads to systematic diversity collapse.

3.1 Objective Indifference

The RLVR objective (Eq. 1) specifies what the policy should produce (correct outputs) but not how probability mass should be distributed among multiple correct solutions. The following theorem formalizes this observation.

Theorem 3.1 (Stationary Points of RLVR)**.** *Let π be any policy such that $\mathrm{supp}(\pi(\cdot \mid x)) \subseteq \mathcal{Y}^+(x)$, i.e., π places all probability mass on correct outputs. Then π is a stationary point of the RLVR objective (Eq. 1).*

Proof. The RLVR objective is $J(\pi) = \sum_{y\in\mathcal{Y}^+(x)} \pi(y \mid x)$. For any policy with $\mathrm{supp}(\pi(\cdot \mid x)) \subseteq \mathcal{Y}^+(x)$, we have $J(\pi) = 1$, the maximum possible value. Since J is linear in π and the constraint set

$$\left\{ \pi \;\middle|\; \sum_{y} \pi(y \mid x) = 1,\ \pi(y \mid x) \geq 0 \right\}$$

is a simplex, any policy achieving $J(\pi) = 1$ is a global maximizer of the RLVR objective. □

Remark 3.1 (Implication of Theorem 3.1)**.** *Theorem 3.1 implies that the following policies are all stationary points of RLVR with identical objective value $J(\pi) = 1$:*

- ***Collapsed:*** *$\pi(y_1 \mid x) = 1$ for some $y_1 \in \mathcal{Y}^+(x)$, and $\pi(y \mid x) = 0$ otherwise.*
- ***Uniform-correct:*** *$\pi(y \mid x) = \frac{1}{|\mathcal{Y}^+(x)|}$ for all $y \in \mathcal{Y}^+(x)$.*
- ***Any mixture:*** *Any distribution supported exclusively on $\mathcal{Y}^+(x)$.*

Thus, the RLVR objective (and by extension GRPO and other variants) provides no gradient signal to distinguish among these policies once all mass is on correct outputs. The choice among stationary points is determined entirely by optimization dynamics, not by the objective itself.

3.2 The Collapse Mechanism

Section 3.1 established that RLVR objectives do not specify how probability mass should be allocated among correct responses. We now show that, under GRPO and related group-relative policy-gradient methods, on-policy sampling resolves this ambiguity by favoring concentration: sampling imbalances are amplified across updates, concentrating mass on a subset of correct responses. We formalize this as a four-step cycle and validate each step in controlled experiments where the full policy distribution is observable (Appendix D); proofs are provided in Appendix C.1.

3.2.1 Setup

Consider training a policy π_θ initialized from π_{ref}. For a prompt x, let $\mathcal{Y}$ be the finite output space, partitioned into correct responses $\mathcal{Y}^+ = \{y_1, \ldots, y_m\}$ and incorrect responses $\mathcal{Y}^-$. Under GRPO, all correct responses share advantage A^+ and all incorrect responses share A^-. We analyze a fixed prompt x to isolate the within-prompt collapse mechanism.

3.2.2 SAMPLING BIAS

The GRPO objective involves an expectation over the policy:

$$\mathcal{L}(\theta) = \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}\left[A(y)\cdot\log\pi_\theta(y|x)\right]. \tag{3}$$

Since this expectation is intractable, GRPO approximates it via Monte Carlo: drawing K i.i.d. rollouts from $\pi_\theta(\cdot|x)$ and computing the empirical gradient. Let N_y denote the number of samples from solution y among the K rollouts.

Proposition 3.1 (Sampling Bias)**.** *Under on-policy sampling, solutions with higher probability are sampled more frequently. For K i.i.d. rollouts from $\pi_\theta(\cdot|x)$, the sample count N_y follows:*

$$N_y \sim \text{Binomial}(K, \pi_\theta(y|x)), \quad \mathbb{E}[N_y] = K\cdot\pi_\theta(y|x). \tag{4}$$

Proof. Each rollout independently produces solution y with probability $\pi_\theta(y|x)$. The count of successes in K independent Bernoulli trials is binomial. □

Higher-probability solutions are therefore overrepresented in sampled batches. Even under a uniform policy, finite-batch multinomial sampling can produce unequal counts across correct responses; Proposition C.1 formalizes this sampling-induced symmetry breaking. Appendix D.2 empirically validates this effect in the controlled environment: under uniform initialization, GRPO amplifies whichever correct response is overrepresented, leading to collapse to an arbitrary winner. The key question is therefore whether the subsequent gradient step corrects this imbalance or amplifies it.

3.2.3 GRADIENT AMPLIFICATION

Since sampling produces unequal counts $\{N_y\}$, we now examine how this imbalance enters the gradient update. The GRPO gradient over the batch is:

$$\nabla_\theta\mathcal{L} = \sum_{i=1}^{K} A_i\cdot\nabla_\theta\log\pi_\theta(y_i|x). \tag{5}$$

Grouping by solution, the gradient contribution of each correct $y\in\mathcal{Y}^+$ is: $N_y\cdot A^+\cdot\nabla_\theta\log\pi_\theta(y|x)$.

Theorem 3.2 (Gradient Amplification)**.** *Under GRPO with on-policy sampling, the expected gradient contribution from a correct solution $y\in\mathcal{Y}^+$ is:*

$$\mathbb{E}\left[\sum_{i:y_i=y} A^+\nabla_\theta\log\pi_\theta(y|x)\right] = A^+\cdot K\cdot\pi_\theta(y|x)\cdot\nabla_\theta\log\pi_\theta(y|x). \tag{6}$$

Thus, expected gradient contribution scales with current probability: high-probability solutions receive larger updates, while solutions with $\pi_\theta(y\mid x)\ll 1/K$ are likely unsampled and lose mass through softmax renormalization.

3.2.4 DIVERGENCE OF PROBABILITY RATIOS

The combination of sampling bias and gradient amplification causes probability ratios to diverge over training.

Theorem 3.3 (Compounding Divergence)**.** *Let π_t denote the policy after t updates. Under GRPO, for two correct solutions $y_1, y_2\in\mathcal{Y}^+(x)$ with $\pi_0(y_1|x) > \pi_0(y_2|x)$, the expected change in their log-probability ratio at step t is:*

$$\mathbb{E}\left[\Delta\log\frac{\pi_t(y_1\mid x)}{\pi_t(y_2\mid x)}\right] = \eta\cdot A^+\cdot K\cdot\big(\pi_t(y_1|x)-\pi_t(y_2|x)\big) > 0, \tag{7}$$

where η is the effective learning rate. The log-ratio is a submartingale with positive drift that increases as the probability gap widens.

This feedback causes small initial asymmetries to compound: the dominant solution receives larger updates, increasing its probability and expected sample count in subsequent batches, widening the gap and reinforcing the collapse.

3.2.5 IRREVERSIBLE PRUNING

The collapse becomes irreversible once minority solutions fall below the sampling threshold.

Corollary 3.1 (Sampling Threshold). *When $\pi_t(y|x) \ll 1/K$, the probability that solution y receives at least one sample in a batch of size K is:*

$$P(N_y \geq 1) = 1 - (1 - \pi_t(y|x))^K \approx K \cdot \pi_t(y|x) \ll 1.$$

Below this threshold, solution y receives effectively zero gradient signal and cannot recover under on-policy sampling.

The four steps trace a complete trajectory: (i) initial asymmetries arise from sampling bias, (ii) GRPO amplifies these asymmetries into the gradient, (iii) the updated policy increases the probability gap, which worsens sampling imbalance in subsequent iterations, and (iv) once low-probability solutions fall below the sampling threshold, recovery becomes unlikely. This mechanism explains the empirical Pass@1 versus Pass@K trade-off (Figure 9): GRPO improves single-sample accuracy while reducing multi-sample coverage.

3.3 CONTROLLED EXPERIMENT

To validate the proposed collapse mechanism, we construct a controlled RLVR environment with a softmax policy over a finite discrete output space, of which three outputs are correct and the remainder are incorrect. This makes the full policy distribution exactly observable throughout training (details in Appendix D). Figure 1 confirms all four stages of the collapse cycle; for details see Appendix D.1.

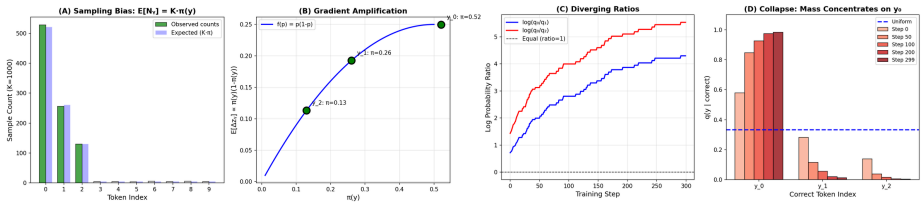

Figure 1: Controlled verification of the GRPO collapse mechanism in the toy RLVR environment. (A) On-policy sampling over-represents high-probability correct tokens, matching $E[N_y] = K \cdot \pi(y)$. (B) Because gradient contribution scales with sample count, dominant correct tokens receive larger expected updates. (C) The log-ratio between dominant and minority correct tokens increases over training, confirming compounding divergence. (D) Probability mass eventually concentrates on the initially preferred correct token, illustrating mode collapse within the correct set.

4 THE OPTIMAL POLICY

Section 3.2 showed that GRPO can systematically collapse toward dominant correct solutions. We now ask: *what should the optimal policy look like when multiple correct solutions exist?* We formalize this question through the conditional distribution over correct outputs and characterize a policy that preserves both correctness and within-correct diversity.

4.1 CONDITIONAL DISTRIBUTION ON CORRECT OUTPUTS

Conditional Correct Distribution

For a policy π, define the correctness mass and conditional distribution over correct outputs:

$$Z_\pi(x) = \sum_{y \in \mathcal{Y}^+(x)} \pi(y \mid x), \qquad q_\pi(y \mid x) = \frac{\pi(y \mid x)}{Z_\pi(x)}. \tag{8}$$

Here $Z_\pi(x)$ is the total probability that the policy produces a correct output. While binary RLVR optimizes $Z_\pi(x)$, it leaves the structure of $q_\pi(\cdot \mid x)$ unconstrained. Diversity collapse corresponds to q_π concentrating on a small subset of $\mathcal{Y}^+(x)$.

4.2 THE UNIFORM-CORRECT POLICY

Uniform-Correct Policy (UCP)

A policy $\pi^\star$ is *uniform-correct* if:

1. $\pi^\star(y \mid x) = 0$ for all $y \notin \mathcal{Y}^+(x)$;
2. $q_{\pi^\star}(y \mid x) = 1/|\mathcal{Y}^+(x)|$ for all $y \in \mathcal{Y}^+(x)$.

This policy places all probability mass on correct outputs and distributes it uniformly among them. We characterize this structure through two complementary criteria: conditional entropy, which resolves the within-correct indifference of the binary RLVR objective, and robustness to shifts that disable an unknown bounded subset of correct outputs. We assume that $\mathcal{Y}$ is finite and $0 < |\mathcal{Y}^+(x)| < |\mathcal{Y}|$. Proofs are given in Appendix C.3.

Theorem 4.1 (Conditional-Entropy Optimality)**.** *Among policies maximizing the binary RLVR objective, the Uniform-Correct Policy uniquely maximizes the conditional entropy:*

$$\pi^\star = \arg \max_{\pi:\, Z_\pi(x)=1} H(q_\pi(\cdot \mid x)) \,.$$

Theorem 4.2 (Robustness under Distribution Shift)**.** *Let $1 \le s < |\mathcal{Y}^+(x)|$, and suppose a distribution shift disables an unknown subset $S \subset \mathcal{Y}^+(x)$ of correct outputs, where $|S| \le s$. Consider the robust objective:*

$$\max_{\pi} \min_{S:\, |S| \le s} \sum_{y \in \mathcal{Y}^+(x) \setminus S} \pi(y \mid x). \tag{9}$$

Then the Uniform-Correct Policy uniquely maximizes the worst-case retained probability mass.

Theorem 4.1 resolves RLVR's indifference among correct solutions by selecting the maximum-conditional-entropy policy, while Theorem 4.2 shows that the same policy is minimax-optimal when an unknown bounded subset of correct outputs becomes inapplicable. These results characterize the target policy; Section 5 operationalizes this target with an objective that jointly promotes correctness mass and within-correct uniformity.

5 UCPO: UNIFORM CORRECT POLICY OPTIMIZATION

Section 4 characterized the Uniform-Correct Policy as the target that resolves RLVR's within-correct indifference. We now derive an objective that targets this policy during training.

5.1 UCPO MOTIVATION

A natural way to counteract diversity collapse is to add a global entropy bonus to the RLVR objective. However, global entropy acts over the full output space: for every positive entropy coefficient, its Gibbs optimum assigns nonzero probability to incorrect outputs. Global entropy therefore couples diversity with a loss of correctness mass. Empirically, it yields small, coefficient-sensitive Pass@K gains (Appendix E). We instead propose *Uniform-Correct Policy Optimization* (UCPO), which augments correctness-mass maximization with a conditional uniformity penalty that encourages the policy to match the uniform distribution over correct responses. UCPO thereby promotes correctness while encouraging a balanced allocation within the correct set.

