Preserving Exploration for LLM Reasoning via Mean Order-Statistic Alignment

Ruotian Peng[1,2,†] **Yi Ren**[3,†] **Longfei Li**[4] **Yandong Wen**[2,*]
[1]Zhejiang University [2]Westlake University [3]University of Oxford [4]Nanjing University
[†]Equal contribution [*]Corresponding author
pengruotian@westlake.edu.cn wenyandong@westlake.edu.cn

Abstract

In this paper, we propose the mean order-statistic profile, a policy-level representation for characterizing rollout diversity in large language models. The profile is a vocabulary-sized vector obtained by sorting each next-token distribution over the vocabulary and averaging across token positions by rank. Its entries are non-increasing with rank: slower decay indicates many likely token choices, suggesting more diverse continuations, whereas faster decay indicates fewer alternatives. Based on this representation, we develop Mean Order-Statistic Alignment (MOSA) to preserve the fine-tuned model's exploration ability, by minimizing the KL divergence between its profile and that of the base model. For efficiency, we approximate each profile using the top-N probabilities and a residual tail bucket, yielding an implementation with minimal computational overhead. Experiments on Countdown, 6 mathematical reasoning benchmarks, and 2 coding benchmarks show that MOSA discovers more diverse plausible solutions, improves both pass@1 and pass@K, and scales to models with up to 32B parameters.

1 Introduction

Reinforcement learning (RL) has become a cornerstone of large language model (LLM) fine-tuning, steering models toward reward-aligned outputs [25, 24, 13]. Yet these gains often come at the cost of exploration ability inherited from base models, as evidenced by degraded pass@K during RL fine-tuning [17, 9, 38, 32] (Figure 1(a),(c)). This degradation commonly appears as premature convergence to a narrow set of outputs, which can limit the model's ability to discover alternative valid solutions and achieve further performance gains [2, 26].

A natural remedy is to anchor the fine-tuned model to the base model through token-wise KL regularization [25, 7, 6, 27]. However, such a regularization leaves an unavoidable trade-off between converging to high-reward solutions and maintaining diversity at the base model's level. Strong constraints can impede adaptation, whereas weak constraints may fail to prevent exploration collapse.

We argue that this trade-off stems from using an overly fine-grained representation for regularization. Token-wise KL penalizes *any* change in the posterior probability assigned to every token. Our goal, however, is to allow task-oriented adaptation while preserving exploration behavior, as expressed in the *overall* probability allocation pattern. This mismatch motivates a representation tailored to summarizing exploration behavior.

To this end, we introduce the *mean order-statistic profile* as a policy-level representation. This representation is obtained by sorting each next-token distribution over the vocabulary and averaging the sorted probabilities by rank across token positions.[1] The resulting vocabulary-sized vector summarizes the policy's probability allocation among high- and low-probability candidates, abstracting

[1]Here, the order statistics are the next-token probabilities sorted in descending order [5].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

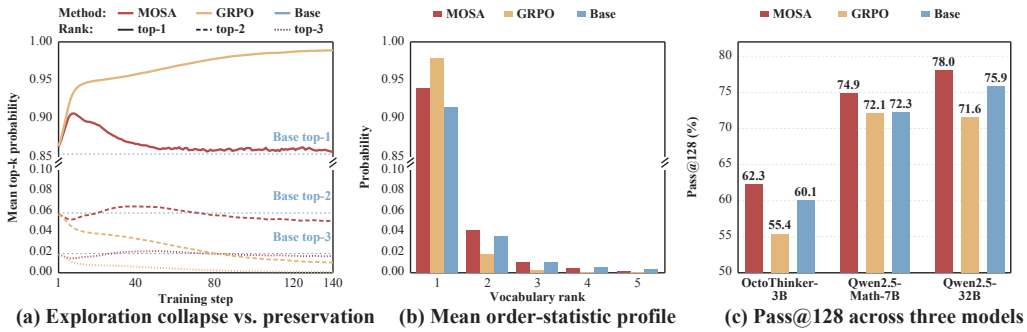


Figure 1: (a) GRPO induces exploration collapse, increasing top-1 mass while diminishing top-2/top-3 mass; in contrast, MOSA preserves the base model's probability allocation. (b) The mean order-statistic profile provides a compact characterization of the model's exploration behavior. (c) MOSA improves pass@128 averaged over 6 math reasoning benchmarks across 3 model backbones.

away token identity. As shown in Figure 1(b), the profile characterizes the policy's probability concentration: gradual decay indicates many likely token choices, suggesting more diverse continuations, whereas sharp decay indicates fewer likely alternatives.

Building on this representation, we develop Mean Order-Statistic Alignment (MOSA) to anchor the fine-tuned model's profile to that of the base model. A key observation is that the profile is itself a valid probability distribution over ranks: each sorted next-token distribution has nonnegative entries summing to one, and averaging preserves these properties. This enables a principled KL-based objective for profile alignment. To improve computational efficiency, MOSA uses a compact approximation that retains the top-N probabilities and their decay structure while aggregating the remaining mass into a residual tail bucket.

We evaluate MOSA on Countdown, 6 mathematical reasoning, and 2 coding benchmarks, using models ranging from 3B to 32B parameters. On Countdown, MOSA substantially improves exploration, achieving a 13.5% absolute gain in pass@128 over token-wise KL and discovering more diverse plausible solutions. This improved exploration does not come at the expense of exploitation, as MOSA also achieves a 9.8% absolute gain in pass@1. On mathematical reasoning benchmarks, the gains persist across the evaluated pass@K range, with even stronger gains observed on the 32B model. Ablation studies show that a small N is sufficient for strong performance and that results remain stable across a 50-fold range of MOSA loss weights.

Our contributions are as follows:

- We introduce the mean order-statistic profile, a policy-level representation of rank-wise probability allocation for characterizing exploration behavior and comparing models.
- We develop MOSA, which aligns the fine-tuned and base profiles through a principled KL-based objective, with a compact top-N approximation for efficient implementation.
- We show that inheriting the base model's exploration behavior through profile alignment can yield substantial gains in solution diversity and pass@1/pass@K during fine-tuning, with benefits extending to models up to 32B parameters.

2 Related Work

RLVR for reasoning. Reinforcement learning from verifiable rewards (RLVR) has become a central approach for improving large language models on math and code reasoning tasks [6, 36, 39, 11]. Subsequent work has largely focused on improving the optimization behavior of GRPO-style training. For example, Dr. GRPO revisits the role of advantage normalization and length bias [19], while GSPO moves the importance-sampling correction from the token level to the sequence level and reports improved stability [40]. Another line of work has questioned the standard use of token-wise KL regularization in RLVR. DAPO removes KL regularization to permit larger policy updates [36], and recent analyses argue more generally that token-wise anchoring can limit the extent to which RL departs from the base model [32, 38]. We share the view that token-level KL can be an overly

restrictive form of regularization for reasoning-oriented RL. We address this by introducing the mean order-statistic profile, a policy-level statistic that summarizes how probability mass is distributed across ranked candidates. Aligning this statistic to the base model's profile via KL divergence shifts the alignment target from local reference-policy proximity to the global shape of the policy's confidence allocation.

Exploration in RLVR. A growing literature studies how to preserve exploration and prevent premature concentration in reasoning-oriented RL. One family of methods encourages exploration through entropy-style regularization, for example by increasing output randomness or upweighting updates on high-entropy tokens that are considered more exploration-relevant [3, 29, 11, 23]. However, these methods operate primarily at the token level. A second family modifies the training objective more directly, for instance by optimizing pass@K-style criteria [2, 28]. However, recent analyses [37] suggest that such objectives may largely act as reweightings of pass@1-style learning signals, rather than directly inducing qualitatively new exploration behavior. A third line of work increases exploration through more diverse data or broader interaction contexts, such as augmenting training with off-policy traces from stronger models [8, 16] or generating alternative solution trajectories for hard problems [35, 17]. These approaches, however, typically rely on additional data generation or external trajectory sources. By contrast, MOSA targets exploration collapse directly through a principled objective: we measure the KL divergence between the current policy's mean order-statistic profile and that of the base model, and minimize it during RL training.

