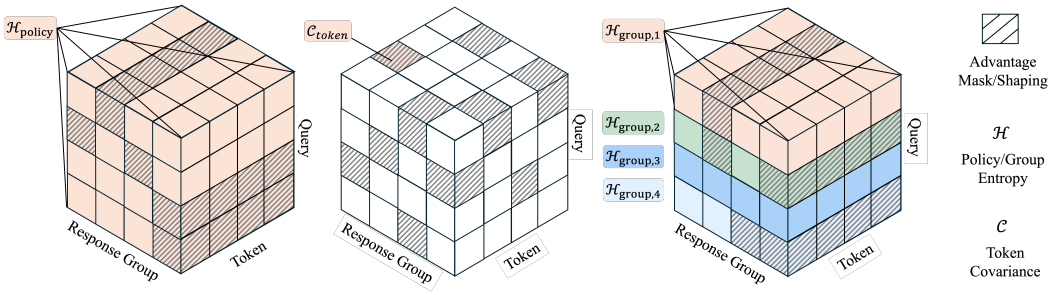
GROUP ENTROPY-CONTROLLED POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Entropy control has become an effective tool in reinforcement learning (RL) of large language models (LLMs), helping balance exploration-exploitation trade-off during alignment process. Such RL paradigm is often conducted on mixtures of heterogeneous tasks, which induce distinct entropy regimes under the same policy, making global or token-level entropy regulation insufficient to corresponding heterogeneous needs of exploration. This heterogeneity further makes GRPO-style normalized advantages induce an entropy-dependent bias, making advantage signals across prompt groups statistically non-comparable. To address this issue, we propose Group Entropy-Controlled Policy Optimization (GEPO), a lightweight extension to GRPO that uses group entropy, estimated from existing grouped samples to perform entropy-conditioned asymmetric advantage shaping. GEPO attenuates positive advantages in low-entropy groups to reduce over-exploitation, and negative advantages in high-entropy groups to preserve exploration, with adaptive thresholds derived from historical entropy statistics. Extensive experiments on two base models across thirteen benchmarks spanning mathematics, physics, science, code generation, and instruction following show that GEPO consistently outperforms GRPO and recent entropy-controlled methods, delivering balanced cross-task improvements while preserving task-specific exploration levels throughout training.



(a) Policy-level Entropy Control (b) Token-level Entropy Control (c) Group-level Entropy Control (Ours)

Figure 1: Comparison of entropy control methods at different granularities. (a): Policy-level methods calculate the entropy of policy over the current batch and unanimously apply it to regulate all responses; (b): Token-level methods calculate the covariance of tokens over the current batch and individually apply it to regulate each token; (c): GEPO calculate group entropy over the responses of each prompt and respectively regulate responses, enabling task-specific exploration-exploitation trade-off across heterogeneous tasks.

1 INTRODUCTION

Reinforcement learning (RL) has recently emerged as an effective post-training paradigm for enhancing both alignment and reasoning capabilities of Large Language Models (LLMs), yielding substantial improvements across a wide range of downstream tasks Shao et al. (2024); Team et al. (2025); Jaech et al. (2024); Hu et al. (2025). Nevertheless, balancing exploration and exploitation remains a longstanding challenge in RL Sutton et al. (1998); Bellemare et al. (2016), which becomes increasingly pronounced in LLM post-training due to the optimization over heterogeneous task mixtures. Specifically, modern LLM post-training is typically performed on mixtures of diverse tasks, spanning instruction following, coding, writing, and mathematical reasoning, which differ

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.

substantially in structure, solution diversity, and uncertainty of policy exploration. Therefore, a central challenge in applying RL to multi-task mixtures is exploration-exploitation trade-off across heterogeneous tasks.

Recent RL methods seek to balance exploration and exploitation through entropy-controlled mechanisms operating either at the policy level (Figure 1(a)) Yang et al. (2025); Shen et al. (2025) or at the token level (Figure 1(b)) Cheng et al. (2026); Cui et al. (2025). However, these approaches primarily treat entropy as a global indicator for sustaining stable RL training, overlooking the optimization imbalance among heterogeneous tasks. Specifically, Group Relative Policy Optimization (GRPO), one of the dominant algorithms in LLM RL, standardizes rewards within each prompt group to construct relative advantages, implicitly treating the resulting advantage signals as unbiased across groups. Yet whether this assumption remains valid when prompt groups exhibit substantially different levels of entropy, and how such entropy heterogeneity affects joint multi-task optimization, remain largely under-explored in existing entropy-controlled methods.

In this work, we make the first attempt to delve deep into this problem, and reveal that the group entropy (*i.e.*, the entropy calculated from the response groups of each task) indeed varies from their corresponding task characteristics, which may implicitly induces a structural bias in GRPO's normalized advantages, causing the optimization of different tasks imbalanced. Based on this insight, we propose Group Entropy-Controlled Policy Optimization (GEPO) (Figure 1(c)), a lightweight and model-agnostic extension to related group-based policy optimization methods that performs entropy-conditioned asymmetric advantage shaping at the group level. The key idea is to use group entropy as a diagnostic signal to identify and mitigate the optimization bias induced by entropy heterogeneity. Specifically, GEPO attenuates positive advantages in low-entropy groups to prevent over-exploitation that would further amplify the entropy gap, while attenuating negative advantages in high-entropy groups to avoid prematurely suppressing exploration. This shaping is asymmetric because we empirically find that low-entropy groups are more susceptible to aggressive intervention, which may trigger length collapse, and therefore require milder attenuation than high-entropy groups. Furthermore, the entropy boundaries for advantage shaping are adaptively derived from historical entropy statistics and smoothed with exponential moving average, enabling automatic calibration across base models and training stages without task-specific tuning.