5.2 OBJECTIVE

For a prompt $x \sim \mathcal{D}$, let $\mathcal{Y}^+(x)$ be the set of correct responses. During training, we sample N rollouts from $\pi_\theta(\cdot|x)$. The correct rollouts $y \in \mathcal{Y}^+(x)$ are effectively samples from the conditional policy $q_\theta(\cdot|x)$:

$$q_\theta(y \mid x) = \frac{\pi_\theta(y \mid x)}{Z_\theta(x)}, \qquad Z_\theta(x) = \sum_{y \in \mathcal{Y}^+(x)} \pi_\theta(y \mid x)$$

where $Z_\theta(x)$ is the total probability mass that the model places on all the correct responses. We also define the uniform distribution over the correct set:

$$u(y \mid x) = \frac{1}{|\mathcal{Y}^+(x)|}.$$

We propose the *Uniform-Correct Policy Optimization* (UCPO) objective:

$$L_{\mathrm{UCPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}}\left[\underbrace{\log Z_\theta(x)}_{\text{correctness mass}} - \tau \underbrace{\mathrm{KL}(u(\cdot \mid x)\|q_\theta(\cdot \mid x))}_{\text{uniformity penalty}}\right] \tag{10}$$

Relative to GRPO, UCPO introduces a single additional term: the conditional uniformity penalty $\mathrm{KL}(u\|q_\theta)$. At $\tau = 0$, UCPO reduces to GRPO. For $\tau > 0$, it breaks GRPO's indifference to within-correct mass allocation and biases the policy toward the uniform-correct target. In particular, the Uniform-Correct policy is the unique optimum of the UCPO objective (Theorem 5.1). This modification incurs negligible additional computational overhead, since the practical reweighting uses the same verifier-accepted rollouts already used by GRPO. As τ increases, the penalty increasingly discourages concentration, pushing q_θ toward the uniform target established in Section 4.

Theorem 5.1 (Uniform-Correct Policy is the unique optimum of UCPO)**.** *Fix a prompt x and consider the UCPO objective:*

$$L_{\mathrm{UCPO}}(x) = \log Z_\theta(x) - \tau\,\mathrm{KL}\big(u(\cdot \mid x)\,\|\,q_\theta(\cdot \mid x)\big).$$

For any $\tau > 0$, the unique maximizer of $L_{\mathrm{UCPO}}(x)$ over policies $\pi_\theta(\cdot \mid x)$ is the Uniform-Correct Policy.

5.3 Gradient Estimation

UCPO counteracts collapse via its gradient structure. Although GRPO is indifferent to within-correct allocation at the objective level, its sampled update effectively weights correct responses according to the current conditional distribution q_θ. UCPO instead shifts gradient mass from dominant correct responses toward underrepresented ones.

Proposition 5.1 (Gradient Decomposition)**.** *For any prompt x with $\mathcal{Y}^+(x) \neq \emptyset$,*

$$\nabla_\theta L_{\mathrm{UCPO}}(x) = \sum_{y\in\mathcal{Y}^+(x)} \Big[(1-\tau)\,q_\theta(y \mid x) + \tau\,u(y \mid x)\Big]\nabla_\theta \log \pi_\theta(y \mid x). \tag{11}$$

The proof is provided in Appendix C.4. The first term can be estimated directly from the n correct rollouts. Exact importance sampling of the uniform term requires inverse sequence-probability weights, which are strongly sensitive to response length. We therefore use the length-normalized policy score (Appendix C.5)

$$s_i = \pi_\theta(y_i \mid x)^{1/T_i} = \exp\Big(\tfrac{1}{T_i}\log \pi_\theta(y_i \mid x)\Big), \qquad \bar{v}_i = \frac{s_i^{-1}}{\sum_{j=1}^n s_j^{-1}},$$

where T_i is the response length, giving the estimator

$$\widehat{g}_{\mathrm{UCPO}} = \sum_{i=1}^n w_i\,\nabla_\theta \log \pi_\theta(y_i \mid x), \qquad w_i = \frac{1-\tau}{n} + \tau\,\bar{v}_i.$$

Theorem 5.1 characterizes the ideal objective; the length-normalized weights avoid the exponential length sensitivity of raw sequence probabilities, replacing the exact uniform component with a length-dependent flattening of the conditional policy (Appendix C.5).

By construction, $\sum_i w_i = 1$. Consequently, underrepresented correct responses receive larger weights, directly counteracting the sampling bias identified in Section 3.2. Equivalently, UCPO redistributes GRPO's total advantage budget nA^+ according to w_i:

$$A_i^{\mathrm{UCPO}} = n \cdot A^+ \cdot w_i. \tag{12}$$

Proposition 5.2 (Mass Preservation)**.** $\sum_{i=1}^n A_i^{\mathrm{UCPO}} = \sum_{i=1}^n A_i^{\mathrm{GRPO}}$.

Thus, UCPO preserves the total advantage mass assigned to correct rollouts and reallocates it towards underrepresented correct responses. The proof is provided in Appendix C.6.

Table 1: Pass@1 and Pass@64 (%) across models and reasoning benchmarks. UCPO improves average Pass@64 across all three model families while remaining competitive at Pass@1. Best results are shown in bold.

DeepSeek-Distill-Qwen-1.5B

	AIME24		AIME25		AMC		MATH		OlympiadBench		Avg	
Method	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**
GRPO	19.06	63.33	**20.21**	46.67	55.27	87.95	71.57	**97.60**	42.63	78.52	**41.75**	74.81
KL-Cov	**19.64**	60.00	18.13	43.33	54.78	91.57	70.47	96.80	42.69	78.22	41.14	73.98
Ent-Reg	19.43	63.33	17.24	46.67	**56.12**	89.16	**72.23**	**97.60**	**43.03**	78.07	41.61	74.97
Ent-Adv	19.58	63.33	15.94	46.67	54.07	91.57	69.37	97.00	41.59	79.11	40.11	75.54
Pass@K	15.47	**66.67**	14.27	**53.33**	47.63	**93.98**	62.74	97.40	35.24	77.93	35.07	77.86
FGRPO	17.92	**66.67**	16.35	50.00	52.79	**93.98**	67.76	97.20	40.57	78.96	39.08	77.36
LAD	15.42	**66.67**	15.66	50.00	48.36	91.57	64.23	97.40	36.27	78.22	35.97	76.77
UCPO	19.01	**66.67**	18.49	**53.33**	52.97	**93.98**	67.93	97.20	40.15	**80.15**	39.71	**78.27**

DeepSeek-Distill-Qwen-7B

	AIME24		AIME25		AMC		MATH		OlympiadBench		Avg	
Method	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**
GRPO	27.71	70.00	22.08	53.33	61.86	93.98	77.99	98.20	47.28	82.96	47.38	79.69
KL-Cov	**31.15**	76.67	**23.07**	**56.67**	**63.42**	95.18	78.21	98.20	47.99	81.93	**48.77**	81.73
Ent-Reg	29.06	76.67	21.30	**56.67**	63.08	93.98	**78.31**	**99.00**	**48.26**	82.37	48.00	81.74
Ent-Adv	29.06	**80.00**	21.15	**56.67**	59.71	96.39	76.08	**99.00**	45.92	**83.56**	46.38	83.12
Pass@K	23.39	70.00	19.11	**56.67**	56.44	93.98	72.59	98.20	41.77	80.74	42.66	79.92
FGRPO	27.55	76.67	21.25	53.33	61.77	95.18	76.25	98.80	46.47	81.93	46.66	81.18
UCPO	30.16	**80.00**	21.51	**56.67**	62.58	**97.59**	76.45	**99.00**	47.39	83.26	47.62	**83.30**

Qwen2.5-7B

	AIME24		AIME25		AMC		MATH		OlympiadBench		Avg	
Method	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**	**@1**	**@64**
GRPO	14.27	43.33	**5.26**	40.00	44.26	90.36	**71.61**	95.40	34.85	72.74	**34.05**	68.37
KL-Cov	12.50	46.67	4.27	46.67	41.87	87.95	67.58	95.40	31.84	74.81	31.61	70.30
Ent-Reg	11.98	50.00	4.11	40.00	42.09	86.75	66.60	94.80	32.02	73.48	31.36	69.01
Ent-Adv	12.71	**60.00**	4.69	43.33	41.74	90.36	68.86	95.00	32.41	73.33	32.08	72.40
Pass@K	10.73	53.33	4.01	43.33	38.52	89.16	61.85	95.60	28.82	74.22	28.79	71.13
FGRPO	12.97	56.67	5.05	40.00	41.74	**91.57**	67.02	95.40	32.21	74.37	31.80	71.60
UCPO	**15.05**	56.67	5.21	**46.67**	**43.66**	**91.57**	70.74	**95.80**	**35.03**	**75.26**	33.94	**73.19**

6 EXPERIMENTS

6.1 SETUP

We train on 10K competition-style math problems from DeepScaleR (Luo et al., 2025; Liao et al., 2025), with 500 reserved for validation. We evaluate three publicly released models under the same RLVR pipeline: DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025), and Qwen2.5-7B (Team, 2024). Evaluation uses AIME 2024/2025 (Li et al., 2024), AMC 2023 (Li et al., 2024), MATH 500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024), reporting Pass@1 and Pass@64. We assess diversity from two complementary views: Pass@64 measures multi-sample coverage, while equation-level diversity measures variation among verifier-accepted completions. We compare UCPO with GRPO and five diversity-preserving baselines: objective entropy regularization (Ent-Reg), entropy-based advantage bonuses (Ent-Adv) (Cheng et al., 2025), covariance-based KL control (KL-Cov) (Cui et al., 2025), Pass@K-oriented training with analytical advantages (Pass@K Training) (Chen et al., 2025), and focal prompt-level advantage scaling (FGRPO) (Plyusov et al., 2026). We additionally compare with concurrent LAD (Li & Li, 2026) on DeepSeek-R1-Distill-Qwen-1.5B. Dataset and implementation details are in Appendix F.

6.2 MAIN RESULTS

Table 1 reports Pass@1 and Pass@64 across five benchmarks and three models (1.5B-7B parameters). Overall, UCPO achieves a stronger correctness-coverage trade-off: it obtains the highest average Pass@64 across all three model families while preserving Pass@1. In contrast, several diversity-oriented baselines improve Pass@K at a substantially larger cost to single-sample accuracy (Pass@1).

Relative to GRPO, UCPO improves average Pass@64 by +3.5–4.8 points across the three models, with gains reaching +10 points on AIME24 for DeepSeek-Distill-Qwen-7B, and +6.7 and +6.0 points on AIME25 and AMC, respectively, for DeepSeek-Distill-Qwen-1.5B.

Crucially, UCPO's gains are not obtained by sacrificing Pass@1. Ent-Adv, the strongest Pass@64 baseline on both 7B models, provides a strong comparison: on DeepSeek-Distill-Qwen-7B and Qwen2.5-7B, UCPO obtains higher average Pass@1 and Pass@64. On the 1.5B model, Ent-Adv retains only a slightly higher Pass@1 (40.11 vs. 39.71), but with a worse Pass@64 (75.54 vs. 78.27). Conversely, the strongest Pass@64 baseline on this model, Pass@K Training, comes close in coverage (77.86 vs. 78.27) but at a substantially lower Pass@1 (35.07 vs. 39.71). **Thus, UCPO achieves higher coverage while retaining substantially more single-sample accuracy, yielding a more favorable correctness-coverage trade-off.** To assess run-to-run variability, we repeat the Qwen2.5-7B experiments and report mean $\pm$ sample standard deviation in Table 7. UCPO achieves the highest average Pass@64 in both runs while better preserving Pass@1 than the diversity-oriented baselines. Compared with Ent-Adv, UCPO improves average Pass@1 by 1.79 percentage points across four benchmarks while also increasing average Pass@64 by 0.45 points.

Table 2: Equation-level diversity among correct solutions. UCPO achieves the highest diversity across benchmarks.

Benchmark	GRPO	KL-Cov	Ent-Reg	Ent-Adv	Pass@K	FGRPO	UCPO
AIME	0.150	0.179	0.157	0.186	0.167	0.168	**0.218**
AMC	0.197	0.176	0.155	0.188	0.190	0.192	**0.217**
MATH	0.173	0.156	0.140	0.159	0.182	0.171	**0.196**
Olympiad	0.208	0.185	0.163	0.193	0.210	0.203	**0.228**
Avg.	0.182	0.174	0.154	0.181	0.187	0.184	**0.215**

6.3 Diversity

We measure within-correct diversity using equation-level diversity (Table 2), defined as the fraction of mathematical expressions in a correct rollout that are unique relative to other correct rollouts for the same prompt (Hu et al., 2025). UCPO achieves the highest diversity across all benchmarks, improving over GRPO by up to 45% on AIME and 13% on MATH, while entropy-based methods do not reliably improve diversity within the correct set and in some cases reduce it. These results show that UCPO's Pass@K gains are accompanied by greater diversity within the correct set, consistent with the controlled experiments in Appendix D: GRPO collapses the conditional distribution over correct solutions, while UCPO maintains a more uniform allocation.

6.4 Pass@K Performance

Figure 7 shows Pass@K curves across all models and benchmarks. UCPO consistently improves performance at larger K while remaining competitive at Pass@1, particularly on AIME 2024/2025. The widening gap at larger K is consistent with broader coverage of correct responses rather than concentration on a few dominant solutions by redistributing gradient signal within the verifier-accepted set.

6.5 Cost and Hyperparameter Sensitivity

UCPO introduces a single hyperparameter τ beyond standard GRPO. Its training cost is effectively identical to GRPO (Appendix G.2). UCPO is also robust across a moderate range of τ values, whereas global entropy regularization is more sensitive to its coefficient (Appendix G.3).