3 Preliminaries

Group Relative Policy Optimization (GRPO) [25] is a widely adopted RLVR method that eliminates the critic model required by PPO [24]. Let $\mathcal{D}$ denote the training distribution over questions. For each question $x \sim \mathcal{D}$, GRPO samples a group of G responses $\{o_1, \ldots, o_G\}$ and assigns each response a verifiable reward $R(o_i)$. The group-normalized advantage is $A_i = (R(o_i) - \mu)/\sigma$, where μ and σ are the group mean and standard deviation of the rewards. The policy is updated by maximizing

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, o_i\sim\pi_{\mathrm{old}}(\cdot|x)}\left[\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\min\Big(\gamma_{i,t}A_i,\ \mathrm{clip}\big(\gamma_{i,t}, 1-\epsilon_{\mathrm{clip}}, 1+\epsilon_{\mathrm{clip}}\big)A_i\Big)\right] - \beta_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}}, \tag{1}$$

where $\gamma_{i,t} = \pi_\theta(o_{i,t} \mid x, o_{i,<t})/\pi_{\mathrm{old}}(o_{i,t} \mid x, o_{i,<t})$, and π_{old} denotes the sampling policy. Here, ϵ_{clip} is the clipping parameter, $\mathcal{L}_{\mathrm{KL}}$ denotes the token-averaged KL divergence between π_θ and π_{base}, and β_{KL} is its coefficient.

Exploration collapse. RL fine-tuning can improve pass@1, yet these gains often come at the cost of exploration ability. This degradation is reflected in lower pass@K scores, which measure the probability of finding at least one correct solution within K independent attempts. Figure 1(c) shows this decline in pass@128 for GRPO relative to the base model. Empirically, following recent findings [23], we observe a pronounced concentration of probability mass during RL training. As shown in Figure 1(a), the mean rank-1 probability increases, while the rank-2 and rank-3 probabilities decrease. This increasing concentration can limit the model's ability to discover alternative valid solutions. We therefore aim to preserve the base model's exploration ability while improving pass@1 during RL fine-tuning.

4 Preserving Exploration via Mean Order-Statistic Alignment

To preserve the base model's exploration ability during RL fine-tuning, we propose **Mean Order-Statistic Alignment (MOSA)**, which aligns the fine-tuned model's rank-wise probability allocation with that of the base model. We first establish the requirements for an exploration-preserving statistic in Section 4.1, then define the mean order-statistic profile in Section 4.2. We use its structure as a probability distribution to derive a principled KL-based alignment objective in Section 4.3, and develop an efficient implementation using a compact top-N approximation in Section 4.4.

4.1 What Should Be Preserved for Exploration?

Preserving the base model's exploration ability requires maintaining probability mass on alternative continuations, allowing repeated sampling to search for diverse valid solutions. Token-wise KL

Table 1: Comparison of existing approaches for RLVR. MOSA introduces a policy-level summary statistic, the mean order-statistic profile, enabling direct alignment of exploration behavior.

Method	Representation	Granularity	Objective
Token-wise KL	Next-token distribution $\pi_\theta(\cdot\|s_{i,t})$	Token-level	Local imitation
Entropy bonus	Token entropy $H(\pi_\theta(\cdot\|s_{i,t}))$	Token-level	Local uncertainty
GSPO	Sequence importance ratio $\left[\frac{\pi_\theta(o \mid x)}{\pi_{\text{old}}(o \mid x)}\right]^{1/\|o\|}$	Sequence-level	Sequence-level stability
MOSA	**Mean order-statistic profile $\hat{\mathbf{c}}_\theta$**	**Policy-level**	**Exploration preservation**

regularization is a natural way to anchor the fine-tuned model to the base model. However, it penalizes changes in the probabilities assigned to each token, which can limit the policy's ability to achieve higher rewards. Preserving the probability allocation pattern does not require retaining the base model's exact token preferences: two next-token distributions can have identical sorted probabilities while assigning them to different tokens. This motivates characterizing probability allocation across candidates ranked by probability, independently of token identity.

In addition, token-wise KL measures local differences between next-token distributions at individual decoding states. Exploration is a higher-level behavior reflected in the probabilities assigned to alternative continuations and complete responses. Capturing this behavior therefore calls for a policy-level summary of probability allocation.

Together, these considerations motivate three desiderata:

- **Identity-free.** Two next-token distributions that differ only by a permutation of vocabulary labels should have the same representation.
- **Rank-sensitive.** The statistic should distinguish changes in probability mass on the highest-ranked candidate, other high-probability alternatives, and the low-probability tail.
- **Policy-level.** The statistic should aggregate probability allocation across decoding states encountered along policy-induced rollouts.

4.2 Mean Order-Statistic Profile

These desiderata naturally lead to order statistics of the next-token distribution. Let i index sampled responses and x_i denote the input associated with response o_i. At position t of response o_i, the decoding state is $s_{i,t} = (x_i, o_{i,<t})$. We sort the probabilities in $\pi_\theta(\cdot \mid s_{i,t})$ in descending order and write

$$\pi_\theta^{(1)}(s_{i,t}) \geq \pi_\theta^{(2)}(s_{i,t}) \geq \cdots \geq \pi_\theta^{(V)}(s_{i,t}),$$

where V is the vocabulary size and $\pi_\theta^{(r)}(s_{i,t})$ denotes the r-th largest probability. This representation retains the probability assigned to each rank while abstracting away token identity: the token occupying a given rank may differ across decoding states and policies.

To obtain a policy-level summary, we aggregate these sorted probabilities over policy-induced rollouts. We define the **mean order-statistic profile** as

$$\mathbf{c}_\theta = \mathbb{E}_{x_i \sim \mathcal{D},\, o_i \sim \pi_\theta(\cdot|x_i)} \left[\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left(\pi_\theta^{(1)}(s_{i,t}), \ldots, \pi_\theta^{(V)}(s_{i,t}) \right) \right].$$

The inner average covers decoding states within each response, while the outer expectation covers prompts and on-policy responses. This profile summarizes the policy's rank-wise probability allocation during generation. The mean rank-1 probability reflects the mass on the highest-ranked candidate, while the mean probabilities at subsequent ranks describe the mass retained on alternatives. Figure 1(b) illustrates how the profile characterizes probability concentration: compared with the base model, the model trained with GRPO exhibits a higher mean rank-1 probability and lower mean probabilities at subsequent ranks. The profile thus captures the concentration of probability mass on the highest-ranked candidate at the expense of alternatives after GRPO training, consistent with GRPO's lower pass@128 relative to the base model in Figure 1(c).

4.3 Mean Order-Statistic Alignment

Let $\mathbf{c}_{\text{base}}$ denote the mean order-statistic profile of the fixed base policy π_{base}. We use $\mathbf{c}_{\text{base}}$ as the alignment reference to preserve the base model's exploration ability during RL fine-tuning. The mean order-statistic profile forms a valid probability distribution over ranks. Each sorted next-token distribution has non-negative entries summing to one, and averaging and expectation preserve these properties. We therefore define the MOSA alignment loss as

$$\mathcal{L}_{\text{MOSA}}(\theta) = D_{\text{KL}}\big(\mathbf{c}_{\text{base}} \| \mathbf{c}_\theta\big) = \sum_{r=1}^{V} c_{\text{base},r} \log \frac{c_{\text{base},r}}{c_{\theta,r}}. \tag{2}$$

This direction discourages the fine-tuned policy from shifting probability mass away from alternatives toward the highest-ranked candidate, while allowing its token preferences to change. We combine this alignment loss with GRPO:

$$\max_\theta \; J_{\text{RL}}(\theta) - \beta_{\text{MOSA}} \mathcal{L}_{\text{MOSA}}(\theta), \tag{3}$$

where J_{RL} is the GRPO objective with $\beta_{\text{KL}} = 0$, and β_{MOSA} controls the alignment strength. The RL objective favors higher-reward responses, while MOSA encourages the policy to retain the base model's probability allocation pattern.

4.4 Efficient Implementation

To improve computational efficiency, we adopt a compact approximation consisting of the top-N probabilities and a residual tail bucket. The top-N probabilities retain the probability allocation among the highest-ranked candidates, while the tail bucket aggregates the remaining mass. For each decoding state $s_{i,t}$, we define

$$\rho_N(\pi_\theta(\cdot \mid s_{i,t})) = \left(\pi_\theta^{(1)}(s_{i,t}), \ldots, \pi_\theta^{(N)}(s_{i,t}), 1 - \sum_{r=1}^{N} \pi_\theta^{(r)}(s_{i,t}) \right).$$

This reduces the profile dimensionality from V to $N+1$ while preserving non-negativity and normalization, allowing us to retain the KL-based alignment objective.