Extensive experiments on Intern-S1-mini and Qwen3.5-9B across thirteen diverse benchmarks demonstrate that GEPO consistently improves both average performance and inter-task balance, achieving state-of-the-art results among GRPO and entropy-aware RL methods. Further analysis further shows that GEPO induces more stable optimization trajectories, and consistently benefits tasks with diverse exploration states. Notably, these gains are obtained with zero additional sampling cost and can be directly incorporated into existing group-based policy optimization methods.

2 Related Work

Policy entropy has long been studied in reinforcement learning as a measure of action uncertainty and exploration, with entropy regularization and maximum-entropy objectives serving as fundamental mechanisms for encouraging exploration in stochastic policies Haarnoja et al. (2018). More recently, entropy dynamics have attracted increasing attention in the post-training of LLMs, where researchers investigate their relationship with exploration, training stability, and reasoning performance Cui et al. (2025). Several studies have investigated the relationship between entropy dynamics and reasoning performance, suggesting that entropy evolution is closely associated with exploration behaviors, training stability, and the risk of premature convergence. From an empirical perspective, Cheng et al. (2026) characterize the role of high-entropy regions in facilitating reflective behaviors, pivotal reasoning steps, and exploration of alternative solution trajectories, while Cui et al. (2025) analyze entropy collapse during RL training and connect excessive entropy reduction with degraded reasoning performance. Beyond empirical observations, Wang et al. (2026) provide a theoretical analysis of entropy evolution under policy optimization and its impact on training dynamics. Motivated by these findings, recent approaches have introduced explicit entropy regulation mechanisms to improve RL training stability and exploration efficiency. Yang et al. (2025) propose entropy control strategies to stabilize long-horizon RL optimization, while Wan et al. (2026) preserve exploration diversity through global and local diversity regularization. However, existing entropy-aware approaches primarily operate at the policy or token level, treating entropy as a global or local property of the policy distribution. The impact of entropy variations across sampled groups, and their interaction with group-wise optimization mechanisms such as GRPO, remains largely unexplored.

3 METHOD

3.1 PRELIMINARY

Group Relative Policy Optimization (GRPO) (Shao et al. (2024)) is a reinforcement learning (RL) optimization framework for large language model (LLM) post-training that eliminates the need for a separate value function by estimating advantages from grouped samples. Given a prompt x, GRPO samples K responses $\{y_1, y_2, \ldots, y_K\}$ from the current policy π_θ, obtains rewards $\{r_1, r_2, \ldots, r_K\}$, and computes normalized advantages:

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^K)}{\text{std}(\{r_j\}_{j=1}^K)}, \quad i = 1, \ldots, K. \tag{1}$$

The GRPO objective maximizes the clipped surrogate:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{x,\{y_i\}} \left[\frac{1}{K} \sum_{i=1}^{K} \frac{1}{T_i} \sum_{t=1}^{T_i} \left(\min \left(\rho_{i,t} A_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) A_i \right) \right) \right], \tag{2}$$

where $\rho_{i,t} = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$ is the importance sampling ratio, and T_i is the length of response y_i.

3.2 GROUP ENTROPY AND OPTIMIZATION DYNAMICS

We define group entropy, a per-response-group signal capturing task-level policy exploration, and study its empirical properties under GRPO grouped sampling. For a prompt x, we define the sequence-level entropy as

$$H(\pi_\theta(\cdot \mid x)) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)}[-\log \pi_\theta(y \mid x)] = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[-\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, x) \right], \tag{3}$$

which directly quantifies policy uncertainty for prompt x: low values imply π_θ concentrates probability on few responses (potential over-exploitation), while high values mean π_θ spreads probability over diverse responses (active exploration with possible optimization instability).

However, computing $H(\pi_\theta(\cdot|x))$ is intractable for LLMs with the large vocabulary size and response length for reasoning tasks. The expectation form of Equation 3 admits a natural Monte Carlo estimator. In GRPO, group responses $\{y_1, \ldots, y_K\}$ are sampled i.i.d. from $\pi_\theta(\cdot|x)$ for each prompt x. Therefore, we define **group entropy** as an estimator of the group-based sequence-level entropy:

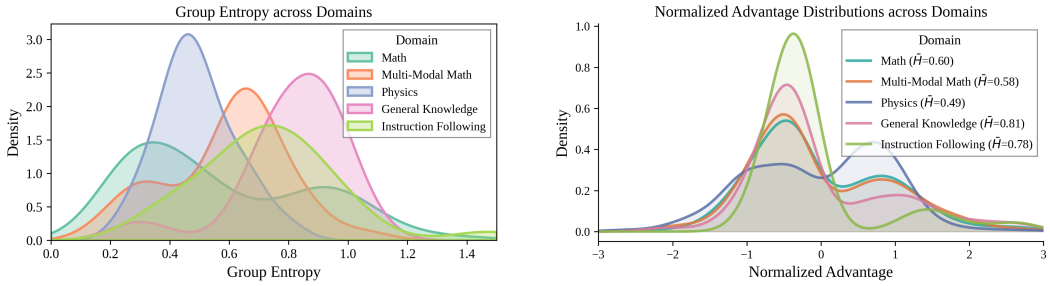
$$\hat{H}_{\text{g}}(x) \ = \ -\frac{1}{K} \sum_{i=1}^{K} \sum_{t=1}^{T_i} \log \pi_\theta(y_{i,t} \mid y_{i,<t}, x). \tag{4}$$