7 Conclusion

We identify a structural limitation of RLVR: its objective is indifferent to how probability mass is distributed among correct solutions, allowing on-policy training to collapse diversity and reduce Pass@K. We characterize the Uniform-Correct Policy as a principled target that resolves RLVR's within-correct indifference under conditional-entropy and robustness criteria. To achieve this structure, we proposed UCPO, to reweight gradients conditional on correctness, amplifying underrepresented correct solutions while tempering dominant ones. Across three model families and mathematical reasoning benchmarks, UCPO improves the correctness–coverage trade-off, achieving the highest average Pass@64 while preserving Pass@1, with gains of up to 10 percentage points on AIME24. These results highlight the value of explicitly shaping within-correct diversity in RLVR.

AI USE STATEMENT

We used generative AI tools to assist with refining the writing of the submission. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies a training-time method for preserving diversity in mathematical reasoning models. Maintaining diverse reasoning strategies may improve reliability in applications such as mathematical problem-solving and automated verification. The experiments use publicly available model checkpoints and mathematical reasoning datasets and do not involve human subjects or the collection of personal or sensitive data. We do not identify risks specific to UCPO beyond those associated with the broader deployment of increasingly capable language models. We follow the licenses and terms of use of the datasets and model checkpoints used in our experiments.

REPRODUCIBILITY STATEMENT

We provide the implementation of UCPO, together with training and evaluation script in the supplementary material. Appendix F documents the datasets, model checkpoints, training configuration, hyperparameters, and evaluation protocol used in our experiments. Together, the released code and the appendix specify the information needed to reproduce the main theoretical and empirical results reported in the paper.

REFERENCES

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

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

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

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

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Jiong Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, Ruiqi Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang,

Wenjun Gao, Wen-Xia Yu, Wentao Zhang, Wangding Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645:633 – 638, 2025. URL https://api.semanticscholar.org/CorpusID:275789950.

Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, et al. Agentic reinforced policy optimization. *arXiv preprint arXiv:2507.19849*, 2025.

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

Bofei Gao, Feifan Song, Zhe Yang, Zefan Cai, Yibo Miao, Qingxiu Dong, Lei Li, Chenghao Ma, Liang Chen, Runxin Xu, et al. Omni-math: A universal olympiad level mathematic benchmark for large language models. *arXiv preprint arXiv:2410.07985*, 2024.

Andre Wang He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting GRPO beyond distribution sharpening. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25548–25560, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1298. URL https://aclanthology.org/2025.emnlp-main.1298/.

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

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

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning. *arXiv preprint arXiv:2509.26209*, 2025.

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

Aayush Karan and Yilun Du. Reasoning with sampling: Your base model is smarter than you think. *arXiv preprint arXiv:2510.14901*, 2025.

Germán Kruszewski, Pierre Erbacher, Jos Rozen, and Marc Dymetman. Whatever remains must be true: Filtering drives reasoning in llms, shaping diversity. In *International Conference on Learning Representations*, volume 2026, pp. 85291–85316, 2026.

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

Wendi Li and Sharon Li. Lad: Learning advantage distribution for reasoning. *ACL 2026*, 2026.

Mengqi Liao, Xiangyu Xi, Ruinian Chen, Jia Leng, Yangen Hu, Ke Zeng, Shuai Liu, and Huaiyu Wan. Enhancing efficiency and exploration in reinforcement learning for llms. *arXiv preprint arXiv:2505.18573*, 2025.

Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Tianjun Zhang, Li Erran Li, et al. Deepscaler: Surpassing o1-preview with a 1.5 b model by scaling rl. *Notion Blog*, 2025.

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

Kangqi Ni, Zhen Tan, Zijie Liu, Pingzhi Li, and Tianlong Chen. Can grpo help llms transcend their pretraining origin?, 2025. URL https://arxiv.org/abs/2510.15990.

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

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

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

Yuda Song, Julia Kempe, and Remi Munos. Outcome-based exploration for llm reasoning. *arXiv preprint arXiv:2509.06941*, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Jiakang Wang, Runze Liu, Fuzheng Zhang, Xiu Li, and Guorui Zhou. Stabilizing knowledge, promoting reasoning: Dual-token constraints for rlvr, 2025a. URL https://arxiv.org/abs/2507.15778.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025b. URL https://arxiv.org/abs/2506.01939.

Fang Wu, Weihao Xuan, Ximing Lu, Mingjie Liu, Yi Dong, Zaid Harchaoui, and Yejin Choi. The invisible leash: Why rlvr may or may not escape its origin, 2026. URL https://arxiv.org/abs/2507.14843.

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

Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, Qian Liu, Ge Zhang, and Zejun Ma. First return, entropy-eliciting explore, 2025. URL https://arxiv.org/abs/2507.07017.

A Related Works

Prior works have documented diversity collapse in RLVR: fine-tuning improves Pass@1 but can reduce Pass@K by contracting the support of correct outputs and sharpening around high-likelihood base-model modes (Matsutani et al., 2025; Wu et al., 2026; He et al., 2025; Yue et al., 2025; Ni et al., 2025). A growing body of work aims to address this collapse, which we organize into analysis-driven methods that diagnose a collapse mechanism and derive a targeted fix, and methods that encourage diversity indirectly through token-level stochasticity or trajectory-level exploration.

A.1 Diagnosing and Addressing Diversity Collapse

Several concurrent works analyze mechanisms underlying diversity collapse and propose targeted interventions. Cui et al. (2025) attribute collapse to entropy decay driven by positive probability-advantage covariance, and propose covariance-based gradient clipping and KL regularization. Gai et al. (2025) show that selection and reinforcement biases in KL-regularized RLVR amplify base-policy imbalance within the correct set, and propose differential smoothing using sequence log-probabilities[1]. Plyusov et al. (2026) show that easy prompts with high success rates drive the strongest concentration of correctness mass, and propose focal advantage scaling to down-weight updates from these prompts. Kruszewski et al. (2026) propose filtering incorrect responses to shape diversity in RLVR reasoning. LAD (Li & Li, 2026) states that expected-advantage maximization distorts relative proportions among correct responses, and replaces it with an f-divergence objective that preserves the previous policy's proportions. By contrast, our work identifies that RLVR advantage-maximization objectives are indifferent to within-correct set allocation. The proposed UCPO reshapes the within-correct distribution toward uniformity by redistributing gradient mass among correct responses.

A.2 Diversity via Token-Level Regularization and Trajectory-Level Exploration

Entropy-based methods aim to mitigate collapse by maintaining stochasticity during training. Global entropy regularization acts over the full output space, which can increase probability mass on incorrect responses and conflict with correctness. More targeted token-level methods preserve local stochasticity through entropy bonuses, high-uncertainty token updates, or selective entropy regularization (Cheng et al., 2025; Wang et al., 2025b;a; Jiang et al., 2025). However, a model can remain uncertain at the token-level while still collapsing to a narrow set of final solutions (Wu et al., 2026). Another line of work promotes diversity by broadening trajectory-level exploration through adaptive rollout allocation, targeted exploration, or external data (Liao et al., 2025; Zheng et al., 2025; Dong et al., 2025). These methods are complementary to UCPO as they expand the set of reasoning trajectories explored during training, while UCPO operates at the conditional-distribution level, reshaping how gradient mass is allocated among correct responses once they are sampled.

B Background: GRPO Formulation

B.1 Group Relative Policy Optimization (GRPO)

GRPO (Shao et al., 2024) estimates RLVR policy gradients using group-relative baselines. For each prompt x, it samples K rollouts $\{y_i\}_{i=1}^{K} \sim \pi_{\text{old}}(\cdot \mid x)$ from the rollout policy, obtains rewards $R_i = R(x, y_i)$, and updates the current policy π_θ using importance ratios. Let $\sigma_R = \text{std}(\{R_j\}_{j=1}^{K})$ denote the within-group reward standard deviation. The group-relative baseline and advantage are

$$\bar{R} = \frac{1}{K}\sum_{i=1}^{K} R_i, \qquad A_i = \frac{R_i - \bar{R}}{\sigma_R + \varepsilon}. \tag{13}$$

[1]Their analysis is for KL-regularized RLVR, whereas our focus is on the RLVR objective itself. Throughout our analysis and experiments, we use $\beta = 0$, isolating collapse arising from the advantage-maximization term rather than reference regularization.

Following PPO (Schulman et al., 2017), GRPO uses the clipped surrogate objective:

$$L_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\{y_i\}_{i=1}^{K}\sim\pi_{\text{old}}(\cdot|x)}\left[\sum_{i=1}^{K}\min\left(r_i(\theta)A_i, \text{clip}(r_i(\theta), 1-\varepsilon, 1+\varepsilon)\, A_i\right)\right] \tag{14}$$

where $r_i(\theta) = \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)}$ is the importance ratio. All correct responses share advantage A^+, and all incorrect responses share A^-. Consequently, GRPO inherits the RLVR objective's indifference to how probability mass is allocated within the correct set.

C Theoretical Details and Proofs

C.1 The Diversity Collapse Mechanism

Sampling Bias: Connection to Pólya Urn and Preferential Attachment

The sampling bias in the collapse mechanism is mathematically related to classical models of preferential attachment: Consider **Pólya Urn Analogy** of an urn containing balls of M colors (representing M correct solutions), with initial counts proportional to $\{\pi_0(y_i)\}$. At each step, draw K balls with replacement, observe their colors, then add balls proportionally to the observed counts. This is the Pólya urn process, known to converge to a random limit where one color dominates with positive probability.

GRPO's on-policy sampling implements an analogous dynamic: the "urn composition" (policy distribution) is updated proportionally to observed samples (gradient updates), creating identical rich-get-richer dynamics.

Unequal sample counts are inevitable

Whether or not the policy initially assigns equal probability to correct solutions, the sample counts $\{N_y\}_{y\in\mathcal{Y}^+}$ will be unequal:

- **Deterministic case:** If $\pi(y_1|x) > \pi(y_2|x)$, then $\mathbb{E}[N_{y_1}] > \mathbb{E}[N_{y_2}]$. Higher-probability solutions are systematically over-represented.
- **Stochastic case:** Even if $\pi(y|x) = 1/m$ for all $y \in \mathcal{Y}^+$, the realized counts will differ due to sampling variance (see Proposition C.1 and empirical illustration in Appendix D).

Proposition C.1 (Stochastic Symmetry Breaking)**.** *Suppose $\pi_0(y|x) = 1/m$ for all $y \in \mathcal{Y}^+(x)$. For any finite batch size K, with positive probability, the sample counts $\{N_y\}$ are not all equal.*

Proof. Under a uniform policy over the m correct responses, the count vector satisfies

$$(N_{y_1}, \ldots, N_{y_m}) \sim \text{Multinomial}\left(K, \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)\right).$$

If K is not divisible by m, exact equality of all counts is impossible. If K is divisible by m, exact equality requires $N_{y_j} = K/m$ for every j, which occurs with probability

$$\mathbb{P}(N_{y_1} = \cdots = N_{y_m} = K/m) = \frac{K!}{((K/m)!)^m}\cdot\frac{1}{m^K} < 1.$$

Thus, unequal counts occur with positive probability in any finite batch. On this event, at least one correct response is overrepresented, i.e., there exists y^* such that $N_{y^*} > K/m$. Across repeated independent batches, the probability that no imbalance ever occurs is zero, so symmetry-breaking occurs almost surely over training. □

C.2 Proof of Theorem 3.2, Theorem 3.3 and Corollary 3.1

Theorem C.1 (Gradient Amplification)**.** *Under GRPO with on-policy sampling, the expected gradient contribution from a correct solution $y \in \mathcal{Y}^+$ is:*

$$\mathbb{E}\left[\sum_{i:y_i=y} A^+\nabla_\theta \log \pi_\theta(y|x)\right] = A^+ \cdot K \cdot \pi_\theta(y|x)\cdot\nabla_\theta \log \pi_\theta(y|x). \tag{15}$$

The expected contribution is proportional to current probability $\pi(y|x)$. Solutions with higher probability receive larger expected updates; solutions with $\pi(y|x) \ll 1/K$ likely remain unsampled and receive no gradient signal, causing their probability to decrease through softmax renormalization.

Proof. The GRPO gradient aggregated over the batch is:

$$\nabla_\theta \mathcal{L} = \sum_{i=1}^{K} A_i \cdot \nabla_\theta \log \pi_\theta(y_i|x).$$

For correct solutions ($y_i \in \mathcal{Y}^+$), the advantage is $A_i = A^+$. Grouping terms by solution identity:

$$\nabla_\theta \mathcal{L}\big|_{\mathcal{Y}^+} = \sum_{y \in \mathcal{Y}^+} N_y \cdot A^+ \cdot \nabla_\theta \log \pi_\theta(y|x),$$

where $N_y = \sum_{i=1}^{K} \mathbf{1}[y_i = y]$ counts occurrences of y in the batch.