At each RL update, we estimate the current policy's compressed profile using the on-policy rollouts already collected for GRPO. Each batch contains B questions and G responses per question, giving $M = BG$ responses in total. We index these responses by $i = 1, \ldots, M$. The current estimate is

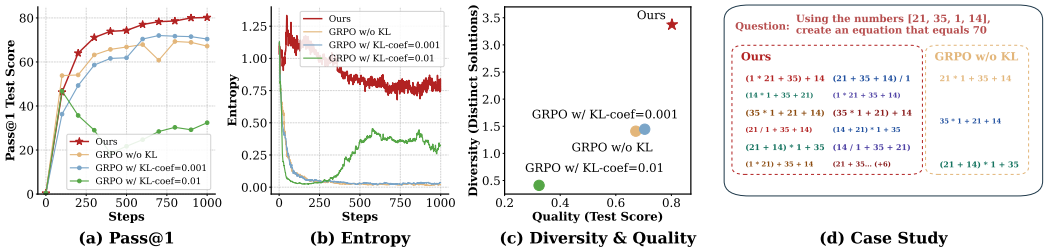
$$\widehat{\mathbf{c}}_\theta = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \rho_N(\pi_\theta(\cdot \mid s_{i,t})).$$

As in the full profile, we average over decoding states within each response and then across responses. Before the first parameter update, we apply the same estimator to the first training batch of rollouts generated by π_{base}. The resulting estimate $\hat{\mathbf{c}}_{\text{base}}$ remains fixed throughout training. In training, we replace the full profiles in $\mathcal{L}_{\text{MOSA}}$ with their compressed batch estimates $\hat{\mathbf{c}}_{\text{base}}$ and $\hat{\mathbf{c}}_\theta$, so the KL divergence is evaluated over $N+1$ rank channels.

For a fixed policy, the group-level bound in Appendix A.1 shows that estimating all $N+1$ coordinates within ε_{est} requires $B = O\big(\varepsilon_{\text{est}}^{-2} \log \frac{N+1}{\delta}\big)$ independent prompt groups with probability at least $1-\delta$, versus a $\log V$ dependence for the full profile. Profile estimation reuses the GRPO rollouts, and the base profile is computed only once. Thus, MOSA requires no additional rollout collection and reduces the profile used for alignment from V to $N+1$ entries.

4.5 Comparison with Existing Approaches

Table 1 summarizes the key differences between MOSA and existing approaches. The central distinction lies in the statistic used and the associated optimization objective. **Token-wise KL** uses each local next-token distribution $\pi_\theta(\cdot \mid s_{i,t})$ as its statistic and minimizes its KL divergence from the reference distribution. **Entropy bonus** summarizes each local distribution by the scalar $H(\pi_\theta(\cdot \mid s_{i,t}))$ and maximizes it. **GSPO** uses the length-normalized sequence-level importance ratio $[\pi_\theta(o \mid x)/\pi_{\text{old}}(o \mid x)]^{1/|o|}$ as its statistic and applies sequence-level clipping for stability. **MOSA** instead uses the mean order-statistic profile $\hat{\mathbf{c}}_\theta$ as its statistic and minimizes $D_{\text{KL}}(\hat{\mathbf{c}}_{\text{base}} \| \hat{\mathbf{c}}_\theta)$, directly aligning rank-wise probability allocation with the base model.



(a) Pass@1 (b) Entropy (c) Diversity & Quality (d) Case Study

Figure 2: Countdown results comparing our method with GRPO and token-wise KL baselines. (a) Our method achieves the highest pass@1 after 200 training steps. (b) MOSA maintains higher policy entropy during training than the evaluated GRPO and token-wise KL baselines. (c) Our method achieves the best quality–diversity trade-off, improving both pass@1 and solution diversity simultaneously. (d) A case study shows that our method discovers more diverse valid solutions for the same question.

5 Experiment

5.1 Countdown Task

We begin by evaluating MOSA on the Countdown task, where the model is given a list of numbers and must construct an expression using the operators $(+, -, \times, \div)$ to reach a target number. The task admits multiple valid solutions, providing a setting to evaluate how MOSA affects both task accuracy and solution diversity during RL fine-tuning.

Setup. We train Qwen2.5-3B [33] using the Countdown dataset from TinyZero [22], comprising 327,680 training problems and 1,024 held-out test problems. A response receives reward 1 if it produces a correct equation in the required reasoning format, 0.1 if only the format is correct, and 0 otherwise. Training uses batch size 256, mini-batch size 64, learning rate 10^{-6}, maximum response length 1,024 tokens, and $G = 5$ responses per prompt. For GRPO w/ KL baselines we evaluate $\beta_{\text{KL}} \in \{0.001, 0.01\}$; for MOSA we use N=10 and $\beta_{\text{MOSA}} = 0.1$. All methods are trained for 1,000 steps. During evaluation, we sample 128 responses per problem to estimate pass@1 and pass@128.

Results Analysis. Figure 2(a) shows that MOSA achieves the highest pass@1 on Countdown, reaching **80.2**, compared with 67.2 for GRPO, 70.4 for GRPO w/ $\text{KL}_{0.001}$, and 32.4 for GRPO w/ $\text{KL}_{0.01}$. These gains suggest that policy-level alignment is less restrictive than token-wise KL regularization, supporting better exploitation. The benefits also extend to exploration of a wider range of solutions. Figure 2(c) shows that MOSA generates an average of **3.37** distinct valid solutions per problem, versus 1.41 for GRPO, 1.45 for GRPO w/ $\text{KL}_{0.001}$, and 0.41 for GRPO w/ $\text{KL}_{0.01}$. Representative examples in Figure 2(d) illustrate this greater variety of valid expressions for the same problem. The entropy measurements in Figure 2(b) provide complementary evidence. MOSA maintains relatively high entropy throughout fine-tuning, whereas baseline entropy drops sharply and remains substantially lower, suggesting that MOSA encourages a less concentrated policy.

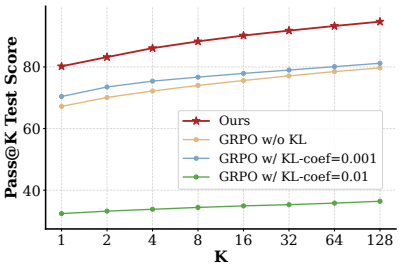
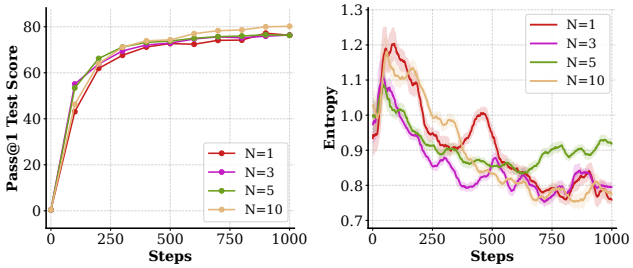


Figure 3: Pass@K performance of MOSA and baselines on Countdown.

To investigate whether this diversity helps solving more problems, we report pass@K, the probability of finding at least one correct solution within K attempts. Figure 3 shows that MOSA achieves **94.7** pass@128, compared with 79.7 for GRPO, 81.2 for GRPO w/ $\text{KL}_{0.001}$, and 36.4 for GRPO w/ $\text{KL}_{0.01}$, supporting MOSA's ability to sustain exploration that is both diverse and effective.



(a) Pass@1 (b) Entropy

Figure 4: Countdown performance under different N.