As a standard Monte Carlo estimator, $\hat{H}_{\text{g}}(x)$ is an unbiased estimate of $H(\pi_\theta(\cdot \mid x))$, with variance decreasing as $O(K^{-1})$. In practice, we normalize by the average response length to obtain the per-token group entropy $\mathcal{H}_{\text{g}}(x) = \hat{H}_{\text{g}}(x)/\bar{T}$, where $\bar{T} = \frac{1}{K}\sum_{i=1}^{K} T_i$, to improve comparability across groups with different response lengths.

We now present two empirical observations from the first rollout of GRPO training. Specifically, group entropy is computed for each response group associated with a prompt, while semantic domains are used only to aggregate and visualize these group-level quantities. These observations motivate the method design in Section 3.3. Detailed experimental settings are in Section 4.1.

Observation 1: group entropy is heterogeneous across domains. Figure 2a shows the domain-conditional distributions of group entropy. Under the same base policy, prompts from different domains occupy markedly different entropy regimes: for example, physics prompts concentrate around $\mathcal{H}_{\text{g}} \approx 0.49$, whereas general knowledge prompts are centered near $\mathcal{H}_{\text{g}} \approx 0.81$. Consequently, a single global entropy statistic cannot adequately characterize the exploration states of all prompts in a multi-domain training mixture.

Observation 2: Domain-level entropy regimes coincide with distinct advantage profiles. Figure 2b shows that domains occupying different group entropy regimes also exhibit markedly different normalized-advantage distributions:



(a) group entropy by domain.

(b) Normalized advantage distributions by domain.

Figure 2: Preliminary observations from the first GRPO rollout before any policy update ($K = 16$ responses per prompt). **(a)** Domain-conditional distributions of group entropy. **(b)** Domain-conditional distributions of GRPO-normalized advantages. Together, the panels show that semantic domains occupy different exploration and credit-assignment regimes within the same training mixture.

- The low-entropy Physics domain ($\bar{\mathcal{H}}_g = 0.49$, accuracy 48%) produces a nearly *symmetric bimodal* advantage distribution (skewness ≈ 0), with balanced positive and negative signals.
- The medium-entropy Math domain ($\bar{\mathcal{H}}_g = 0.60$, accuracy 39%) exhibits moderate right skew (≈ 0.75), with more negative than positive advantages.
- The high-entropy Instruction Following domain ($\bar{\mathcal{H}}_g = 0.78$, accuracy 16%) yields extreme right skew (skewness $= 2.32$, kurtosis $= 4.84$): the vast majority of responses cluster near a small negative advantage, while the rare correct responses become extreme positive outliers.

The observed asymmetry is not caused by entropy alone, but arises mechanically from the lower group accuracy that accompanies the high-entropy regime in this rollout. When group accuracy $p \to 0$, the zero-mean constraint forces positive advantages to concentrate on rare correct responses with magnitude $|A_+| = \sqrt{(1-p)/p} \gg 1$, while negative advantages are diluted across the majority with $|A_-| = \sqrt{p/(1-p)} \ll 1$ (see Appendix A for the full derivation). Beyond this intra-group asymmetry, we analyze a more fundamental *inter-group* issue: standard group normalization does not guarantee comparable advantage signals across groups with different entropy levels.

To formalize this, let ν_x denote a reference measure over the response space $\mathcal{A}_x$ (*e.g.*, the uniform distribution over feasible responses), and define the *reference reward cumulative distribution function (CDF)* $F_x^{\nu}(t) := \nu_x(\{a : \mu_x(a) \le t\})$ and the *policy-weighted reward CDF* $F_x^{\pi}(t) := \pi_x(\{a : \mu_x(a) \le t\})$, where $\mu_x(a) = \mathbb{E}[R \mid x, a]$.

Proposition 3.1 (Entropy-Dependent Distribution Distortion)**.** *For any prompt x with conditional policy entropy $H(x) = -\sum_a \pi_x(a) \log \pi_x(a)$ and response space of size $N_x = |\mathcal{A}_x|$, the Kolmogorov distance between the policy-weighted and reference reward distributions satisfies*

$$\sup_t |F_x^{\pi}(t) - F_x^{\nu}(t)| \;\le\; \mathrm{TV}(\pi_x, \nu_x) \;\le\; \sqrt{\frac{1}{2}(\log N_x - H(x))}, \tag{5}$$

where the second inequality holds when ν_x is the uniform distribution over $\mathcal{A}_x$, following from Pinsker's inequality (Tsybakov (2009)) applied to $\mathrm{KL}(\pi_x \| \nu_x) = \log N_x - H(x)$.

Proposition 3.1 shows that low policy entropy permits a larger worst-case deviation between the policy-weighted and reference reward distributions. Thus, low-entropy prompts can be more sensitive to policy-induced reweighting of the response space, although the bound does not imply that the maximal deviation is attained for every prompt.