Taking expectations and applying Proposition 3.1:

$$\begin{aligned}\mathbb{E}\left[N_y \cdot A^+ \cdot \nabla_\theta \log \pi_\theta(y|x)\right] &= A^+ \cdot \mathbb{E}[N_y] \cdot \nabla_\theta \log \pi_\theta(y|x) \\ &= A^+ \cdot K \cdot \pi_\theta(y|x) \cdot \nabla_\theta \log \pi_\theta(y|x).\end{aligned}$$

The expected gradient contribution is proportional to $\pi_\theta(y|x)$:

- High-probability solutions: $\pi_\theta(y|x)$ large $\Rightarrow$ large expected contribution
- Low-probability solutions: $\pi_\theta(y|x)$ small $\Rightarrow$ small expected contribution
- Very low probability ($\pi_\theta(y|x) \ll 1/K$): $\mathbb{E}[N_y] \ll 1$, so y is likely unsampled and contributes nothing

For an unsampled solution y_{unsamp}:

$$\Delta z_{y_{\text{unsamp}}} = 0 \quad \text{(no gradient signal)}$$

After the update, the softmax normalization means:

$$\pi'_\theta(y_{\text{unsamp}}|x) = \frac{\exp(z_{y_{\text{unsamp}}})}{\sum_{y'} \exp(z_{y'} + \Delta z_{y'})} < \pi_\theta(y_{\text{unsamp}}|x)$$

The unsampled solution's probability decreases as the probability mass shifts toward solutions that were sampled. This creates a multiplicative dynamic: solutions that are already more probable receive larger updates because they are more probable and sampled more often. After each gradient step, the probability distribution over $\mathcal{Y}^+$ becomes more concentrated: the gap between high- and low-probability correct solutions widens. Since $\nabla_\theta \log \pi_\theta(y|x)$ points in the direction that increases $\pi_\theta(y|x)$, solutions with higher current probability receive updates of larger expected magnitude, amplifying the initial disparity. □

Theorem C.2 (Compounding Divergence)**.** *Let π_t denote the policy after t updates. Under GRPO, for two correct solutions $y_1, y_2 \in \mathcal{Y}^+(x)$ with $\pi_0(y_1|x) > \pi_0(y_2|x)$, the expected change in their log-probability ratio at step t is:*

$$\mathbb{E}\left[\Delta \log \frac{\pi_t(y_1|x)}{\pi_t(y_2|x)}\right] = \eta \cdot A^+ \cdot K \cdot \big(\pi_t(y_1|x) - \pi_t(y_2|x)\big) > 0, \tag{16}$$

where η is the effective learning rate (and $\Delta f_t := f_{t+1} - f_t$ denotes the one-step change). The log-ratio is a submartingale with positive drift that increases as the probability gap widens.

Proof. Under softmax parameterization, $\pi_t(y|x) = \exp(z_y)/Z$ where $Z = \sum_{y'} \exp(z_{y'})$. The log-probability ratio is:

$$\log \frac{\pi_t(y_1|x)}{\pi_t(y_2|x)} = z_{y_1} - z_{y_2}.$$

From Theorem 3.2, the expected logit update is:

$$\mathbb{E}[\Delta z_y] = \eta \cdot A^+ \cdot K \cdot \pi_t(y|x),$$

where η absorbs learning rate and gradient-to-logit conversion factors.

The expected change in the log-ratio is:

$$\begin{aligned}\mathbb{E}\left[\Delta \log \frac{\pi_t(y_1|x)}{\pi_t(y_2|x)}\right] &= \mathbb{E}[\Delta z_{y_1}] - \mathbb{E}[\Delta z_{y_2}] \\ &= \eta \cdot A^+ \cdot K \cdot \pi_t(y_1|x) - \eta \cdot A^+ \cdot K \cdot \pi_t(y_2|x) \\ &= \eta \cdot A^+ \cdot K \cdot \big(\pi_t(y_1|x) - \pi_t(y_2|x)\big).\end{aligned}$$

When $\pi_t(y_1|x) > \pi_t(y_2|x)$, this quantity is strictly positive. Moreover, as training progresses:

- $\pi_t(y_1|x)$ increases (receiving larger updates)
- $\pi_t(y_2|x)$ decreases (receiving smaller updates, or losing mass to normalization)
- The gap $\pi_t(y_1|x) - \pi_t(y_2|x)$ widens
- The drift in Eq. 16 increases

This establishes that the log-ratio diverges with accelerating speed: the collapse is not merely linear but compounds over training. □

Corollary C.1 (Sampling Threshold)**.** *When $\pi_t(y|x) \ll 1/K$, the probability that solution y receives at least one sample in a batch of size K is:*

$$P(N_y \geq 1) = 1 - (1 - \pi_t(y|x))^K \approx K \cdot \pi_t(y|x) \ll 1$$

Below this threshold, solution y receives effectively zero gradient signal and cannot recover under on-policy sampling.

Proof. The probability that y is never sampled in K trials is:

$$P(N_y = 0) = (1 - \pi_t(y|x))^K$$

Thus:

$$P(N_y \geq 1) = 1 - (1 - \pi_t(y|x))^K$$

For small $\pi_t(y|x)$, using $(1-p)^K \approx e^{-Kp} \approx 1 - Kp$ for $Kp \ll 1$:

$$P(N_y \geq 1) \approx K\pi_t(y|x)$$

When $\pi_t(y|x) \ll 1/K$, we have $P(N_y \geq 1) \ll 1$, meaning y is almost never sampled.

Without samples, the gradient contribution from y is zero, and $\pi_t(y|x)$ can only decrease (due to the softmax normalization as other solutions increase). The expected waiting time for at least one sample is $1/\mathbb{P}(N_y \geq 1) = \Omega(1/(Kp))$. This makes the collapse irreversible. □

C.3 Proofs for uniform-correct optimality

Theorem C.3 (Conditional-Entropy Optimality)**.** *Among policies maximizing the binary RLVR objective, the Uniform-Correct Policy uniquely maximizes the conditional entropy:*

$$\pi^\star = \arg \max_{\pi:\, Z_\pi(x)=1} H(q_\pi(\cdot \mid x)).$$

Proof. Under binary rewards,

$$\mathbb{E}_{y\sim\pi}[R(x,y)] = \sum_{y \in \mathcal{Y}^+(x)} \pi(y \mid x) = Z_\pi(x),$$

so the RLVR-optimal policies are exactly those satisfying $Z_\pi(x) = 1$.

For any such policy, all probability mass lies on $\mathcal{Y}^+(x)$, and therefore $q_\pi(y \mid x) = \pi(y \mid x)$ on the correct set. Hence the remaining optimization is

$$\max_{q \in \Delta(\mathcal{Y}^+(x))} H(q).$$

Let u denote the uniform distribution on $\mathcal{Y}^+(x)$. Then

$$H(q) = \log |\mathcal{Y}^+(x)| - \mathrm{KL}(q \| u) \leq \log |\mathcal{Y}^+(x)|,$$

with equality if and only if $q = u$. Thus the unique maximizer is uniform over $\mathcal{Y}^+(x)$ and assigns zero mass elsewhere, i.e., $\pi = \pi^*$. □

Theorem C.4 (Robustness under Distribution Shift)**.** *Let $1 \leq s < |\mathcal{Y}^+(x)|$, and suppose a distribution shift disables an unknown subset $S \subset \mathcal{Y}^+(x)$ of correct outputs, where $|S| \leq s$. Consider the robust objective:*

$$\max_{\pi} \min_{S:\, |S| \leq s} \sum_{y \in \mathcal{Y}^+(x) \setminus S} \pi(y \mid x). \tag{17}$$

Then the Uniform-Correct Policy uniquely maximizes the worst-case retained probability mass.

Proof. Any probability mass assigned outside $\mathcal{Y}^+(x)$ cannot increase the retained correct mass, so an optimizer must satisfy $Z_\pi(x) = 1$. We may therefore restrict attention to distributions supported on $\mathcal{Y}^+(x)$. Let π be any distribution over $\mathcal{Y}^+(x)$ with $\sum_{y \in \mathcal{Y}^+(x)} \pi(y|x) = 1$. The adversary choosing S will select the s outputs with highest probability mass.

For the uniform distribution $\pi^*(y|x) = 1/|\mathcal{Y}^+|$, the worst-case retained mass is:

$$1 - \frac{s}{|\mathcal{Y}^+(x)|}.$$

For any non-uniform distribution, let

$$\pi_{\max} = \max_{y \in \mathcal{Y}^+} \pi(y|x) > \frac{1}{|\mathcal{Y}^+(x)|}.$$

The adversary can achieve retained mass strictly less than $1 - s/|\mathcal{Y}^+(x)|$ by targeting high-probability outputs.

Sorting probabilities in decreasing order:

$$\pi(y_{(1)}|x) \geq \pi(y_{(2)}|x) \geq \dots$$

the adversary removes $\{y_{(1)}, \dots, y_{(s)}\}$, achieving retained mass:

$$\sum_{j>s} \pi(y_{(j)}|x) = 1 - \sum_{j=1}^{s} \pi(y_{(j)}|x) \leq 1 - \frac{s}{|\mathcal{Y}^+(x)|},$$

with equality if and only if π is uniform. Thus, the uniform distribution over $\mathcal{Y}^+(x)$ uniquely maximizes the worst-case retained probability mass. □

C.4 Proof of Gradient Decomposition Proposition 5.1

Proposition C.2 (Gradient Decomposition, restated)**.** *Let x be a prompt with $\mathcal{Y}^+(x) \neq \emptyset$. Define the correctness mass $Z_\theta(x) = \sum_{y \in \mathcal{Y}^+(x)} \pi_\theta(y \mid x)$, the conditional distribution $q_\theta(y \mid x) = \pi_\theta(y \mid x)/Z_\theta(x)$, and the uniform distribution $u(y \mid x) = 1/|\mathcal{Y}^+(x)|$. Then the gradient of the UCPO objective (excluding the reference KL term) is:*

$$\nabla_\theta L_{\mathrm{UCPO}}(x) = \sum_{y \in \mathcal{Y}^+(x)} \Big[(1-\tau)\, q_\theta(y \mid x) + \tau\, u(y \mid x)\Big] \nabla_\theta \log \pi_\theta(y \mid x).$$

Proof. We compute the gradients of each term in $L_{\mathrm{UCPO}}(x) = \log Z_\theta(x) - \tau\, \mathrm{KL}(u \| q_\theta)$ separately.

Step 1: Gradient of $\log Z_\theta(x)$. By the chain rule,

$$\begin{aligned}\nabla_\theta \log Z_\theta(x) &= \frac{1}{Z_\theta(x)} \nabla_\theta Z_\theta(x) \\ &= \frac{1}{Z_\theta(x)} \sum_{y \in \mathcal{Y}^+(x)} \nabla_\theta \pi_\theta(y \mid x) \\ &= \frac{1}{Z_\theta(x)} \sum_{y \in \mathcal{Y}^+(x)} \pi_\theta(y \mid x) \nabla_\theta \log \pi_\theta(y \mid x) \\ &= \sum_{y \in \mathcal{Y}^+(x)} q_\theta(y \mid x) \nabla_\theta \log \pi_\theta(y \mid x).\end{aligned}$$

Step 2: Gradient of $\mathrm{KL}(u \| q_\theta)$. Expanding the KL divergence,

$$\begin{aligned}\mathrm{KL}(u \| q_\theta) &= \sum_{y \in \mathcal{Y}^+(x)} u(y \mid x) \log \frac{u(y \mid x)}{q_\theta(y \mid x)} \\ &= \sum_{y \in \mathcal{Y}^+(x)} u(y \mid x) \log u(y \mid x) - \sum_{y \in \mathcal{Y}^+(x)} u(y \mid x) \log q_\theta(y \mid x).\end{aligned}$$

The first sum is a constant (entropy of u), so

$$\nabla_\theta \mathrm{KL}(u \| q_\theta) = - \sum_{y \in \mathcal{Y}^+(x)} u(y \mid x) \nabla_\theta \log q_\theta(y \mid x).$$

Step 3: Gradient of $\log q_\theta(y \mid x)$. Since $q_\theta(y \mid x) = \pi_\theta(y \mid x) / Z_\theta(x)$,

$$\nabla_\theta \log q_\theta(y \mid x) = \nabla_\theta \log \pi_\theta(y \mid x) - \nabla_\theta \log Z_\theta(x).$$

Step 4: Combining. Substituting Step 3 into Step 2:

$$\begin{aligned}\nabla_\theta \mathrm{KL}(u \| q_\theta) &= - \sum_y u(y) \Big[\nabla_\theta \log \pi_\theta(y) - \nabla_\theta \log Z_\theta(x) \Big] \\ &= - \sum_y u(y) \nabla_\theta \log \pi_\theta(y) + \nabla_\theta \log Z_\theta(x),\end{aligned}$$

where we used $\sum_y u(y) = 1$.

Therefore,

$$\nabla_\theta \big[- \tau \, \mathrm{KL}(u \| q_\theta) \big] = \tau \sum_y u(y) \nabla_\theta \log \pi_\theta(y) - \tau \nabla_\theta \log Z_\theta(x).$$

Step 5: Full gradient. Adding the gradients from Steps 1 and 4:

$$\begin{aligned}\nabla_\theta L_{\mathrm{UCPO}}(x) &= \nabla_\theta \log Z_\theta(x) + \tau \sum_y u(y) \nabla_\theta \log \pi_\theta(y) - \tau \nabla_\theta \log Z_\theta(x) \\ &= (1 - \tau) \nabla_\theta \log Z_\theta(x) + \tau \sum_y u(y) \nabla_\theta \log \pi_\theta(y) \\ &= (1 - \tau) \sum_y q_\theta(y) \nabla_\theta \log \pi_\theta(y) + \tau \sum_y u(y) \nabla_\theta \log \pi_\theta(y) \\ &= \sum_{y \in \mathcal{Y}^+(x)} \Big[(1 - \tau) q_\theta(y) + \tau u(y) \Big] \nabla_\theta \log \pi_\theta(y).\end{aligned}$$

□

C.5 Derivation of the Length-Normalized Inverse-Score Weights

From Proposition 5.1, the ideal UCPO gradient can be written as

$$\nabla_\theta L_{\mathrm{UCPO}} = (1 - \tau) \, \mathbb{E}_{y \sim q_\theta} [\nabla_\theta \log \pi_\theta(y \mid x)] + \tau \, \mathbb{E}_{y \sim u} [\nabla_\theta \log \pi_\theta(y \mid x)]. \tag{18}$$

We estimate these terms from the n correct rollouts $\{y_i\}_{i=1}^{n} \sim q_\theta(\cdot \mid x)$ among the N rollouts sampled from $\pi_\theta(\cdot \mid x)$.