Ablation on N. We vary N to examine whether a compact profile retains enough information for effective alignment. As shown in Figure 4,

Table 2: Pass@1 and pass@128 results across 3 model backbones on 6 mathematical reasoning benchmarks.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg. P@1	Avg. P@128
Qwen2.5-Math-7B								
Base Model	12.9	5.0	38.1	56.0	16.3	25.6	25.7	72.3
+GRPO	26.6	12.2	62.4	76.5	33.1	38.6	41.6	72.1
+GRPO w/ KL	27.6	12.5	61.3	76.6	33.8	38.7	41.8	72.0
+GSPO	31.3	11.4	62.3	77.0	34.6	38.9	42.6	74.0
+DAPO	30.9	12.3	63.0	76.8	34.3	39.1	42.7	74.3
+APO	23.9	9.6	60.8	74.5	32.5	37.1	39.7	74.2
+KL-Cov	29.7	11.1	61.6	76.6	34.1	38.8	42.0	73.7
+SimKO	31.2	11.2	62.3	77.1	35.0	39.2	42.7	74.1
+Entropy-Adv	24.8	11.8	63.0	76.5	33.3	39.4	41.5	71.4
+MOSA$_{(\mathrm{GRPO})}$	31.9	12.4	63.3	78.4	36.0	39.9	**43.6**	**74.9**
+MOSA$_{(\mathrm{GSPO})}$	31.6	12.0	62.2	77.8	35.9	39.5	**43.2**	**75.3**
+MOSA$_{(\mathrm{DAPO})}$	32.9	14.2	64.0	78.1	35.2	39.7	**44.0**	**75.7**
Qwen2.5-32B								
Base Model	9.7	4.9	42.1	64.5	27.0	28.1	29.4	75.9
+GRPO	27.3	16.1	65.1	82.7	43.1	44.5	46.5	71.6
+GSPO	27.7	12.7	65.0	82.0	42.9	43.6	45.7	73.6
+Entropy-Adv	26.4	15.3	67.5	82.6	42.6	43.9	46.4	71.9
+MOSA$_{(\mathrm{GRPO})}$	28.8	17.1	68.8	84.1	43.3	47.2	**48.2**	**78.0**
OctoThinker-3B-Hybrid-Base								
Base Model	1.4	0.5	14.4	30.8	8.8	10.7	11.1	60.1
+GRPO	3.4	0.9	29.8	56.3	20.6	22.6	22.3	55.4
+GRPO w/ KL	3.3	1.1	30.1	57.8	20.7	23.9	22.8	57.9
+GSPO	2.9	0.8	28.3	56.9	21.1	22.4	22.1	58.9
+Entropy-Adv	2.1	1.8	27.8	57.7	20.8	23.4	22.3	56.4
+MOSA$_{(\mathrm{GRPO})}$	4.1	1.7	28.9	59.3	21.2	24.5	**23.3**	**62.3**

all four settings, $\mathtt{N} \in \{1, 3, 5, 10\}$, yield broadly similar final pass@1. However, N=1 exhibits more pronounced entropy fluctuations, whereas N=10 converges slightly more slowly than N=3 or N=5. These results support our compact top-N design: retaining only a few ranks is sufficient for strong task performance. We therefore use N=3 in subsequent experiments to obtain stable entropy dynamics while keeping the profile compact.

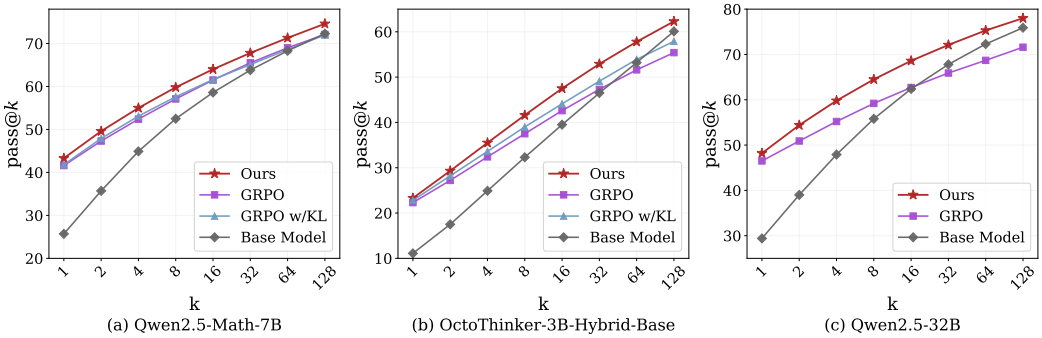
5.2 Reasoning on Math Tasks

We next evaluate MOSA on mathematical reasoning tasks to assess its effectiveness in more complex reasoning settings and across model backbones ranging from 3B to 32B parameters.

Setup. We experiment with a diverse set of models, including Qwen2.5-Math-7B [34], Qwen2.5-32B [33], and OctoThinker-3B-Hybrid-Base [31]. For the math tasks, models are trained on the MATH dataset [12]. Across backbones, we compare MOSA with GRPO, GRPO with token-wise KL, GSPO [40], and Entropy-Adv [3]. On Qwen2.5-Math-7B, we additionally compare with DAPO [36], APO [30], KL-Cov [4], and SimKO [23], and integrate MOSA with GSPO and DAPO. Evaluations are conducted on a variety of benchmarks: MATH-500 [12], Minerva Math [15], Olympiad-Bench [10], AMC and AIME. We evaluate pass@K for K up to 128 using the unbiased estimator

$$\mathrm{pass@}\,K = \mathbb{E}_{x\sim\mathcal{D}_{\mathrm{eval}}}\left[1 - \frac{\binom{n-n_{\mathrm{correct}}(x)}{K}}{\binom{n}{K}}\right],$$

where $\mathcal{D}_{\mathrm{eval}}$ is the empirical distribution over evaluation problems, and $n_{\mathrm{correct}}(x)$ is the number of correct responses among n responses generated for problem x. More details in Appendix B.



(a) Qwen2.5-Math-7B (b) OctoThinker-3B-Hybrid-Base (c) Qwen2.5-32B

Figure 5: Pass@K curves averaged over 6 math benchmarks. Across all values of K, MOSA consistently outperforms GRPO on all 3 backbones and GRPO w/ KL on the backbones for which this baseline is evaluated.

5.2.1 Performance Analysis

MOSA consistently improves pass@1 on math reasoning. Table 2 reports average pass@1 across 6 benchmarks and 3 model backbones. On Qwen2.5-Math-7B, MOSA reaches **43.6**, compared with 25.7 for the base model, 41.6 for GRPO, and 41.8 for GRPO w/ KL. It also outperforms exploration-oriented baselines, including DAPO, APO, KL-Cov, and SimKO, by **+0.9**, **+3.9**, **+1.6**, and **+0.9**, respectively. These results show that MOSA achieves stronger exploitation in complex mathematical reasoning tasks. The benefit of MOSA is verified at larger model scale: on Qwen2.5-32B, MOSA achieves **48.2**, compared with 29.4 for the base model and 46.5 for GRPO. Results on OctoThinker-3B-Hybrid-Base further show that MOSA's pass@1 gains generalize to other backbones.

MOSA produces more diverse correct reasoning strategies. Following prior work [41], we use Claude-4.6-Sonnet [1] as the evaluator to score 16 correct responses per problem on AIME 2024/2025 with Qwen2.5-Math-7B (prompt in Appendix D). Figure 6 shows relative diversity gains of **+15.6%** over GRPO and **+11.6%** over GRPO w/ KL, alongside higher test accuracy. The judged solutions differ in reasoning strategy rather than merely in phrasing, indicating that MOSA's accuracy gains are accompanied by semantic diversity among correct solutions. Additional human, reference-free, and multi-judge evaluations across 3 model backbones are provided in Appendix C.1.

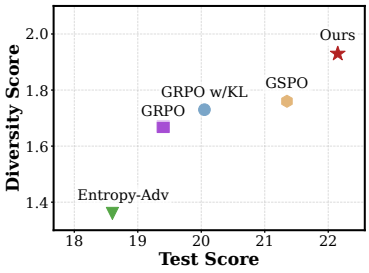


Figure 6: Quality-Diversity tradeoff.

MOSA significantly improves pass@K. Figure 5 shows average pass@K over 6 math benchmarks. On Qwen2.5-Math-7B, GRPO and GRPO w/ KL achieve pass@128 scores of 72.1 and 72.0, close to the base model's 72.3. MOSA instead reaches **74.9**, also outperforming additional exploration-oriented methods, including DAPO, APO, and SimKO. Together with the strategy-diversity results, these gains suggest that MOSA's broader range of reasoning strategies helps it solve challenging math problems. MOSA's improvement over GRPO becomes more significant on Qwen2.5-32B: GRPO, GSPO, and Entropy-Adv all fall below the base model's pass@128 of 75.9, whereas MOSA reaches **78.0**, outperforming GRPO by **6.4** points. The improvement also extends to OctoThinker-3B-Hybrid-Base, where MOSA outperforms both the base model and GRPO.