Proposition 3.2 (Structural Bias of Normalized Advantages)**.** *Let $Z_x^{\pi}(a) = (\mu_x(a) - m_x^{\pi})/\sigma_x^{\pi}$ denote the advantage under policy-weighted moments, and $Z_x^{\nu}(a) = (\mu_x(a) - m_x^{\nu})/\sigma_x^{\nu}$ the advantage under reference moments. Assume $|\mu_x(a)| \le M$ and $\sigma_x^{\pi}, \sigma_x^{\nu} \ge \sigma_0 > 0$ for all x, a. Then there exists a constant $C = C(M, \sigma_0) > 0$ such that*

$$\sup_{a \in \mathcal{A}_x} |Z_x^{\pi}(a) - Z_x^{\nu}(a)| \;\le\; C\,\mathrm{TV}(\pi_x, \nu_x) \;\le\; C\sqrt{\frac{1}{2}(\log N_x - H(x))}. \tag{6}$$

Proposition 3.2 further shows that per-group centering and scaling do not guarantee a common statistical reference frame across prompts: the potential discrepancy between policy-weighted and reference-normalized advantages remains entropy dependent.

Optimization consequences. The intra-group skewness (Observation 2) and inter-group structural bias (Propositions 3.1–3.2) jointly create entropy-dependent optimization pressures at two levels:

Within each high-entropy group, the advantage asymmetry produces two concurrent pathologies. First, *diffuse and noisy suppression*: incorrect responses receive small negative advantages across many diverse reasoning paths, yielding unreliable suppression that prematurely penalizes exploratory trajectories with valuable partial reasoning. Second, *inconsistent reinforcement*: rare correct responses receive large positive advantages, but different rollouts often surface different reasoning paths, so the reinforcement signal does not accumulate coherently over training steps.

Across groups, the structural bias (Proposition 3.2) further compounds this imbalance: low-entropy groups, whose advantages are most distorted relative to the true reward landscape, nonetheless produce the strongest and most consistent gradient signals, while high-entropy groups—despite their advantages being closer to the reference frame—contribute weaker, noisier gradients. As a result, the batch-level optimization is systematically dominated by the already-converging low-entropy tasks. This creates a self-reinforcing cycle: as low-entropy groups improve, their accuracy rises and entropy drops further, widening the entropy gap and causing high-entropy groups to receive diminishing effective learning signal—ultimately leading to persistent inter-task imbalance.

This analysis motivates an entropy-conditioned intervention that rebalances the advantage signals across entropy regimes, rather than treating all normalized advantage distributions identically.

3.3 Group Entropy-Controlled Policy Optimization

Consequently, we propose our group entropy-controlled policy optimization method. For a prompt x with group entropy $\mathcal{H}_{\text{g}}(x)$ and advantages $\{A_i\}_{i=1}^{K}$, we define a group entropy-based coefficient to shape the advantage as

$$\hat{A}_i = \omega(A_i, \mathcal{H}_{\text{g}})A_i = \begin{cases} \alpha_{\text{low}} \cdot A_i & \text{if } A_i > 0 \text{ and } \mathcal{H}_{\text{g}}(x) < \mathcal{H}_{\text{low}}^{(t)}, \\ \alpha_{\text{high}} \cdot A_i & \text{if } A_i < 0 \text{ and } \mathcal{H}_{\text{g}}(x) > \mathcal{H}_{\text{high}}^{(t)}, \\ A_i & \text{otherwise,} \end{cases} \tag{7}$$

where $\alpha_{\text{high}}, \alpha_{\text{low}} \in (0, 1)$ are the scaling coefficients, and $\mathcal{H}_{\text{low}}^{(t)}$ and $\mathcal{H}_{\text{high}}^{(t)}$ denote the adaptive lower and upper entropy thresholds at training step t (defined in the following paragraph).

Compared with global policy entropy, Equation 7 offers several notable advantages. First, it is also task-aware: even within the same domain, prompts may vary considerably in difficulty, and group entropy provides a finer-grained characterization at the task level. Second, group entropy is naturally compatible with the grouped sampling structure of GRPO, allowing it to be incorporated without introducing additional sampling procedures or computational overhead beyond that already required by GRPO. Finally, by conditioning on group entropy, the advantage shaping attenuates the structural bias characterized by Propositions 3.1–3.2: it attenuates the over-exploitative signals from low-entropy groups that would otherwise drive premature convergence, and prevents the diluted negative signals in high-entropy groups from suppressing exploration before the policy discover correct reasoning paths, thereby promoting more balanced learning across entropy regimes.

Asymmetric Advantage Shaping. We impose the constraint $\alpha_{\text{high}} < \alpha_{\text{low}}$ based on the following empirical observation: in the low-entropy regime, the model is already confident and the policy distribution is sharply peaked. In this regime, we empirically find that aggressively penalizing negative advantages (which would further increase entropy) may cause length collapse—a pathological behavior where the model dramatically shortens its responses to reduce per-token uncertainty. By using a higher α_{low} for positive advantage attenuation in the low-entropy regime, we apply a gentler nudge that encourages exploration without triggering this instability. This asymmetry reflects the difference in the fragility of the two entropy regimes.