The first term is an expectation under q_θ and is estimated directly:

$$\mathbb{E}_{y\sim q_\theta}[\nabla_\theta \log \pi_\theta(y \mid x)] \approx \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log \pi_\theta(y_i \mid x). \tag{19}$$

The second term is an expectation under the uniform distribution u, while the available correct rollouts are sampled from q_θ.

Ideal importance weights. For any function f,

$$\mathbb{E}_{y\sim u}[f(y)] = \mathbb{E}_{y\sim q_\theta}\left[\frac{u(y \mid x)}{q_\theta(y \mid x)} f(y)\right].$$

Since $u(y \mid x) = 1/|\mathcal{Y}^+(x)|$ is constant over the correct set,

$$\frac{u(y \mid x)}{q_\theta(y \mid x)} \propto \frac{1}{q_\theta(y \mid x)} \propto \frac{1}{\pi_\theta(y \mid x)}.$$

Thus, exact importance sampling of the uniform term would use inverse sequence-probability weights.

Length-normalized practical weights. For an autoregressive completion,

$$\pi_\theta(y \mid x) = \prod_{t=1}^{T(y)} \pi_\theta(y_t \mid x, y_{<t}),$$

so raw inverse sequence probabilities are strongly sensitive to response length. We therefore use the length-normalized policy score

$$s_\theta(y \mid x) = \pi_\theta(y \mid x)^{1/T(y)} = \exp\left(\frac{1}{T(y)} \log \pi_\theta(y \mid x)\right). \tag{20}$$

For each sampled correct rollout, let

$$s_i = s_\theta(y_i \mid x), \qquad v_i = s_i^{-1}, \qquad \bar{v}_i = \frac{v_i}{\sum_{j=1}^{n} v_j}.$$

Because the samples are drawn from q_θ, these self-normalized inverse-score weights correspond, for fixed θ, to the induced target

$$\widetilde{u}_\theta(y \mid x) \propto q_\theta(y \mid x) s_\theta(y \mid x)^{-1} \propto \pi_\theta(y \mid x)^{(T(y)-1)/T(y)}.$$

Thus, the practical weighting replaces exact uniformity with a length-dependent flattening of the conditional policy, avoiding the exponential length sensitivity of raw inverse sequence probabilities.

The corresponding practical update is

$$\widehat{g}_{\mathrm{UCPO}} = \sum_{i=1}^{n} w_i \nabla_\theta \log \pi_\theta(y_i \mid x),$$

where

$$w_i = \frac{1-\tau}{n} + \tau\, \bar{v}_i. \tag{21}$$

Correct responses with lower length-normalized policy scores receive larger weights, shifting gradient mass away from dominant responses and toward underrepresented ones.

C.6 Proofs for Proposition 5.2 and Theorem 5.1

Proposition C.3 (Mass Preservation, restated)**.** *Let A_{base} be the GRPO advantage for correct responses, and let n be the number of correct rollouts. Define UCPO advantages as $A_i^{\text{UCPO}} = A_{\text{base}} \cdot n \cdot w_i^{\text{blend}}$ where*

$$w_i^{\text{blend}} = \frac{1-\tau}{n} + \tau \cdot \frac{v_i}{\sum_{j=1}^{n} v_j}.$$

Then:

$$\sum_{i=1}^{n} A_i^{\text{UCPO}} = \sum_{i=1}^{n} A_i^{\text{GRPO}} = n \cdot A_{\text{base}}.$$

Proof. We verify that the blended weights sum to one:

$$\begin{aligned}
\sum_{i=1}^{n} w_i^{\text{blend}} &= \sum_{i=1}^{n} \left(\frac{1-\tau}{n} + \tau \cdot \frac{v_i}{\sum_{j=1}^{n} v_j} \right) \\
&= (1-\tau) \cdot \frac{n}{n} + \tau \cdot \frac{\sum_{i=1}^{n} v_i}{\sum_{j=1}^{n} v_j} \\
&= (1-\tau) + \tau \\
&= 1.
\end{aligned}$$

Therefore,

$$\sum_{i=1}^{n} A_i^{\text{UCPO}} = A_{\text{base}} \cdot n \cdot \sum_{i=1}^{n} w_i^{\text{blend}} = A_{\text{base}} \cdot n \cdot 1 = n \cdot A_{\text{base}}.$$

Standard GRPO assigns A_{base} to each of the n correct responses, so $\sum_i A_i^{\text{GRPO}} = n \cdot A_{\text{base}}$. □

Theorem C.5 (Uniform-correct policy uniquely maximizes the UCPO objective)**.** *For a fixed prompt x, consider the UCPO objective without the reference KL term:*

$$L_{\text{UCPO}}(x) = \log Z_\theta(x) - \tau \, \text{KL}\big(u(\cdot \mid x) \,\|\, q_\theta(\cdot \mid x)\big),$$

where

$$Z_\theta(x) = \sum_{y \in \mathcal{Y}^+(x)} \pi_\theta(y \mid x), \qquad q_\theta(y \mid x) = \frac{\pi_\theta(y \mid x)}{Z_\theta(x)} \quad \text{for } y \in \mathcal{Y}^+(x),$$

and

$$u(y \mid x) = \frac{1}{|\mathcal{Y}^+(x)|} \quad \text{for } y \in \mathcal{Y}^+(x).$$

For any $\tau > 0$, the unique maximizer of $L_{\text{UCPO}}(x)$ over policies $\pi_\theta(\cdot \mid x)$ is the uniform-correct policy

$$\pi^*(y \mid x) = \begin{cases} \frac{1}{|\mathcal{Y}^+(x)|}, & y \in \mathcal{Y}^+(x), \\ 0, & y \notin \mathcal{Y}^+(x). \end{cases}$$

Proof. For any policy $\pi_\theta(\cdot \mid x)$, since $\pi_\theta(\cdot \mid x)$ is a probability distribution over the full output space,

$$Z_\theta(x) = \sum_{y \in \mathcal{Y}^+(x)} \pi_\theta(y \mid x) \leq 1.$$

Therefore,

$$\log Z_\theta(x) \leq 0,$$

with equality if and only if $Z_\theta(x) = 1$, i.e., if and only if

$$\pi_\theta(y \mid x) = 0 \qquad \text{for all } y \notin \mathcal{Y}^+(x).$$

Next, by non-negativity of KL divergence,

$$\text{KL}\big(u(\cdot \mid x) \,\|\, q_\theta(\cdot \mid x)\big) \geq 0,$$

so

$$-\tau \,\mathrm{KL}\big(u(\cdot \mid x) \,\|\, q_\theta(\cdot \mid x)\big) \le 0,$$

with equality if and only if

$$q_\theta(\cdot \mid x) = u(\cdot \mid x).$$

Combining the two inequalities,

$$L_{\mathrm{UCPO}}(x) = \log Z_\theta(x) - \tau \,\mathrm{KL}\big(u(\cdot \mid x) \,\|\, q_\theta(\cdot \mid x)\big) \le 0.$$

Equality holds if and only if both conditions are satisfied:

1. $Z_\theta(x) = 1$, so all probability mass lies on $\mathcal{Y}^+(x)$; and
2. $q_\theta(\cdot \mid x) = u(\cdot \mid x)$, so the conditional distribution over correct responses is uniform.

These two conditions uniquely determine

$$\pi_\theta(y \mid x) = \begin{cases} \frac{1}{|\mathcal{Y}^+(x)|}, & y \in \mathcal{Y}^+(x), \\ 0, & y \notin \mathcal{Y}^+(x), \end{cases}$$

which is exactly the uniform-correct policy π^*. Hence π^* is the unique maximizer of $L_{\mathrm{UCPO}}(x)$. □

C.7 Choice of Divergence

The UCPO objective penalizes deviation from the uniform target using $\mathrm{KL}(u \,\|\, q_\theta)$ rather than $\mathrm{KL}(q_\theta \,\|\, u)$. While both divergences are minimized when $q_\theta = u$, their gradients differ in a way that is critical for escaping mode collapse. We derive the gradient of each and explain why $\mathrm{KL}(u \| q_\theta)$ is the appropriate choice.

C.7.1 Entropy Maximization: $\mathrm{KL}(q_\theta \,\|\, u)$

When u is uniform over $\mathcal{Y}^+(x)$, we have $\log u(y) = -\log |\mathcal{Y}^+(x)|$ for all $y \in \mathcal{Y}^+(x)$. Expanding the KL divergence:

$$\begin{aligned} \mathrm{KL}(q_\theta \| u) &= \sum_{y \in \mathcal{Y}^+(x)} q_\theta(y) \log \frac{q_\theta(y)}{u(y)} \\ &= \sum_y q_\theta(y) \log q_\theta(y) - \sum_y q_\theta(y) \log u(y) \\ &= -H(q_\theta) + \log |\mathcal{Y}^+(x)|. \end{aligned}$$

Thus, minimizing $\mathrm{KL}(q_\theta \| u)$ is equivalent to maximizing the entropy $H(q_\theta)$.

The entropy of q_θ is:

$$H(q_\theta) = -\sum_{y \in \mathcal{Y}^+(x)} q_\theta(y) \log q_\theta(y).$$

Taking the gradient with respect to θ:

$$\begin{aligned} \nabla_\theta H(q_\theta) &= -\sum_y \nabla_\theta \big[q_\theta(y) \log q_\theta(y) \big] \\ &= -\sum_y \big[\nabla_\theta q_\theta(y) \cdot \log q_\theta(y) + q_\theta(y) \cdot \nabla_\theta \log q_\theta(y) \big] \\ &= -\sum_y \nabla_\theta q_\theta(y) \cdot \log q_\theta(y) - \sum_y q_\theta(y) \nabla_\theta \log q_\theta(y). \end{aligned}$$

Since $\sum_y q_\theta(y) = 1$, we have $\sum_y \nabla_\theta q_\theta(y) = 0$. Using $\nabla_\theta q_\theta(y) = q_\theta(y)\nabla_\theta \log q_\theta(y)$:

$$\begin{aligned}\nabla_\theta H(q_\theta) &= -\sum_y q_\theta(y)\nabla_\theta \log q_\theta(y) \cdot \log q_\theta(y) - \sum_y q_\theta(y)\nabla_\theta \log q_\theta(y) \\ &= -\sum_y q_\theta(y)\big[1 + \log q_\theta(y)\big]\nabla_\theta \log q_\theta(y) \\ &= -\mathbb{E}_{y\sim q_\theta}\big[(1 + \log q_\theta(y))\nabla_\theta \log q_\theta(y)\big]. \end{aligned} \tag{22}$$

Since the gradient is weighted by q_θ, already-suppressed responses receive little corrective signal, making recovery from concentration difficult. This preserves the feedback loop identified in Section 3.2.

C.7.2 Uniform-Weighted Divergence: $\mathrm{KL}(u \,\|\, q_\theta)$

Expanding the KL divergence:

$$\begin{aligned}\mathrm{KL}(u\|q_\theta) &= \sum_{y\in\mathcal{Y}^+(x)} u(y) \log \frac{u(y)}{q_\theta(y)} \\ &= \sum_y u(y)\log u(y) - \sum_y u(y)\log q_\theta(y) \\ &= -H(u) - \sum_y u(y)\log q_\theta(y).\end{aligned}$$

Since $H(u)$ is constant with respect to θ:

$$-\mathrm{KL}(u\|q_\theta) = \sum_y u(y)\log q_\theta(y) + \text{const}.$$

Substituting $q_\theta(y) = \pi_\theta(y)/Z_\theta(x)$ where $Z_\theta(x) = \sum_{y\in\mathcal{Y}^+(x)} \pi_\theta(y)$:

$$\begin{aligned}-\mathrm{KL}(u\|q_\theta) &= \sum_y u(y)\big[\log \pi_\theta(y) - \log Z_\theta(x)\big] + \text{const} \\ &= \sum_y u(y)\log \pi_\theta(y) - \log Z_\theta(x) + \text{const}.\end{aligned}$$

Taking the gradient:

$$\nabla_\theta\big[-\mathrm{KL}(u\|q_\theta)\big] = \sum_y u(y)\nabla_\theta \log \pi_\theta(y) - \nabla_\theta \log Z_\theta(x). \tag{23}$$

For the second term, we use:

$$\begin{aligned}\nabla_\theta \log Z_\theta(x) &= \frac{1}{Z_\theta(x)} \sum_{y\in\mathcal{Y}^+(x)} \nabla_\theta \pi_\theta(y) \\ &= \sum_y \frac{\pi_\theta(y)}{Z_\theta(x)} \nabla_\theta \log \pi_\theta(y) \\ &= \sum_y q_\theta(y)\nabla_\theta \log \pi_\theta(y).\end{aligned}$$

Substituting:

$$\nabla_\theta\big[-\mathrm{KL}(u\|q_\theta)\big] = \mathbb{E}_{y\sim u}\big[\nabla_\theta \log \pi_\theta(y)\big] - \mathbb{E}_{y\sim q_\theta}\big[\nabla_\theta \log \pi_\theta(y)\big]. \tag{24}$$

The gradient consists of two terms:

- $\mathbb{E}_{y\sim u}[\nabla_\theta \log \pi_\theta(y)]$: Increases probability of all correct responses equally, regardless of their current probability.
- $-\mathbb{E}_{y\sim q_\theta}[\nabla_\theta \log \pi_\theta(y)]$: Decreases probability of responses in proportion to their current dominance.