MOSA integrates effectively with other RL algorithms. We apply MOSA to GSPO and DAPO on Qwen2.5-Math-7B to assess its compatibility with other RL objectives. Adding MOSA improves average pass@1 by **+0.6** over GSPO and **+1.3** over DAPO, while improving average pass@128 by **+1.3** and **+1.4**, respectively. DAPO+MOSA achieves the highest average pass@1 of **44.0** and pass@128 of **75.7**. These improvements show that mean order-statistic alignment can complement different RL objectives, improving both exploration and exploitation.

5.2.2 Ablation Studies

We vary β_{MOSA} over the range $[0.01, 0.5]$ to assess MOSA's sensitivity to the alignment strength. As shown in Table 3, MOSA consistently outperforms GRPO on both pass@1 and pass@128 across all values, confirming robustness to this hyperparameter. We observe a mild trade-off: smaller weights yield slightly higher pass@1 (e.g., $\beta_{\text{MOSA}} = 0.05$ achieves 43.6 vs. 42.9 for $\beta_{\text{MOSA}} = 0.1$), while larger weights produce better pass@128 (e.g., $\beta_{\text{MOSA}} = 0.5$ achieves 74.9 vs. 73.6 for

Table 3: Effect of β_{MOSA} on pass@1 and pass@128, evaluated on math benchmarks using Qwen2.5-Math-7B

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Base Model	12.9 / 64.0	5.0 / 42.3	38.1 / 93.9	56.0 / 95.4	16.3 / 64.0	25.6 / 74.4	25.7 / 72.3
+GRPO	26.6 / 70.0	12.2 / 47.1	62.4 / 91.4	76.5 / 94.0	33.1 / 59.2	38.6 / 71.1	41.6 / 72.1
+MOSA ($\beta_{\mathrm{MOSA}} = 0.01$)	31.3 / 65.8	12.8 / 50.0	62.3 / 96.5	78.0 / 95.2	35.3 / 63.6	39.7 / 72.6	43.2 / 74.0
+MOSA ($\beta_{\mathrm{MOSA}} = 0.03$)	30.6 / 67.0	11.7 / 47.2	63.2 / 95.8	77.7 / 95.6	35.3 / 64.0	39.6 / 72.1	43.0 / 73.6
+MOSA ($\beta_{\mathrm{MOSA}} = 0.05$)	31.9 / 70.6	12.3 / 47.3	63.3 / 97.1	78.4 / 96.4	36.0 / 64.7	39.9 / 73.5	43.6 / 74.9
+MOSA ($\beta_{\mathrm{MOSA}} = 0.08$)	31.6 / 68.0	12.1 / 47.7	63.0 / 96.1	77.4 / 96.2	34.9 / 64.3	39.4 / 72.1	43.1 / 74.1
+MOSA ($\beta_{\mathrm{MOSA}} = 0.1$)	31.7 / 66.5	11.9 / 52.0	62.3 / 97.0	77.3 / 95.4	34.8 / 64.3	39.4 / 73.3	42.9 / 74.8
+MOSA ($\beta_{\mathrm{MOSA}} = 0.5$)	31.1 / 68.7	12.7 / 52.6	62.5 / 95.1	77.2 / 96.2	34.9 / 64.0	39.3 / 73.0	42.9 / 74.9

$\beta_{\mathrm{MOSA}} = 0.03$). This pattern is consistent with our design goal: a stronger alignment better preserves the base model's rank-wise probability structure, maintaining more diverse solution paths at the cost of a marginal reduction in pass@1 accuracy. Overall, the differences are small, indicating that MOSA is robust to the choice of β_{MOSA}.

5.3 Code Reasoning Task

We further assess MOSA's effectiveness on code reasoning tasks by fine-tuning DeepSeek-R1-Distill-Qwen-7B on the DeepCoder [21] training dataset. More details in Appendix B.3. Table 4 presents the results on code reasoning benchmarks. On LiveCodeBench, MOSA achieves significant improvements over the base model, gaining +4.12 in pass@1 and +5.38 in pass@16, and outperforming GRPO by +0.76 and +1.08, respectively.

Table 4: Performance comparison on code reasoning benchmarks for DeepSeek-R1-Distill-Qwen-7B.

Method	LiveCodeBench		HumanEval+	
	Pass@1	Pass@16	Pass@1	Pass@16
Base Model	30.85	47.31	81.02	94.48
GRPO	34.21	51.61	80.71	94.48
MOSA	**34.97**	**52.69**	**81.71**	**95.71**

On HumanEval+, MOSA achieves the highest pass@1 (81.71) and pass@16 (95.71). GRPO's pass@1 is slightly below the base model (80.71 versus 81.02). These results show that MOSA's exploration-preserving alignment also benefits code reasoning, improving both pass@1 and pass@16 beyond GRPO, even on a benchmark where GRPO does not surpass the base model.

6 Conclusion

In this work, we revisit exploration preservation in RL fine-tuning by asking what behavior should be inherited from the base model. We argue that exploration is poorly captured by token-wise measurements and is better represented as a policy-level pattern of probability allocation across candidate continuations. To this end, we introduce the mean order-statistic profile and propose MOSA, which efficiently aligns this profile between the fine-tuned and base models via a KL-based objective with a "top-N + tail" approximation. Experiments show that MOSA mitigates mode collapse, improves useful diversity, and enhances both pass@1 and pass@K across reasoning benchmarks. Overall, our results suggest that exploration-preserving fine-tuning requires anchoring through the right characterization of policy behavior. While MOSA scales effectively up to 32B models in our experiments, evaluating its behavior on larger models (e.g., 70B) remains a future direction.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62606411). The authors thank Westlake HPC Center for technical assistance, and declare no competing interests.

Author Contributions: Ruotian Peng conceived and led the project, designed and implemented the MOSA framework, conducted the experiments, and drafted the manuscript. Yi Ren contributed to the development and discussion of the ideas and assisted with theoretical proofs. Longfei Li helped improve the writing of the introduction and abstract. Yandong Wen supervised the project, co-conceived and co-designed the MOSA framework, and offered extensive feedback on the manuscript.

References

[1] Anthropic. Claude sonnet 4.6 system card. https://www.anthropic.com/news/claude-sonnet-4-6, February 2026. Accessed: 2026-2-17.

[2] 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.

[3] 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.

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

[5] Herbert A David and Haikady N Nagaraja. *Order statistics*. John Wiley & Sons, 2004.

[6] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

[7] DeepSeek-AI. Deepseek-v4: Towards highly efficient million-token context intelligence, 2026.

[8] Yihong Dong, Xue Jiang, Yongding Tao, Huanyu Liu, Kechi Zhang, Lili Mou, Rongyu Cao, Yingwei Ma, Jue Chen, Binhua Li, et al. Rl-plus: Countering capability boundary collapse of llms in reinforcement learning with hybrid-policy optimization. *arXiv preprint arXiv:2508.00222*, 2025.

[9] Andre He, Daniel Fried, and Sean Welleck. Rewarding the unlikely: Lifting grpo beyond distribution sharpening. *arXiv preprint arXiv:2506.02355*, 2025.

[10] Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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. *arXiv preprint arXiv:2402.14008*, 2024.

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

[12] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

[13] Jian Hu. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

[14] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

[15] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in neural information processing systems*, 35:3843–3857, 2022.

[16] Jiazheng Li, Hong Lu, Kaiyue Wen, Zaiwen Yang, Jiaxuan Gao, Hongzhou Lin, Yi Wu, and Jingzhao Zhang. Questa: Expanding reasoning capacity in llms via question augmentation. *arXiv preprint arXiv:2507.13266*, 2025.

[17] Xiao Liang, Zhongzhi Li, Yeyun Gong, Yelong Shen, Ying Nian Wu, Zhijiang Guo, and Weizhu Chen. Beyond pass@ 1: Self-play with variational problem synthesis sustains rlvr. *arXiv preprint arXiv:2508.14029*, 2025.

[18] Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/43e9d647ccd3e4b7b5baab53f0368686-Abstract-Conference.html.