Adaptive Entropy Thresholds. Fixed entropy thresholds are particularly brittle in multi-task LLM post-training. Suitable entropy range depends on the task mixture, initial policy, and training stage: different base models can exhibit substantially different entropy scales and controllability on the same tasks, and the entropy distribution further shifts as the policy improves. Therefore, we define

entropy thresholds relative to the empirical group entropy distribution of the current batch. At step t, we compute the batch mean $\mu_H^{(t)}$ and standard deviation $\sigma_H^{(t)}$ of group entropy, and set

$$\hat{\mathcal{H}}_{\text{low}}^{(t)} = \mu_H^{(t)} - \beta_{\text{low}}\sigma_H^{(t)}, \qquad \hat{\mathcal{H}}_{\text{high}}^{(t)} = \mu_H^{(t)} + \beta_{\text{high}}\sigma_H^{(t)}. \tag{8}$$

Here, β_{low} and β_{high} control the width of the lower and upper entropy margins, respectively. We then apply exponential moving averages to obtain temporally stable thresholds:

$$\mathcal{H}_{\text{low}}^{(t+1)} = (1-\gamma)\mathcal{H}_{\text{low}}^{(t)} + \gamma\hat{\mathcal{H}}_{\text{low}}^{(t)}, \qquad \mathcal{H}_{\text{high}}^{(t+1)} = (1-\gamma)\mathcal{H}_{\text{high}}^{(t)} + \gamma\hat{\mathcal{H}}_{\text{high}}^{(t)}. \tag{9}$$

The smoothing coefficient $\gamma \in (0, 1)$ governs the trade-off between responsiveness and stability. This yields a distribution-aware and temporally smooth entropy band that adapts to model-specific entropy scales and changing exploration regimes.

4 Experiments

4.1 Settings

We select Intern-S1-mini and Qwen3.5-9B as our base models. Intern-S1-mini is a lightweight open-source multimodal reasoning model based on the same techniques as Intern-S1 Bai et al. (2025), and Qwen3.5-9B is an open-source text-based reasoning model. We use an in-house multimodal dataset in the training process. The dataset is designed to cover diverse task categories and knowledge domains, including mathematics, instruction following, engineering reasoning, physics, and chemistry. Our implementation is based on LMDeploy Contributors (2023a) and Xtuner Contributors (2023b). During rollouts, we set the temperature to 1 and sample 16 responses per prompt. Training follows an off-policy RL setup with a batch size of 256 and a minibatch size of 32. For hyperparameter settings, the scaling coefficients $\alpha_{\text{high}}, \alpha_{\text{low}}$ are set as 0.2 and 0.5. The adaptive entropy thresholds coefficients $\beta_{\text{high}}, \beta_{\text{low}},$ and γ are set as 0.3, 0.2 and 0.01.

We evaluate model performance on diverse benchmarks spanning mathematical reasoning, scientific reasoning, code generation, instruction following, and multimodal understanding, including GPQA Rein et al. (2023), LiveCodeBench Jain et al. (2024), IFBench Pyatkin et al. (2025), AIME2025, IMO-Bench-AnswerBench, CMPhysBench Wang et al. (2025), MMMU-Pro Yue et al. (2025), MathVista Lu et al. (2024), Physics Feng et al. (2025), SFE Zhou et al. (2025), MicroVQA Burgess et al. (2025), and ChartQAPro Masry et al. (2025). This comprehensive benchmark suite evaluates general reasoning and robustness across diverse domains, task formats, and multimodal settings.

We compare our approach against GRPO, AEPO Shen et al. (2025), Clip-Cov, and KL-Cov Cui et al. (2025). For Clip-Cov, and KL-Cov, we use the same set of parameters as for Qwen2.5-7B from Cui et al. (2025). Besides, all other sampling and training parameters are consistent with the GEPO implementation. To avoid the issues of token dropping, we clip the importance sampling weight as proposed in CISPO Chen et al. (2025), which are also implemented in baseline methods.

4.2 Main Results

Performance Across Multiple Benchmarks. As shown in Table 1, existing methods exhibit notable learning imbalance across tasks. AEPO achieves strong improvements on mathematics but yields limited gains or even underperformance on physics, instruction following, and scientific reasoning. This suggests that its optimization pressure is disproportionately concentrated on certain task categories at the expense of others. Similarly, GRPO shows moderate improvements on some tasks but limited or negative transfer on others, *e.g.*, CMPhysBench and MMLU-Pro. Clip-Cov and KL-Cov provide stronger coverage-aware regularization and improve several individual benchmarks, but their gains remain uneven across task categories. On Intern-S1-mini, Clip-Cov improves IMO-Bench and CMPhysBench, while KL-Cov further improves code generation and ChartQAPro. However, both methods still underperform GEPO in average performance, achieving 51.5 and 51.8 compared with GEPO's 54.2, and show clear regressions on tasks such as Physics and SFE.

In contrast, GEPO achieves the best performance on 7 out of 13 benchmarks on Intern-S1-mini, spanning mathematical reasoning, physics, instruction following, and scientific understanding. For example, GEPO attains substantial gains on challenging reasoning benchmarks such as AIME25, IMO-Bench, and GPQA, while simultaneously improving instruction following and scientific reasoning. GEPO maintains strong gains across most benchmarks and delivers the best overall average, demonstrating that the group entropy mechanism balances optimization across heterogeneous tasks.