Together, these terms redistribute gradient signal from overrepresented to underrepresented responses, breaking the self-reinforcing feedback loop.

C.7.3 Connection to UCPO

Combining $\mathrm{KL}(u\|q_\theta)$ with the correctness objective $\log Z_\theta(x)$ yields the UCPO gradient (Proposition 5.1):

$$\nabla_\theta L_{\mathrm{UCPO}} = \sum_{y\in\mathcal{Y}^+(x)} \left[(1-\tau)\, q_\theta(y) + \tau\, u(y)\right] \nabla_\theta \log \pi_\theta(y \mid x).$$

At $\tau = 0$, updates are fully q_θ-weighted (GRPO). As τ increases, updates shift toward uniform weighting, directly counteracting the sampling bias that drives collapse.

D Controlled Empirical Validation of Collapse Dynamics

Section 3.2 establishes that diversity collapse under RLVR arises from the interaction between objective indifference and optimization dynamics, which together induce a self-reinforcing concentration of probability mass on a subset of correct solutions. While this mechanism is derived analytically, its empirical manifestation in large-scale language models is difficult to observe directly due to partial observability and sequence-level complexity.

To isolate and validate the proposed mechanism, we construct a controlled RLVR environment in which the policy is fully observable and all quantities can be computed exactly. The goal of this experiment is to causally verify each component of the collapse mechanism under minimal assumptions.

Setup. We define a discrete output space of 20 tokens, among which a subset of 3 tokens are designated as correct:

$$\mathcal{Y}^+ = \{y_1, y_2, y_3\}, \qquad \mathcal{Y}^- = \mathcal{Y} \setminus \mathcal{Y}^+.$$

The policy $\pi_\theta(y)$ is parameterized as a softmax over logits z_y:

$$\pi_\theta(y) = \frac{\exp(z_y)}{\sum_{y'} \exp(z_{y'})}.$$

To reflect realistic pretrained biases, the initial policy assigns non-uniform probability mass within the correct set, while distributing the remaining mass across incorrect outputs. This ensures that symmetry is broken at initialization, allowing us to study how small imbalances evolve under training.

Training Dynamics. At each iteration:

- We sample K outputs from π_θ (on-policy sampling).
- Rewards are assigned deterministically:

$$R(y) = \mathbf{1}[y \in \mathcal{Y}^+].$$

- The policy is updated using either GRPO, UCPO, and Global entropy regularization.

Tracked Metrics. Since the full distribution is explicitly available, we analyze training dynamics through the following quantities:

(i) Sampling statistics. The number of times each output is sampled: N_y.

(ii) Conditional distribution over correct solutions. We define the normalized distribution restricted to correct outputs:

$$q_\theta(y \mid \mathcal{Y}^+) = \frac{\pi_\theta(y)}{\sum_{y'\in\mathcal{Y}^+} \pi_\theta(y')}, \quad y \in \mathcal{Y}^+.$$

(iii) Entropy of correct solutions. To measure diversity within the correct set:

$$H(q_\theta) = -\sum_{y\in\mathcal{Y}^+} q_\theta(y \mid \mathcal{Y}^+) \log q_\theta(y \mid \mathcal{Y}^+).$$

(iv) Total correctness mass. The probability assigned to correct outputs:

$$Z_\theta = \sum_{y \in \mathcal{Y}^+} \pi_\theta(y).$$

(v) Relative dominance. We monitor pairwise probability ratios between correct solutions to capture mode concentration:

$$\frac{\pi_\theta(y_i)}{\pi_\theta(y_j)}, \quad y_i, y_j \in \mathcal{Y}^+.$$

D.1 GRPO Induces Self-Reinforcing Collapse

We validate each step of the collapse mechanism (Section 3.2), with the corresponding evidence shown in Figure 1.

1. **Sampling bias:** Figure 1(a) shows that higher-probability solutions are sampled more frequently, with empirical counts closely matching theoretical expectation:

   $$\mathbb{E}[N_y] = K \cdot \pi_\theta(y).$$

   Even under near-uniform initialization (Appendix D.2), stochastic sampling produces unequal counts, confirming the inherent bias of on-policy estimation.

2. **Gradient amplification:** Figure 1(b) plots $f(p) = p(1-p)$, the function proportional to the expected gradient magnitude for a softmax policy. The green dots mark the initial probabilities of the three correct tokens. The key observation is that tokens already at higher probability receive a larger expected gradient update. Under GRPO, all correct samples receive the same normalized advantage A^+, so the total expected update to $\log \pi(y)$ scales as:

   $$\Delta \log \pi_\theta(y) \propto \eta \cdot A^+ \cdot K \cdot \pi_\theta(y).$$

   This confirms that GRPO amplifies sampling asymmetries rather than correcting them.

3. **Probability ratios diverge monotonically.** Figure 1(c) tracks the evolution of probability ratios between correct solutions. The log-ratio between dominant and minority solutions increases steadily over training:

   $$\log \frac{\pi_\theta(y_1)}{\pi_\theta(y_2)} \uparrow .$$

   This demonstrates the positive feedback loop predicted in Theorem 3.3, that the expected change in log-ratio is proportional to the difference in probabilities, which is always positive when $y_0 \geq y_1$.

4. **Collapse is irreversible once below the sampling threshold.** Figure 1(d) shows the conditional distribution

   $$q_\theta(y \mid \text{correct}) = \frac{\pi_\theta(y)}{Z_\theta}$$

   at five snapshots during training (steps $0, 50, 100, 200, 299$). The blue dashed line marks the uniform target $1/3$. By step 299, virtually all mass has migrated to y_0. The solutions y_1 and y_2 fall so far below the sampling threshold ($\mathbb{E}[N_y] \ll 1$) that they receive negligible gradient signal and can no longer recover. This confirms the irreversibility predicted in Corollary 3.1.

D.2 Collapse Is Not Limited to Skewed Initializations

In Figure 1, we use a skewed initialization of $4:2:1$ over correct tokens to make the collapse dynamics visually clear. However, even under near-uniform initialization, stochastic sampling produces unequal counts, confirming the inherent bias of on-policy estimation, consistent with Proposition 3.1. To verify this we evaluate three initialization profiles spanning the range from maximally symmetric to strongly biased. The reference policy assigns background mass 0.01 to each incorrect token (before normalization) across all profiles:

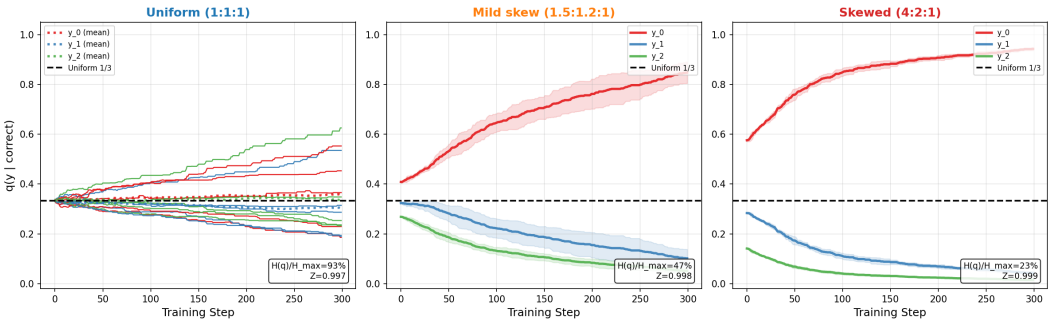

Figure 2: Collapse dynamics under GRPO in a controlled RLVR environment with three correct tokens and three initialization profiles. Each panel tracks the conditional distribution $q(y \mid \mathcal{Y}^+)$ over 300 training steps across 5 random seeds; the dashed blue line marks the uniform target $1/3$. Even under uniform initialization, stochastic sampling breaks symmetry and GRPO collapses to a single dominant token. With increasing initialization skew, collapse becomes faster and more deterministic. In all cases, total correctness mass Z remains high, showing that GRPO improves correctness while collapsing within-correct diversity.

- **Uniform (1:1:1):** $q_\theta(y_0) = q_\theta(y_1) = q_\theta(y_2) = 0.333$. Tests the null hypothesis that collapse requires an initial imbalance.
- **Mild skew (1.5:1.2:1):** $q_\theta(y_0) = 0.41,\ q_\theta(y_1) = 0.32,\ q_\theta(y_2) = 0.27$. Reflects the small asymmetry.
- **Skewed (4:2:1):** $q_\theta(y_0) = 0.57,\ q_\theta(y_1) = 0.29,\ q_\theta(y_2) = 0.14$.

We run 5 independent seeds per profile and track the conditional entropy $H(q_\theta)$ and correctness mass Z_θ throughout training. Results are shown in Figure 2. Under skewed and mild-skew initialization, the dominant token y_0 wins deterministically across all 5 seeds: GRPO always amplifies the initially more-probable correct token, confirming the collapse mechanism. Final conditional entropy is $H(q_\theta)/H_{\max} \approx 16\%$ (skewed) and 32% (mild skew), both far below the uniform target.

Under uniform initialization, collapse still occurs in every seed, but the winner is arbitrary: across 5 seeds, y_0 collapses to the dominant mode in 3 runs, y_1 in 1 run, and y_2 in 1 run. This is a direct consequence of Proposition C.1 (stochastic symmetry breaking): even when $\pi_0(y \mid x) = 1/|Y^+|$, finite-batch sampling can produce unequal counts with positive probability, and symmetry-breaking occurs almost surely over repeated sampling steps, seeding the positive feedback loop. GRPO then amplifies whichever token happened to be over-represented in the first batch.

These results establish that mode collapse is not an artifact of a skewed initialization, it is an intrinsic consequence of the interaction between on-policy sampling variance and GRPO's indifference to within-correct diversity. The skewed profile collapses faster and more predictably; the uniform profile collapses more slowly and to an arbitrary mode.

D.3 GRPO Induces Concentration within the Correct Set

We further analyze the training dynamics of GRPO, UCPO, and global entropy regularization in depth (Figure 3). As shown in Figure 3(a), GRPO drives the conditional distribution $q_\theta(y \mid \mathcal{Y}^+)$ toward a single dominant solution. Correspondingly, Figure 3(d) shows rapid decay in conditional entropy $H(q_\theta)$, while Figure 3(e) shows that total correctness mass Z remains high.

This indicates that GRPO improves correctness by concentrating probability mass rather than maintaining coverage over multiple valid solutions. As a result, diversity within the correct set collapses despite high sampling efficiency.

D.4 UCPO Preserves Diversity via Gradient Redistribution

UCPO modifies training dynamics by redistributing gradient signal within the correct set according to the reweighting mechanism derived in Section 5.

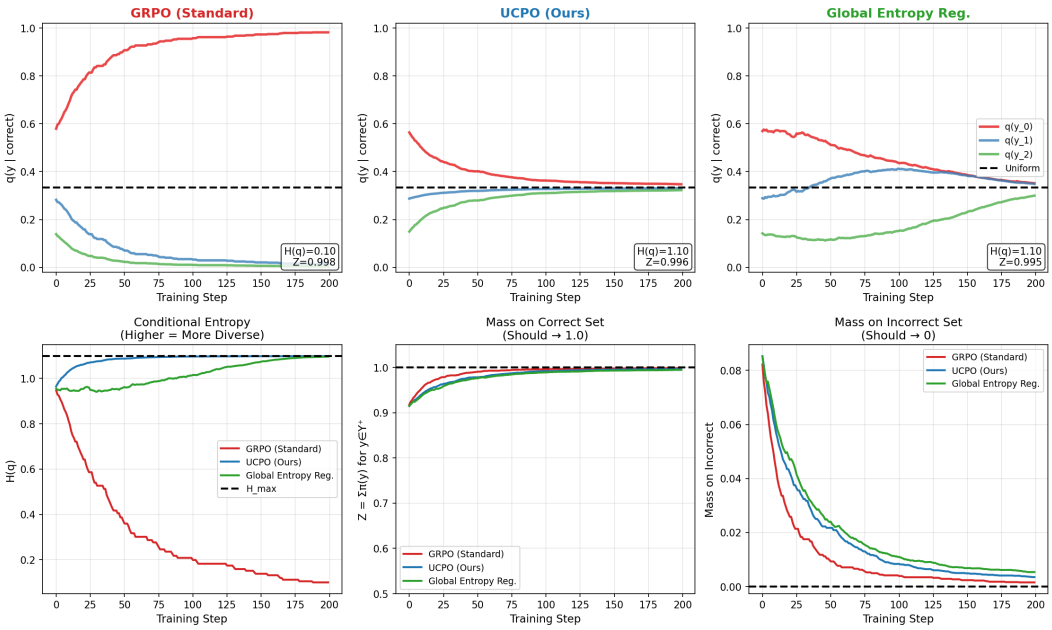

Figure 3: Training dynamics under GRPO, UCPO, and global entropy regularization. GRPO drives the conditional distribution over correct tokens toward a single dominant mode, causing conditional entropy to collapse. UCPO instead maintains a near-uniform distribution within the correct set while preserving total correctness mass. Global entropy regularization increases entropy only by assigning probability mass to incorrect tokens, demonstrating that global entropy acts on the wrong support.