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

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

[21] Michael Luo, Sijun Tan, Roy Huang, Ameen Patel, Alpay Ariyak, Qingyang Wu, Xiaoxiang Shi, Rachel Xin, Colin Cai, Maurice Weber, Ce Zhang, Li Erran Li, Raluca Ada Popa, and Ion Stoica. Deepcoder: A fully open-source 14b coder at o3-mini level. https://pretty-radio-b75.notion.site/DeepCoder-A-Fully-Open-Source-14B-Coder-at-O3-mini-Level-1cf81902c14680b3bee5eb349a512a51, 2025. Notion Blog.

[22] Jiayi Pan, Junjie Zhang, Xingyao Wang, Lifan Yuan, Hao Peng, and Alane Suhr. Tinyzero. https://github.com/Jiayi-Pan/TinyZero, 2025. Accessed: 2025-01-24.

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

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

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

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

[27] GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghua Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao, Hongwei Liu, Hongxi Yan, Huan Liu, Huilong Chen, Ji Li, Jiajing Zhao, Jiamin Ren, Jian Jiao, Jiani Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, Mingshu Zhai, Pengfan Du, Qian Dong, Shangde Lei, Shangqing Tu, Shangtong Yang, Shaoyou Lu, Shijie Li, Shuang Li, Shuang-Li, Shuxun Yang, Sibo Yi, Tianshu Yu, Wei Tian, Weihan Wang, Wenbo Yu, Weng Lam Tam, Wenjie Liang, Wentao Liu, Xiao Wang, Xiaohan Jia, Xiaotao Gu, Xiaoying Ling, Xin Wang, Xing Fan, Xingru Pan, Xinyuan Zhang, Xinze Zhang, Xiuqing Fu, Xunkai Zhang, Yabo Xu, Yandong Wu, Yida Lu, Yidong Wang, Yilin Zhou, Yiming Pan, Ying Zhang, Yingli Wang, Yingru Li, Yinpei Su, Yipeng Geng, Yitong Zhu, Yongkun Yang, Yuhang Li, Yuhao Wu, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yuxuan Zhang, Zezhen Liu, Zhen Yang, Zhengda Zhou, Zhongpei Qiao,

Zhuoer Feng, Zhuorui Liu, Zichen Zhang, Zihan Wang, Zijun Yao, Zikang Wang, Ziqiang Liu, Ziwei Chai, Zixuan Li, Zuodong Zhao, Wenguang Chen, Jidong Zhai, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models, 2025. URL https://arxiv.org/abs/2508.06471.

[28] Christian Walder and Deep Karkhanis. Pass@ k policy optimization: Solving harder reinforcement learning problems. *arXiv preprint arXiv:2505.15201*, 2025.

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

[30] Tianyi Wang, Long Li, Hongcan Guo, Yibiao Chen, Yixia Li, Yong Wang, Yun Chen, and Guanhua Chen. Anchored policy optimization: Mitigating exploration collapse via support-constrained rectification, 2026. URL https://arxiv.org/abs/2602.05717.

[31] Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *arXiv preprint arXiv:2506.20512*, 2025. Preprint.

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

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

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

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

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

[37] Yang Yu. Pass@ k metric for rlvr: A diagnostic tool of exploration, but not an objective. *arXiv preprint arXiv:2511.16231*, 2025.

[38] 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.

[39] Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. In *Second Conference on Language Modeling*, 2025.

[40] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[41] Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, et al. Flowrl: Matching reward distributions for llm reasoning. *arXiv preprint arXiv:2509.15207*, 2025.

A More Theoretical Analysis

A.1 Estimator Reliability under Rank-Space Compression

MOSA relies on estimating the rollout-level profile from a finite batch of on-policy samples. For each rollout, define

$$\mathbf{Z}_i = \frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\rho_N(\pi_\theta(\cdot \mid s_{i,t})) \in \Delta^N, \qquad \hat{\mathbf{c}}_\theta = \frac{1}{M}\sum_{i=1}^{M}\mathbf{Z}_i.$$

Here, the statistical unit is the rollout, not the token position. Although token positions within a rollout are correlated through the autoregressive generation process, the per-rollout average $\mathbf{Z}_i$ is still a bounded random vector. Therefore, concentration can be applied across independently sampled rollouts $\mathbf{Z}_1, \ldots, \mathbf{Z}_M$, without assuming independence among positions inside each rollout. Since each coordinate of $\mathbf{Z}_i$ lies in $[0, 1]$, Hoeffding's inequality gives, for each rank channel r,

$$\Pr\left(|\hat{c}_{\theta,r} - c_{\theta,r}| \geq \epsilon_{\text{est}}\right) \leq 2\exp\left(-2M\epsilon_{\text{est}}^2\right).$$

A union bound over the $N+1$ profile coordinates gives

$$\Pr\left(\|\hat{\mathbf{c}}_\theta - \mathbf{c}_\theta\|_\infty \geq \epsilon_{\text{est}}\right) \leq 2(N+1)\exp\left(-2M\epsilon_{\text{est}}^2\right).$$

Therefore, estimating all profile coordinates within ϵ_{est} with probability at least $1-\delta$ requires

$$M = O\left(\frac{1}{\epsilon_{\text{est}}^2}\log\frac{N+1}{\delta}\right).$$

This dimension dependence is the statistical benefit of rank-space compression. A full vocabulary-level or full order-statistic profile would require a coordinate-uniform estimate over V entries, giving a $\log V$ dependence instead of $\log(N+1)$. MOSA therefore estimates a much lower-dimensional population statistic while retaining the top-rank geometry that controls finite-sample branching. This bound also inspires us that generating more roll-outs and estimating $\hat{\mathbf{c}}_\theta$ in a batch-wise manner will enable a better estimation of this quantity.

A.2 Gradient Structure: Rank-Space Feedback

MOSA produces a different gradient geometry from token-wise KL. Let $s_{i,t} = (x_i, o_{i,<t})$ denote the decoding state at position t of response o_i, and let $\mathbf{p}_{i,t} = \pi_\theta(\cdot \mid s_{i,t})$. The batch profile is

$$\hat{\mathbf{c}}_\theta = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\rho_N(\mathbf{p}_{i,t}),$$

where M is the number of sampled responses. The gradient with respect to a local distribution is

$$\nabla_{\mathbf{p}_{i,t}}\mathcal{L}_{\text{MOSA}} = \frac{1}{M|o_i|}J_{\rho_N}(\mathbf{p}_{i,t})^\top\left(-\frac{\hat{\mathbf{c}}_{\text{base}}}{\hat{\mathbf{c}}_\theta}\right).$$

where $J_{\rho_N}(\mathbf{p}_{i,t})$ is the Jacobian matrix and the vector division is elementwise. The vector $-\hat{\mathbf{c}}_{\text{base}}/\hat{\mathbf{c}}_\theta$ is shared across all visited states. Thus, MOSA first measures a batch-level deviation in rank space and then broadcasts the same correction to every state through its local ranking.

This shared correction has a simple logit-space interpretation. Let $\mathbf{p}_{i,t} = \text{softmax}(\mathbf{z}_{i,t})$, and let $\text{rank}_{i,t}(v)$ denote the rank of token v at state $s_{i,t}$. Define

$$q_{i,t,v} = \begin{cases} -\hat{c}_{\text{base,rank}_{i,t}(v)}/\hat{c}_{\theta,\text{rank}_{i,t}(v)}, & \text{rank}_{i,t}(v) \leq N, \\ -\hat{c}_{\text{base,N+1}}/\hat{c}_{\theta,\text{N+1}}, & \text{rank}_{i,t}(v) > N. \end{cases}$$

Then we have

$$\frac{\partial\mathcal{L}_{\text{MOSA}}}{\partial z_{i,t,v}} = \frac{1}{M|o_i|}p_{i,t,v}\left(q_{i,t,v} - \sum_{u=1}^{V}p_{i,t,u}q_{i,t,u}\right).$$

Therefore, the correction force depends on the token's current rank channel, not on its vocabulary identity. The model is free to change which token occupies each rank; MOSA only controls how much probability mass is allocated to rank channels on average.

When $\hat{\mathbf{c}}_\theta = \hat{\mathbf{c}}_{\text{base}}$, all rank coefficients are equal to -1; hence $q_{i,t,v}$ is constant over the vocabulary and the logit gradient vanishes. MOSA therefore does not continuously push the policy toward

higher entropy or toward base-model token preferences. It only corrects deviations from the target exploration profile.