Table 1: Main results on thirteen benchmarks across baselines and our methods

	MATH			Physics		Instruction Following	Code
	AIME25	IMO-Bench	MathVista	Physics	CMPhysBench		
Intern-S1-mini	74.6	42.3	70.5	37.4	29.2	29.8	39.4
+ GRPO	75.7	45.5	70.0	37.8	28.0	34.3	38.9
+ AEPO	79.5	47.5	**71.5**	36.1	28.2	32.2	38.3
+ Clip-Cov	76.7	49.8	69.8	32.4	30.6	31.2	38.3
+ KL-Cov	76.4	49.8	71.2	32.6	**31.2**	33.5	**40.0**
+ GEPO (ours)	**82.0**	**52.8**	71.0	**39.1**	30.4	**38.8**	39.4
Qwen3.5-9B	82.5	58.5	85	51.0	35.8	62.2	57.7
+ GRPO	90.1	67.3	**86.4**	56.8	41.2	80.7	**68.0**
+ AEPO	84.2	63.8	85.3	53.2	41.5	68.0	60.6
+ Clip-Cov	**91.9**	69.3	85.5	**59.0**	42.9	**81.3**	64.0
+ KL-Cov	89.9	66.8	85.4	55.9	41.1	80.2	64.6
+ GEPO (ours)	91.4	**69.5**	84.9	58.8	**43.2**	80.7	67.4

	Knowledge & Science				Visual & Chart		**Avg.**
	MMLU-Pro	GPQA	MMMU-Pro	SFE	ChartQAPro	MicroVQA	
Intern-S1-mini	75.0	65.1	55.0	45.0	47.5	50.4	50.9
+ GRPO	74.3	64.3	55.7	44.7	48.5	51.8	51.5
+ AEPO	74.7	62.5	57.4	**46.5**	46.7	**52.5**	51.8
+ Clip-Cov	75.0	65.7	55.8	43.9	48.6	51.5	51.5
+ KL-Cov	75.1	64.7	55.8	42.8	**49.7**	50.8	51.8
+ GEPO (ours)	**77.1**	**69.2**	**59.8**	43.9	49.0	51.5	**54.2**
Qwen3.5-9B	81.6	81.9	70.1	53.1	63.4	62.9	65.0
+ GRPO	83.4	**85.3**	74.8	57.1	64.9	62.9	70.7
+ AEPO	82.5	81.6	70.9	55.2	64.0	61.4	67.1
+ Clip-Cov	83.3	85.0	76.3	57.7	62.9	63.1	70.9
+ KL-Cov	**83.9**	83.3	73.5	55.6	**65.4**	62.8	69.9
+ GEPO (ours)	83.1	83.7	**76.6**	**60.3**	62.4	**64.1**	**71.2**

For Qwen3.5-9B, GEPO further demonstrates its generalization capability, achieving the highest average score. Clip-Cov is a particularly strong baseline on this model, obtaining the best results on AIME25, Physics, and instruction following; KL-Cov leads on MMLU-Pro and ChartQAPro. Nevertheless, GEPO matches or surpasses these coverage-based baselines in overall performance and achieves the best results on IMO-Bench, CMPhysBench, MMMU-Pro, SFE, and MicroVQA. In contrast, AEPO shows significant degradation on Qwen3.5-9B compared to GRPO, suggesting that its uniform entropy control is particularly fragile when transferred to a base model with different entropy characteristics. GEPO's adaptive thresholds calibrate to a new model's entropy scale without task-specific tuning, verifying the model-agnostic adaptive entropy mechanism.

Training Stability. Beyond final performance, we examine the training dynamics through validation curves (Figure 3). GEPO exhibits smooth and monotonically improving validation performance across both models and all benchmarks, with notably less variance compared to the baselines.

For Intern-S1-mini, GEPO shows a clear upward trend on all four benchmarks with minimal fluctuation, whereas GRPO oscillates without clear improvement and AEPO suffers intermittent performance drops. The contrast is even more pronounced on Qwen3.5-9B: GRPO undergoes catastrophic performance collapse around step 170, with AIME25 accuracy dropping from $\sim$85% to $\sim$40% and simultaneous degradation on GPQA, MMMU-Pro, and MathVista. This collapse is symptomatic of unconstrained entropy dynamics under GRPO—without task-aware regulation, the policy entropy can grow unboundedly (as shown in Figure 4), eventually leading to degenerate outputs. AEPO avoids complete collapse but exhibits persistent oscillations and generally underperforms GEPO, indicating that uniform entropy control is insufficient for heterogeneous task mixtures. Clip-Cov and KL-Cov improve stability over unconstrained GRPO by introducing covaiance-aware constraints, but they still exhibit task-dependent fluctuations and plateau earlier on several validation benchmarks. This suggests that controlling covaiance between action probability and the change in logits alone does not fully resolve the instability caused by heterogeneous task difficulty and task-specific exploration requirements. Besides, KL-Cov shows training collapse on Qwen3,5-9B after 100 training steps,