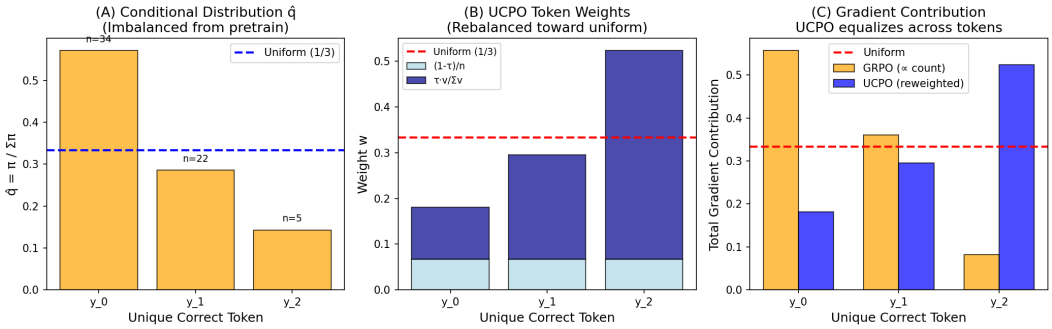

Figure 4: UCPO reweights gradient mass within the correct set. Starting from an imbalanced conditional distribution over correct tokens, UCPO assigns larger weight to underrepresented correct tokens and smaller weight to dominant ones. Unlike GRPO, whose total gradient contribution is proportional to sample count, UCPO approximately equalizes total contribution across unique correct tokens while preserving the total advantage mass.

As shown in Figure 3(a), UCPO maintains a balanced conditional distribution: $q_\theta(y \mid \mathcal{Y}^+) \approx$ uniform. Figure 3(d) shows that conditional entropy remains high throughout training, while Figure 3(e) shows that correctness mass Z is preserved.

To isolate the mechanism of UCPO, we further analyze gradient contributions per solution. Figure 4(b) shows that UCPO equalizes gradient contributions across correct solutions and removes the implicit bias toward dominant modes while preserving total gradient mass. In contrast, under GRPO contributions scale with sample count (Figure 4(a)), leading to dominance of frequently sampled solutions.

These results demonstrate that UCPO counteracts the sampling-induced imbalance by redistributing updates within the correct set, preventing concentration while maintaining correctness.

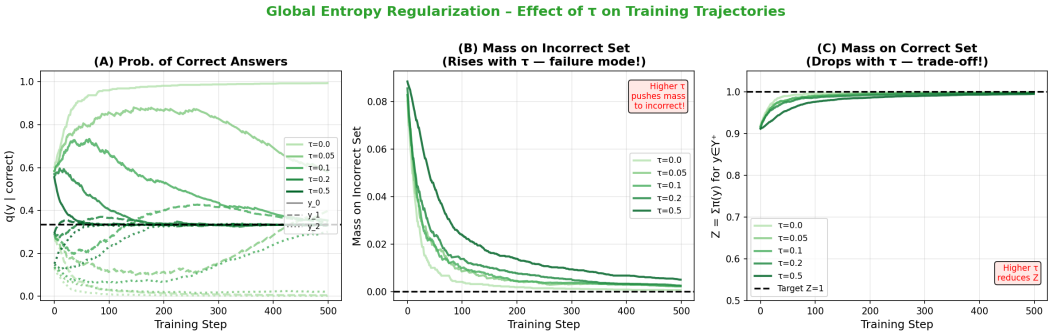

Figure 5: Effect of global entropy regularization strength in the controlled environment. (A) As the entropy coefficient τ_{ent} increases, the conditional distribution over correct tokens remains diverse but at a cost. (B) Mass on incorrect tokens grows with τ_{ent}, confirming that the entropy bonus incentivizes incorrect outputs equally. (C) Correctness mass Z_θ decreases as τ_{ent} increases, demonstrating the fundamental trade-off: global entropy regularization improves diversity only by sacrificing correctness, as it acts on the entire output space rather than within the correct set.

D.5 Global Entropy Regularization Increases Entropy on the Wrong Support

We analyze global entropy regularization in Figure 3 and Figure 5.

Figure 3(d) shows that entropy regularization increases overall entropy, but Figure 3(f) reveals that this increase is partly driven by probability mass assigned to incorrect outputs.

Figure 5(b) further shows that as τ_{ent} increases, the equilibrium mass on incorrect tokens grows monotonically. Since the entropy bonus is applied uniformly across all outputs, it provides equal incentive to increase the probability on incorrect and underrepresented correct solutions.

As a result, Figure 5(c) shows that improvements in entropy $H(q_\theta)$ come at the cost of reduced correctness mass Z_θ. This confirms that global entropy regularization increases entropy on the wrong support and cannot preserve diversity within the correct set.

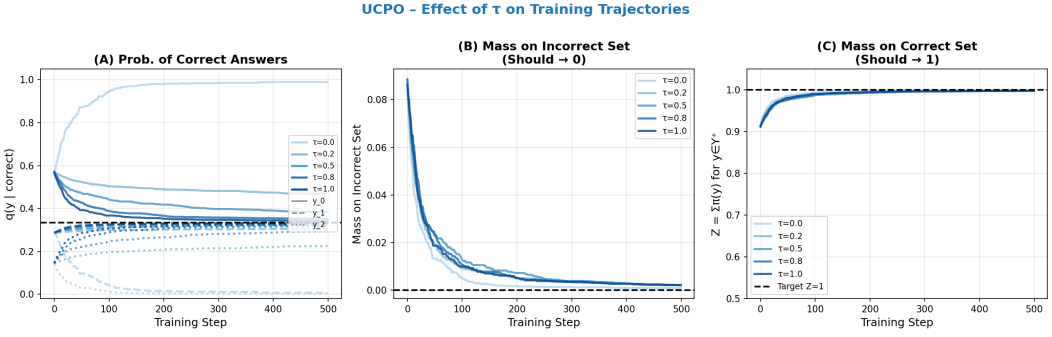

Figure 6: Effect of UCPO interpolation strength τ in the controlled environment. (A) Increasing τ progressively maintains diversity within the correct set, preventing concentration onto a single mode. (B) Mass on incorrect tokens remains negligible across all τ values, confirming that UCPO operates exclusively within the correct set. (C) Correctness mass Z_θ stays near 1.0 for all τ, demonstrating that UCPO achieves diversity without the correctness trade-off observed under global entropy regularization.

D.6 UCPO Maintains Correctness–Diversity Trade-off

Figure 6 shows that increasing τ improves diversity while preserving correctness. Unlike entropy regularization, the mass on incorrect outputs remains negligible across all τ, and Z_θ remains close to 1.0. This demonstrates that UCPO can increase diversity without incurring the correctness trade-off observed in global entropy methods.

E Why Global entropy regularization fails

A natural way to encourage diversity is to add a global entropy bonus to the RLVR objective:

$$J_{\text{ent}}(\pi) = \mathbb{E}_{y\sim\pi}[R(x,y)] + \tau_{\text{ent}} H(\pi(\cdot \mid x)). \tag{25}$$

For any finite $\tau_{\text{ent}} > 0$, the unique distributional optimum is the Gibbs policy

$$\pi_\tau(y \mid x) \propto \exp(R(x,y)/\tau_{\text{ent}}).$$

Under binary rewards, all correct outputs therefore receive equal probability, so the conditional distribution over correct outputs is uniform. However, the optimal policy also assigns total mass

$$1 - Z_{\pi_\tau}(x) = \frac{|\mathcal{Y}^-(x)|}{|\mathcal{Y}^+(x)|e^{1/\tau_{\text{ent}}} + |\mathcal{Y}^-(x)|} > 0 \tag{26}$$

to incorrect outputs for every finite $\tau_{\text{ent}} > 0$. Thus, global entropy reaches the Uniform-Correct Policy only in the limit $\tau_{\text{ent}} \to 0^+$. Unlike conditional uniformity, global entropy therefore couples diversity with a loss of correctness mass. This distinction surfaces in the optimization dynamics.

First, the entropy term is computed over the full output space $\mathcal{Y}$, including incorrect solutions, rather than within the correct set $\mathcal{Y}^+(x)$. The reward term pushes mass toward correct outputs, while the entropy term pushes mass toward all outputs indiscriminately. This creates a structural trade-off: diversity is encouraged on the wrong support. Empirically, global entropy regularization increases overall entropy but does not improve diversity within the correct set (Section 6.3), with much of the entropy increase arising from mass allocated to incorrect outputs (Appendix D.5).

Second, under GRPO with binary rewards, when all rollouts for a prompt are correct (or all incorrect), the group-normalized advantages are $A_i = 0$ for every rollout. In this case, the reward term contributes no gradient, and the update is driven entirely by the entropy bonus. The policy is then pushed to spread mass across all outputs, including incorrect ones, rather than preserving diversity within the correct set. This failure mode is particularly common for easier prompts, where the model achieves high accuracy early in training. Empirically, global entropy regularization yields small and inconsistent Pass@K gains and is sensitive to the entropy coefficient (Table 3). These observations motivate separating the two goals: maximizing correctness mass while directly controlling the conditional distribution over correct responses.

Table 3: Performance at Pass@1 and Pass@64 across benchmarks for different global entropy coefficients for Global Entropy Regularization on DeepSeek-R1-Distill-Qwen-1.5B. All other hyperparameters are held fixed.

	AIME		AMC		MATH		OlympiadBench		Avg	
τ_{ent}	@1	@64	@1	@64	@1	@64	@1	@64	@1	@64
0.0 (GRPO)	19.06	63.33	55.27	87.95	71.57	97.60	42.63	78.52	47.13	81.85
0.0005	19.79	53.33	54.84	92.77	69.45	97.20	41.69	77.93	46.44	80.31
0.001	19.43	63.33	56.12	89.16	72.23	97.60	43.03	78.07	47.70	82.04
0.003	19.38	60.00	54.71	90.36	71.45	97.20	42.58	77.93	47.03	81.37
0.005	19.48	63.33	55.25	90.36	72.08	97.20	43.86	77.04	47.67	81.98

F Implementation Details

F.1 Datasets

We use the RL training dataset from Liao et al. (2025), derived from DeepScaleR 40k corpus (Luo et al., 2025). The training pool consists of competition-style mathematical reasoning problems drawn

Parameter	Value	Parameter	Value
Pretrained Model	DeepSeek-R1-Distill-Qwen-1.5B	Device	4 × Nvidia-H100
Prompts per batch	64	Generations per prompt	8
Gradient update per RL step	1	Max prompt length	1024
Max response length	3072	Learning rate	5×10^{-6}
Clip ratio low	0.2	Clip ratio high	0.2
Epochs	2	β	0.0
Rollout temperature	1	τ	0.2

Table 4: Training configuration for DeepSeek-R1-Distill-Qwen-1.5B

Parameter	Value	Parameter	Value
Pretrained Model	DeepSeek-R1-Distill-Qwen-7B	Device	8 × Nvidia-H100
Prompts per batch	64	Generations per prompt	8
Gradient update per RL step	1	Max prompt length	1024
Max response length	3072	Learning rate	2×10^{-6}
Clip ratio low	0.2	Clip ratio high	0.2
Epochs	2	β	0.0
Rollout temperature	1	τ	0.2

Table 5: Training configuration for DeepSeek-R1-Distill-Qwen-7B

from MATH (Hendrycks et al., 2021), AIME (1983–2023) problem sets, AMC (pre-2023) contests, and Omni-MATH (Gao et al., 2024), with difficulty rebalancing to reduce trivial problems and improve RL training efficiency. The final training set contains 10,000 problems, with an additional 500 problems reserved for validation.

F.2 Models

We evaluate on three models spanning different scales and pretraining approaches: DeepSeek-R1 series: DeepSeek-R1-Distill-Qwen-7B and DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025) as well as Qwen2.5-7B (Team, 2024). These models span both general model and distillation models, while also covering different parameter scales (1.5B to 7B). All models are initialized from publicly released checkpoints and fine-tuned using the same RLVR training pipeline.

F.3 Implementation and Evaluation Details

Hyperparameters are listed in Table 4, 5 and 6. For evaluation we report Pass@1 and Pass@K ($K = 64$) across five benchmarks. Pass@K measures the fraction of problems for which at least one of K sampled completions is accepted by the verifier. For inference for each benchmark problem, we sample rollouts using temperature 1.0 and maximum generation length of 3072 for DeepSeek-R1-Distill-Qwen-1.5B and DeepSeek-R1-Distill-Qwen-7B and 4096 for Qwen-2.5-7B.

G Additional Results

G.1 Pass@K Experiments

Figures 7 present the Pass@K curves for $K \in \{1, 2, 4, 8, 16, 32, 64\}$ on DeepSeek-R1-Distill-Qwen-1.5B, DeepSeek-R1-Distill-Qwen-7B and Qwen-2.5-7B, respectively, across all five benchmarks (AIME24, AIME25, AMC, MATH, OlympiadBench). We compare UCPO against all baselines: GRPO, KL-Cov, Ent-Reg, Ent-Adv, Pass@K Training, and FGRPO. UCPO curves maintain improved performance through $K = 64$. The gap between UCPO and GRPO widens monotonically with K especially across harder datasets such as AIME. On the 7B model, this gap is particularly pronounced on AIME24 (up to $+10$ points at $K = 64$).