Finally, MOSA is not additive over positions. Because all states contribute to the same $\hat{c}_\theta$, MOSA induces cross-position coupling through a low-dimensional rank-space bottleneck. Ignoring changes in the top-N ordering, the mixed second derivative between two distinct positions (i, t) and (i', t')

$$\nabla^2_{\mathbf{p}_{i,t},\mathbf{p}_{i',t'}} \mathcal{L}_{\mathrm{MOSA}} = \frac{1}{M^2|o_i||o_{i'}|} J_{\rho_N}(\mathbf{p}_{i,t})^\top \operatorname{diag}\left(\left\{ \frac{\hat{c}_{\mathrm{base},r}}{\hat{c}^2_{\theta,r}} \right\}_{r=1}^{N+1} \right) J_{\rho_N}(\mathbf{p}_{i',t'}).$$

MOSA therefore acts as a global rank-space controller rather than a collection of independent local imitation penalties.

B More Experiment Details

B.1 Training Details

For the math reasoning experiments, all models are trained using the AdamW optimizer [20] with a learning rate of 10^{-6}, a batch size of 1024, and a PPO mini-batch size of 256. For each input problem, we sample 8 responses with a temperature of 1.0. The coefficient β_{MOSA} is set to 0.05 for Qwen2.5-Math-7B, 0.5 for OctoThinker-3B, and 0.1 for Qwen2.5-32B. For GRPO w/ KL regularization, we set the KL coefficient to 0.001. All experiments are conducted on NVIDIA H200 GPUs. For the math task, training runs for 20 epochs, requiring approximately 112 GPU hours in total.

B.2 Evaluation Details

For the Countdown task, we use a temperature of 1.0 and a top-p value of 0.95, with the maximum response length set to 1024 tokens. For the math tasks, we use a temperature of 0.6 and a top-p value of 0.95. The maximum response length is set to 4096 tokens for all models, except for Qwen2.5-32B, for which we use 8192 tokens. To compute the average pass@1, we sample 512 independent runs on AIME24, AIME25, and AMC23. This provides a more comprehensive evaluation and reduces the variance caused by the relatively small sizes of these benchmarks. For larger benchmarks, including OlympiadBench, Minerva Math and MATH500, we sample 128 independent runs.

B.3 Code Reasoning Details

For code reasoning tasks, we fine-tune DeepSeek-R1-Distill-Qwen-7B on the DeepCoder [21] training dataset with a learning rate of 10^{-6}, a batch size of 128, and a PPO mini-batch size of 32. We sample 8 responses per problem at temperature 0.6, with a maximum response length of 8,192 tokens. β_{MOSA} is set to 0.01. We evaluate on LiveCodeBench [14] and HumanEval+ [18]. For each problem, we generate 16 responses with temperature 0.6 and top-p 0.95, and report average accuracy (pass@1) and pass rate (pass@16). The maximum response length for evaluation is 8,192 tokens, consistent with the training configuration.

C Additional Experiment

C.1 Additional Evaluation of Solution Diversity

We evaluate the diversity produced by MOSA from three complementary perspectives: human strategy annotations, reference-free lexical and structural metrics, and semantic judgments from multiple LLMs.

Human evaluation of strategy diversity. We manually evaluate 32 generated solutions per problem on a 100-problem subset of Countdown. To distinguish semantic diversity from superficial variation, we consider three levels of diversity. Operator-pattern diversity counts distinct combinations of arithmetic operators. Expression-layout diversity further distinguishes operand orderings and parenthesizations within the same operator pattern. Strategy-level diversity groups solutions according to their underlying problem-solving strategy based on human assessment. For each metric, we count the number of distinct groups per problem and average the result across the 100 problems.

Table 5: Human evaluation of solution diversity on a 100-problem Countdown subset, with 32 generated solutions per problem. Values report the average number of distinct operator patterns, expression layouts, and human-identified strategies per problem. Bold denotes the best result.

Method	Operator-Pattern	Expression-Layout	Strategy-Level
GRPO	1.16	1.27	1.22
GRPO w/ KL	1.07	1.34	1.44
Entropy-Adv	1.01	1.13	1.12
MOSA	**1.26**	**2.26**	**2.16**

Table 6: Reference-free lexical and structural diversity across three model backbones. Lower Self-BLEU indicates less overlap, while higher pairwise distances indicate greater output diversity. Bold denotes the best result within each model block.

Model	Method	Self-BLEU ↓	Edit Dist. ↑	ROUGE-L Dist. ↑	Jaccard Dist. ↑
Qwen2.5-Math-7B	GRPO	0.46	0.51	0.37	0.52
	GRPO w/ KL	0.48	0.49	0.35	0.49
	GSPO	0.40	0.56	0.41	0.57
	Entropy-Adv	0.54	0.41	0.30	0.41
	MOSA	**0.34**	**0.60**	**0.45**	**0.60**
OctoThinker-3B	GRPO	0.37	0.57	0.43	0.56
	GRPO w/ KL	0.38	0.54	0.40	0.55
	GSPO	0.29	0.61	0.46	0.63
	Entropy-Adv	0.34	0.58	0.43	0.57
	MOSA	**0.22**	**0.71**	**0.54**	**0.71**
Qwen2.5-32B	GRPO	0.64	0.35	0.24	0.35
	GSPO	0.36	0.58	0.43	0.57
	Entropy-Adv	0.54	0.40	0.29	0.41
	MOSA	**0.35**	**0.60**	**0.44**	**0.60**

Table 5 shows that MOSA achieves the highest diversity at all three levels. In particular, MOSA produces 2.16 distinct human-identified strategies per problem, compared with 1.22 for GRPO, 1.44 for GRPO w/ KL, and 1.12 for Entropy-Adv. The larger gain in strategy-level and expression-layout diversity indicates that MOSA generates different solution structures rather than merely changing local notation.

Reference-free lexical and structural diversity. We next evaluate diversity without relying on either human annotation or an LLM judge. For each problem, we compute Self-BLEU and three average pairwise distances among the generated solutions: normalized edit distance, ROUGE-L distance, and Jaccard distance. Self-BLEU measures overlap between responses and is lower when the solution set is more diverse. The three distance metrics measure differences in token sequences, longest-common-subsequence structure, and token sets, respectively, with higher values indicating greater diversity. We first average each metric within a problem and then report the mean across problems.

As shown in Table 6, MOSA achieves the lowest Self-BLEU and the highest pairwise distances across all three backbones. It also consistently outperforms GSPO, the strongest competing baseline under these metrics. These results establish that the diversity gains are observable directly from the generated solutions and do not depend on a particular semantic judge.

Semantic diversity under multiple LLM judges. Finally, we evaluate semantic strategy diversity using three independent judges: Claude-4.6-Sonnet, DeepSeek-V4-Flash, and Qwen3.5-397B-A17B. Each judge receives the same set of 16 correct responses and assigns a diversity score from 1 to 5 using the common evaluation rubric in Appendix D. We evaluate Qwen2.5-Math-7B and Qwen2.5-32B on AIME 2024 and AIME 2025, and OctoThinker-3B on AMC. The final score for each judge is averaged across problems, and the overall score is the arithmetic mean across the three judges.

Table 7: Semantic diversity of 16 correct responses per problem, evaluated independently by Claude-4.6-Sonnet, DeepSeek-V4-Flash, and Qwen3.5-397B-A17B on a 1–5 scale. The average is computed across the three judges. Bold denotes the best result within each model block.

Method	Claude-4.6	DeepSeek-V4	Qwen3.5-397B	Average
Qwen2.5-Math-7B				
GRPO	1.67	1.73	1.25	1.55
GRPO w/ KL	1.73	1.74	1.29	1.59
GSPO	1.76	1.81	1.37	1.65
Entropy-Adv	1.36	1.78	1.32	1.48
MOSA	**1.93**	**1.94**	**1.46**	**1.78**
OctoThinker-3B				
GRPO	1.31	1.66	1.45	1.47
GRPO w/ KL	1.32	1.46	1.39	1.39
GSPO	1.37	1.63	1.44	1.48
Entropy-Adv	1.29	1.46	1.50	1.42
MOSA	**1.39**	**2.21**	**1.79**	**1.80**
Qwen2.5-32B				
GRPO	1.10	1.61	1.15	1.29
GSPO	1.27	1.66	1.33	1.42
Entropy-Adv	1.26	1.51	1.30	1.36
MOSA	**1.34**	**1.77**	**1.41**	**1.51**

Table 7 shows that MOSA achieves the highest average semantic-diversity score on every backbone. The improvement is consistent across proprietary and open-source judges: MOSA obtains average scores of 1.78, 1.80, and 1.51 on Qwen2.5-Math-7B, OctoThinker-3B, and Qwen2.5-32B, respectively. This consistency reduces dependence on the preferences of any single evaluator.