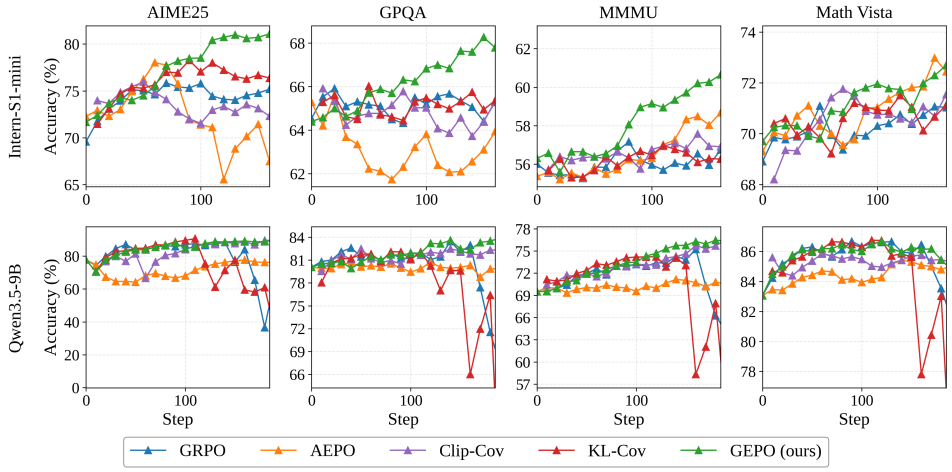


Figure 3: Typical training validation curves. For Intern-S1-mini and Qwen3.5-9B, GEPO improves smoothly over training, while baselines show strong oscillations or late-stage performance collapse.

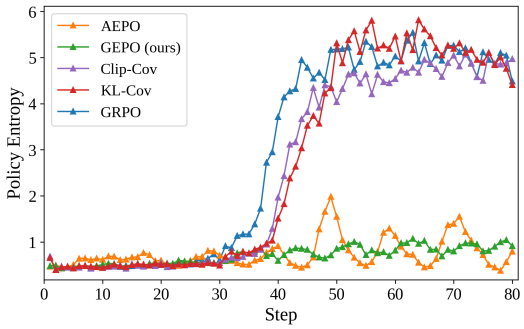


Figure 4: Policy entropy dynamics of Intern-S1-mini during training. GEPO maintains a stable, bounded entropy trajectory, while baselines show runaway entropy growth or persistent oscillations.

which is likely due to the fact that KL-Cov uses a fixed entropy threshold for all tasks, which may not be suitable for all tasks. In contrast, GEPO maintains stable and steadily improving performance throughout training on both models, demonstrating robustness against the training instabilities that affect the baselines. The stability of GEPO can be attributed to its group entropy control with adaptive thresholds: by independently regulating the exploration-exploitation balance for each prompt group, GEPO prevents the runaway entropy dynamics that trigger catastrophic collapse while maintaining sufficient exploration to sustain learning progress across all tasks.

Training Dynamics of Entropy. We further analyze how entropy evolves during training to understand the mechanism of GEPO. As shown in Figure 4, GEPO maintains the healthy policy entropy throughout training, avoiding the premature entropy collapse that can lead to mode-seeking behavior and loss of generalization. In contrast, GRPO suffers from runaway entropy growth, causing the model to produce degenerate and incoherent reasoning outputs, while Clip-Cov and KL-Cov show similar patterns of instability, with high entropy due to the lack of task-specific regulation. AEPO, while avoiding complete collapse, exhibits substantial entropy oscillations throughout training, reflecting the instability of its uniform entropy control under heterogeneous task distributions.

More importantly, the group entropy dynamics (Figure 5) reveal a key distinction between GEPO and AEPO. AEPO, which applies a uniform entropy control signal, drives the entropy of different tasks toward a similar convergence pattern, effectively imposing a one-size-fits-all exploration regime regardless of task characteristics. GEPO, in contrast, preserves differentiated exploration levels across tasks: tasks requiring broader exploration maintain relatively higher group entropy, while tasks with more deterministic solution patterns settle at lower entropy levels. This heterogeneous entropy profile is consistent with the design of GEPO, which uses group entropy as a prompt-level diagnostic rather than forcing a uniform entropy target across all tasks. This explains why GEPO

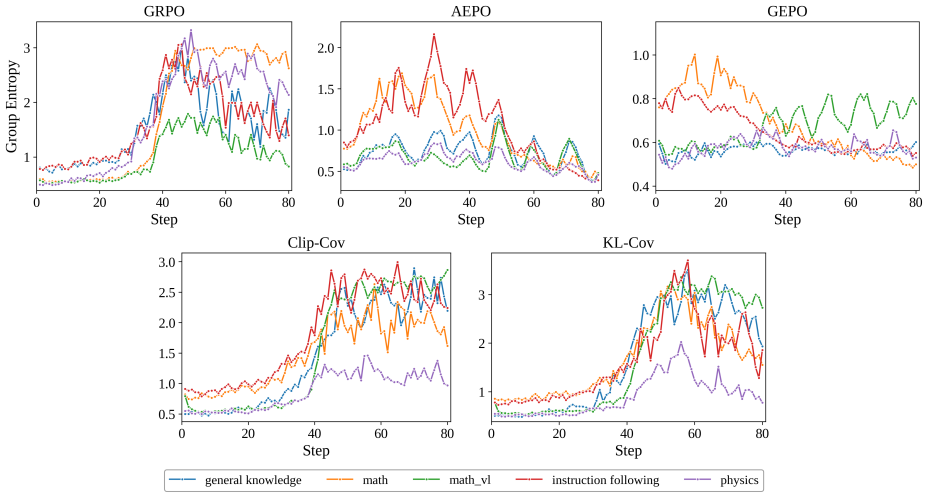


Figure 5: Task entropy dynamics during training process of Intern-s1-mini. GEPO maintains stable, domain-specific entropy regimes, while the baselines either destabilize or homogenize them.