Parameter	Value	Parameter	Value
Pretrained Model	Qwen2.5-7B	Device	8 × Nvidia-H100
Prompts per batch	64	Generations per prompt	8
Gradient update per RL step	1	Max prompt length	1024
Max response length	4096	Learning rate	1×10^{-6}
Clip ratio low	0.2	Clip ratio high	0.2
Epochs	2	β	0.0
Rollout temperature	1	τ	0.2

Table 6: Training configuration for Qwen2.5-7B

Table 7: UCPO maintains substantially higher average Pass@1 than the diversity-oriented baselines. Although its average Pass@64 margin over the strongest baseline is modest, UCPO ranks highest in both runs, indicating a stable aggregate advantage and a better correctness–coverage trade-off. Qwen2.5-7B results across two independent training runs. Entries are mean ± sample SD (%). Bold marks the highest mean in each column.

	AIME24		AMC		MATH		OlympiadBench		Avg.	
Method	**Pass@1**	**Pass@64**	**Pass@1**	**Pass@64**	**Pass@1**	**Pass@64**	**Pass@1**	**Pass@64**	**Pass@1**	**Pass@64**
GRPO	14.09 ± 0.26	45.00 ± 2.36	$\mathbf{44.05} \pm 0.29$	91.17 ± 1.14	$\mathbf{71.66} \pm 0.07$	95.10 ± 0.42	$\mathbf{34.90} \pm 0.07$	73.26 ± 0.73	$\mathbf{41.17} \pm 0.10$	76.13 ± 0.95
KL-Cov	12.37 ± 0.18	46.67 ± 0.00	42.12 ± 0.36	90.96 ± 4.25	67.69 ± 0.16	94.90 ± 0.71	31.78 ± 0.09	74.07 ± 1.05	38.49 ± 0.06	76.65 ± 0.62
Ent-Reg	12.47 ± 0.69	55.00 ± 7.07	41.78 ± 0.44	87.95 ± 1.70	66.79 ± 0.27	94.90 ± 0.14	32.06 ± 0.06	74.44 ± 1.36	38.27 ± 0.14	78.07 ± 2.57
Ent-Adv	12.68 ± 0.04	$\mathbf{58.35} \pm 2.33$	42.18 ± 0.63	90.36 ± 0.00	68.69 ± 0.24	95.20 ± 0.28	32.41 ± 0.01	73.85 ± 0.73	38.99 ± 0.08	79.44 ± 0.33
Pass@K	11.04 ± 0.44	53.33 ± 0.00	38.12 ± 0.56	88.55 ± 0.86	62.10 ± 0.36	$\mathbf{96.10} \pm 0.71$	28.88 ± 0.09	74.66 ± 0.62	35.03 ± 0.08	78.16 ± 0.12
FGRPO	12.94 ± 0.04	55.00 ± 2.36	41.89 ± 0.22	90.96 ± 0.86	66.97 ± 0.07	95.70 ± 0.42	32.36 ± 0.21	74.15 ± 0.31	38.54 ± 0.08	78.95 ± 0.78
UCPO	$\mathbf{14.61} \pm 0.63$	56.67 ± 0.00	42.88 ± 1.10	$\mathbf{92.17} \pm 0.85$	70.78 ± 0.06	95.60 ± 0.28	34.86 ± 0.24	$\mathbf{75.11} \pm 0.21$	40.78 ± 0.48	$\mathbf{79.89} \pm 0.09$

G.2 Cost Overhead

Figure 8 shows that UCPO and GRPO have effectively identical per-step training time. UCPO's only additional computation is the $O(n)$ scalar advantage reweighting (Eq. 21), which is negligible relative to rollout generation and the policy gradient backward pass. UCPO requires no additional forward passes, no reference model, and no extra rollouts.

G.3 Sensitivity to τ

UCPO introduces a single hyperparameter $\tau \in [0, 1]$ that interpolates between standard GRPO ($\tau = 0$) and full inverse score weighting ($\tau = 1$). We evaluate sensitivity on DeepSeek-R1-Distill-Qwen-1.5B for $\tau \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$, with GRPO ($\tau = 0$) as the baseline.

Table 8: Performance and average response length across UCPO τ values on DeepSeek-R1-Distill-Qwen-1.5B. All other hyperparameters are held fixed.

	AIME		AMC		MATH		OlympiadBench		Avg		Avg. Len
τ	@1	@64	@1	@64	@1	@64	@1	@64	@1	@64	
0.0 (GRPO)	19.06	63.33	55.27	87.95	71.57	97.60	42.63	78.52	47.13	81.85	2041.38
0.1	19.38	66.67	53.65	91.57	68.89	97.60	41.26	77.48	45.79	83.33	2114.73
0.2	19.01	66.67	52.97	93.98	67.93	97.20	40.15	80.15	45.02	84.50	2130.32
0.3	19.01	66.67	54.10	93.98	68.95	97.40	40.51	79.41	45.64	84.36	2107.27
0.4	18.96	63.33	53.11	93.98	73.54	97.00	43.24	76.44	47.21	82.69	2104.25
0.5	18.23	56.67	52.03	91.57	70.22	97.40	41.98	77.78	45.61	80.85	2037.49

Results are shown in Table 8. The diversity gains (Pass@64) are robust over GRPO ($\tau = 0$) by $+0.8$ to $+2.7$ points for $\tau \leq 0.4$. Pass@1 remains competitive for $\tau \leq 0.4$. At $\tau = 0.5$, Pass@64 begins to degrade, particularly on AIME (56.67, below GRPO). At larger τ, the update places greater weight on the inverse length-normalized component. The resulting weights can become more uneven across correct rollouts, increasing estimator variance and potentially over-correcting toward underrepresented responses.

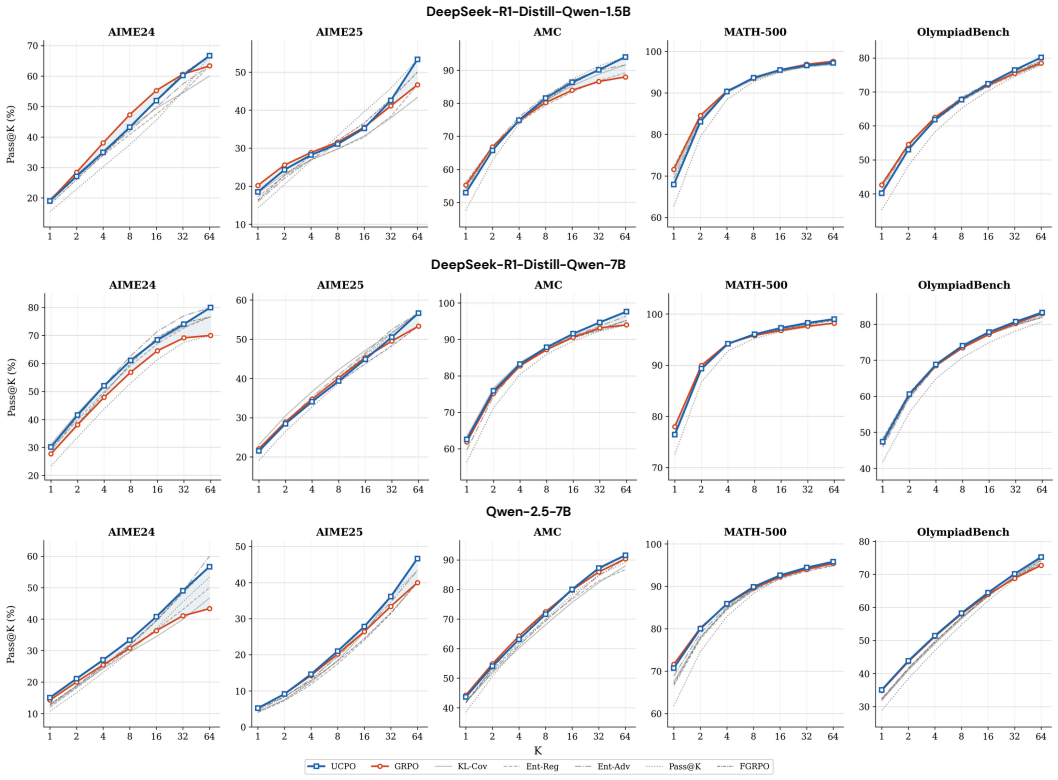

Figure 7: Pass@K curves across mathematical reasoning benchmarks and models. UCPO consistently achieves higher Pass@K, indicating better preservation of diverse correct solution paths. This behavior is consistent with UCPO redistributing gradient signal within the correct set, counteracting the collapse mechanism.

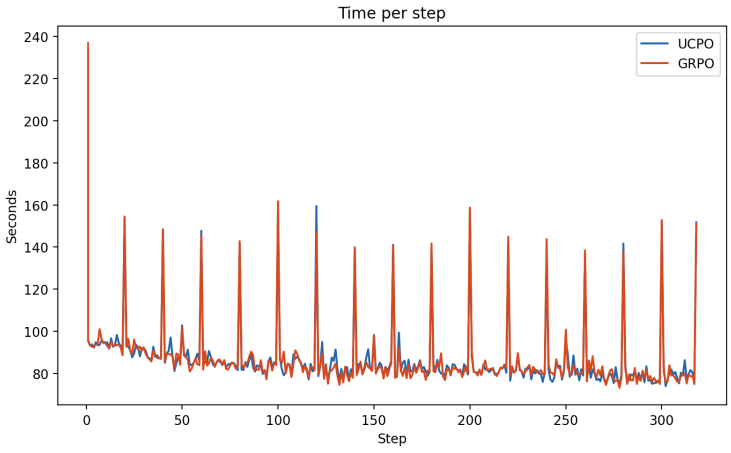

Figure 8: Wall-clock time per training step for GRPO and UCPO under identical configuration (DeepSeek-R1-Distill-Qwen-1.5B, 4×H100). The two methods exhibit nearly identical training cost throughout. Periodic spikes correspond to checkpoint saving.

These results suggest that τ provides a robust operating range. The correctness-coverage trade-off is mild. We use $\tau = 0.2$ as the default throughout our experiments as it yields the strongest Pass@64 improvements.

We also evaluated global entropy regularization across different entropy coefficients (Table 3). Compared with UCPO, global entropy regularization underperforms at Pass@64 and is more sensitive to its coefficient. This shows that global entropy is a less effective mechanism for improving multi-sample coverage, consistent with its action over the full output space rather than the conditional distribution over verifier-accepted responses.

H Equation Diversity

We compute equational diversity following (Hu et al., 2025). For each prompt, we consider only correct rollouts. From each rollout, we extract formulas appearing inside LaTeX math delimiters (`$...$`, `\(...\)`, `\[...\]`) within the first L characters of the response. For rollout i, let F_i denote the set of extracted formulas.

We define the equation uniqueness score for rollout i as

$$D_{\text{eq}}^{(i)} = \frac{\left|F_i \setminus \bigcup_{j \neq i} F_j\right|}{\max\left(1, |F_i|\right)}.$$

This score measures the fraction of formulas in rollout i that do not appear in any other correct rollout for the same prompt. We average $D_{\text{eq}}^{(i)}$ across correct rollouts to obtain a prompt-level score, and then average across prompts within each dataset. Higher values indicate greater equation-level diversity among correct solutions.

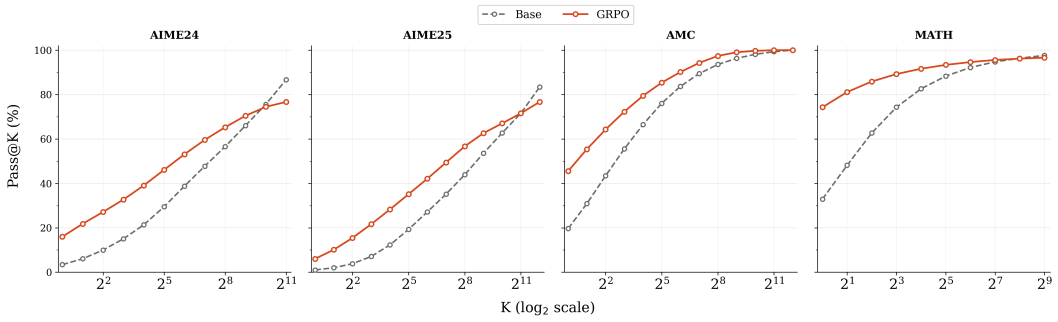

Figure 9: Pass@K comparison between the base model and GRPO on Qwen-2.5-7B. While GRPO improves Pass@1, it underperforms the base model at larger K, indicating contraction of the solution space.

I Limitations

UCPO reshapes the policy distribution over correct responses using verifier-accepted rollouts observed during training. Consequently, it cannot directly recover correct solution modes that are never sampled. If a correct reasoning strategy has already fallen below the sampling threshold, UCPO can only reinforce it after it reappears in the on-policy samples. This limitation is shared by on-policy RLVR methods and suggests that UCPO is complementary to stronger exploration mechanisms, improved sampling strategies, or verifier-guided data augmentation.

Our analysis is restricted to RLVR settings with binary verifier rewards, where correctness is explicitly defined. Extending the same principle to settings with noisy, partial, or subjective feedback, such as preference modeling, instruction following, or open-ended generation, is less straightforward because the notion of a correct set may be ill-defined or unstable.

Our experiments focus on models in the 1.5B–7B scale range. While these models provide a natural testbed for studying RLVR, they do not cover substantially larger models. Whether the same training dynamics and diversity benefits persist for substantially larger models remains an open empirical

question. Finally, the practical UCPO update uses length-normalized inverse-score weighting and therefore induces a flattened conditional target rather than the exact uniform distribution characterized by the population objective.