Across the three evaluation families, MOSA consistently produces a broader set of solution structures and semantic strategies. It also outperforms Entropy-Adv throughout, showing that profile alignment preserves useful diversity more effectively than entropy-based optimization in the evaluated settings. Taken together, these results empirically connect preservation of the order-statistic profile to meaningful trajectory diversity: the profile does not directly encode semantics, but maintaining its non-concentrated structure enables the model to retain and realize distinct solution strategies under finite sampling.

C.2 Conditional Profile Analysis on Matched States

The rollout-level profile in Section 4.2 is averaged over states visited by each policy. To separate conditional changes from state-distribution drift, we evaluate the base, GRPO, and MOSA policies on the same fixed states. We collect prefixes from Qwen2.5-Math-7B base-model rollouts on 30 AIME 2024 problems, sampling 32 responses per problem, and evaluate every policy on these prefixes without generating new prefixes. For each visited position, we sort the next-token probabilities and average the resulting rank statistics across all matched positions.

Table 8: Mean conditional order-statistic probabilities on matched base-model prefixes.

Method	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5	Rank 6	Rank 7	Rank 8	Rank 9	Rank 10
Base	0.91730	0.03565	0.01021	0.00509	0.00331	0.00236	0.00176	0.00136	0.00112	0.00094
GRPO	0.96897	0.02201	0.00416	0.00153	0.00082	0.00053	0.00037	0.00028	0.00022	0.00015
MOSA	**0.92531**	**0.04228**	**0.01195**	**0.00525**	**0.00301**	**0.00197**	**0.00141**	**0.00108**	**0.00083**	**0.00066**

On the matched prefixes, GRPO increases the mean rank-1 probability from 0.91730 to 0.96897 and reduces the cumulative rank-2–10 mass from 0.06180 to 0.03007. MOSA remains close to the base profile, with rank-1 probability 0.92531 and cumulative rank-2–10 mass 0.06844. Thus, the concentration observed after GRPO is present even when all policies are evaluated conditionally on the same states, while MOSA preserves probability mass on alternative rank channels.

C.3 Variance Across Independent Training Runs

We measure training variability using three independent random seeds for GRPO and MOSA on Qwen2.5-Math-7B. All runs use the mathematical reasoning training protocol described in Appendix B, including the same MATH training data, optimizer, learning rate, batch size, number of sampled responses, and evaluation procedure. We report the mean and standard deviation across seeds for each benchmark, together with 95% confidence intervals for the benchmark-averaged metrics.

Table 9: Mean $\pm$ standard deviation across three independent seeds on Qwen2.5-Math-7B. Per-benchmark columns report pass@1.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg. P@1	Avg. P@128
GRPO	26.23±0.35	11.47±0.64	61.47±1.07	76.60±0.17	33.57±0.40	38.80±0.17	41.37±0.25	72.20±0.26
MOSA	31.33±0.81	12.43±0.06	63.27±0.06	78.10±0.30	35.57±0.45	39.77±0.42	43.37±0.21	75.07±0.21

The 95% confidence intervals for the benchmark-averaged pass@1 are $(40.74, 41.99)$ for GRPO and $(42.85, 43.88)$ for MOSA; for pass@128, they are $(71.54, 72.86)$ and $(74.55, 75.58)$, respectively. The improvements are consistent across the three independent runs.

C.4 KL vs. MSE as the Alignment Objective

We compare two choices of objective for aligning the mean order-statistic profile to the base model: KL divergence and mean squared error (MSE). Table 10 shows that both objectives improve over the GRPO baseline, confirming that aligning the order-statistic profile is beneficial regardless of the specific loss function. However, KL divergence yields consistently stronger results across benchmarks (+2.0% pass@1 and +2.8% pass@128 over GRPO on average, vs. +1.1% and +1.7% for MSE). We attribute this gap to the structure of the target statistic: the mean order-statistic profile is a valid probability distribution, so KL divergence is the natural and principled choice for measuring discrepancy in profile space, whereas MSE treats all rank channels as equally weighted real-valued coordinates and ignores this distributional structure.

Table 10: Effect of alignment objective (KL divergence vs. MSE) on pass@1 and pass@128 (Qwen2.5-Math-7B)

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
Base Model	12.9 / 64.0	5.0 / 42.3	38.1 / 93.9	56.0 / 95.4	16.3 / 64.0	25.6 / 74.4	25.7 / 72.3
+GRPO	26.6 / 70.0	12.2 / 47.1	62.4 / 91.4	76.5 / 94.0	33.1 / 59.2	38.6 / 71.1	41.6 / 72.1
+MOSA (MSE)	28.1 / 65.3	11.6 / 44.2	63.0 / 96.1	77.2 / 96.7	36.4 / 65.5	39.9 / 74.8	42.7 / 73.8
+MOSA (KL)	31.9 / 70.6	12.4 / 47.3	63.3 / 97.1	78.4 / 96.4	36.0 / 64.7	39.9 / 73.5	43.6 / 74.9

D Prompt for diversity evaluation.

Diversity Evaluation Prompt

System: You are evaluating the **DIVERSITY** of solution approaches for a mathematics competition problem. Focus on detecting even **SUBTLE** differences in methodology that indicate different problem-solving strategies.
PROBLEM:
{problem}
16 SOLUTION ATTEMPTS:
{formatted_responses}
EVALUATION CRITERIA - Rate diversity from 1 to 5:
Score 1 - Minimal Diversity:

- 14+ responses use essentially identical approaches
- Same mathematical setup, same variable choices, same solution path
- Only trivial differences: arithmetic, notation, wording
- Indicates very low exploration/diversity in the generation process

Score 2 - Low Diversity:

- 11–13 responses use the same main approach
- 1–2 alternative approaches appear but are rare
- Minor variations within the dominant method: different substitutions, orderings
- Some exploration but heavily biased toward one strategy

Score 3 - Moderate Diversity:

- 7–10 responses use the most common approach
- 2–3 distinct alternative approaches present
- Noticeable variation in problem setup or mathematical techniques
- Balanced mix showing reasonable exploration

Score 4 - High Diversity:

- 4–6 responses use the most common approach
- 3–4 distinct solution strategies well-represented
- Multiple mathematical techniques and problem framings
- Strong evidence of diverse exploration strategies

Score 5 - Maximum Diversity:

- No single approach dominates, ≤ 3 responses use same method
- 4+ distinctly different solution strategies
- Wide variety of mathematical techniques and creative approaches
- Excellent exploration and generation diversity

IMPORTANT: Focusing on the **DIVERSITY** of the attempted approaches. Return **ONLY** a number from 1 to 5.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly present our motivation and main approach in the abstract and introduction, and we also outline our contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section 6 (Conclusion).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide the full set of assumptions and complete proofs in both the main text and the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed hyperparameters in the Experiment section and describe the implementation procedure in the Methods section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Due to anonymization constraints during the review process, we do not provide public access to code and data at this time. We plan to release them upon acceptance. Meanwhile, we provide detailed implementation descriptions and experimental settings in the paper to facilitate reproducibility.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification:We provide the relevant training and evaluation details in the Experimental Details section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report pass@1 estimated from 128 samples per problem, without additional runs or statistical significance analysis.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide details about the compute resources used for our experiments, including the type of hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research presented in this paper complies with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential positive societal impacts of our work in the paper.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not involve the release of models or datasets that pose a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available datasets and models, and properly cite their original sources. We respect the corresponding licenses and terms of use.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release new datasets, models, or code as assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects, and therefore IRB approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are used only as evaluation targets or data generators, and are not an integral or novel component of the core methodology.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.