Table 2: Ablation results on asymmetric advantage shaping and bidirectional entropy control with Intern-S1-mini.

	MATH			Physics		Inst. Follow.	Code
	AIME25	IMO-Bench	MathVista	Physics	CMPhysBench		
GEPO	**82.0**	**52.8**	71.2	**39.1**	30.4	**38.8**	**39.4**
w/o low entropy control	79.1	48.3	**73.1**	34.3	**32.2**	34.2	36.0
w/o high entropy control	76.3	50.0	70.7	30.0	28.4	35.4	36.0
w/o asymmetric adv. shaping	79.4	49.8	73.4	33.4	31.2	35.3	36.0

	Knowledge & Sci.				Visual & Chart		**Avg.**
	MMLU-Pro	GPQA	MMMU-Pro	SFE	ChartQAPro	MicroVQA	
GEPO	**77.1**	**69.2**	55.8	43.9	49.0	**51.5**	**54.2**
w/o low entropy control	77.1	68.3	56.9	**46.1**	48.3	50.4	52.6
w/o high entropy control	74.6	64.5	56.0	46.3	47.9	50.5	51.3
w/o asymmetric adv. shaping	76.6	67.7	**60.1**	45.3	**49.7**	50.9	53.0

achieves balanced performance improvements across diverse tasks (Dimension 1) with stable training dynamics (Dimension 2): the group entropy mechanism respects the inherent heterogeneity of multi-task learning and provides task-specific regulation without explicit task annotations.

Ablation Study. We conduct ablation studies, as shown in Table 2, removing any component leads to performance degradation, confirming that all components are essential. We observe that high entropy control contributes the most, as without it the policy suffers from unconstrained entropy growth and loses exploitation capability on hard reasoning tasks such as Physics and AIME25. Low entropy control prevents premature entropy collapse and maintains exploration diversity, particularly benefiting IMO-Bench and instruction following. In addition, asymmetric advantage shaping with outperforms symmetrical Advantage Shaping, as applying a gentler nudge.

5 Conclusion

We introduce Group Entropy-Controlled Policy Optimization (GEPO), a robust multi-task RL algorithm for LLMs. GEPO uses rollout-estimated group entropy for asymmetric advantage shaping, moderating low-entropy reinforcement and high-entropy suppression via adaptive entropy boundaries. It controls the exploration–exploitation trade-off without task labels or extra sampling. On two base models and thirteen benchmarks, GEPO achieves the best average performance, stable optimization trajectories, and balanced cross-task improvements. Our analysis shows preserving differentiated exploration regimes surpasses uniform entropy regularization, verifying group entropy for scalable, stable multi-task RL post-training.

REFERENCES

Lei Bai, Zhongrui Cai, Maosong Cao, Weihan Cao, Chiyu Chen, Haojiong Chen, Kai Chen, Pengcheng Chen, Ying Chen, Yongkang Chen, Yu Cheng, Yu Cheng, Pei Chu, Tao Chu, Erfei Cui, Ganqu Cui, Long Cui, Ziyun Cui, Nianchen Deng, Ning Ding, Nanqin Dong, Peijie Dong, Shihan Dou, Sinan Du, Haodong Duan, Caihua Fan, Ben Gao, Changjiang Gao, Jianfei Gao, Songyang Gao, Yang Gao, Zhangwei Gao, Jiaye Ge, Qiming Ge, Lixin Gu, Yuzhe Gu, Aijia Guo, Qipeng Guo, Xu Guo, Conghui He, Junjun He, Yili Hong, Siyuan Hou, Caiyu Hu, Hanglei Hu, Jucheng Hu, Ming Hu, Zhouqi Hua, Haian Huang, Junhao Huang, Xu Huang, Zixian Huang, Zhe Jiang, Lingkai Kong, Linyang Li, Peiji Li, Pengze Li, Shuaibin Li, Tianbin Li, Wei Li, Yuqiang Li, Dahua Lin, Junyao Lin, Tianyi Lin, Zhishan Lin, Hongwei Liu, Jiangning Liu, Jiyao Liu, Junnan Liu, Kai Liu, Kaiwen Liu, Kuikun Liu, Shichun Liu, Shudong Liu, Wei Liu, Xinyao Liu, Yuhong Liu, Zhan Liu, Yinquan Lu, Haijun Lv, Hongxia Lv, Huijie Lv, Qidang Lv, Ying Lv, Chengqi Lyu, Chenglong Ma, Jianpeng Ma, Ren Ma, Runmin Ma, Runyuan Ma, Xinzhu Ma, Yichuan Ma, Zihan Ma, Sixuan Mi, Junzhi Ning, Wenchang Ning, Xinle Pang, Jiahui Peng, Runyu Peng, Yu Qiao, Jiantao Qiu, Xiaoye Qu, Yuan Qu, Yuchen Ren, Fukai Shang, Wenqi Shao, Junhao Shen, Shuaike Shen, Chunfeng Song, Demin Song, Diping Song, Chenlin Su, Weijie Su, Weigao Sun, Yu Sun, Qian Tan, Cheng Tang, Huanze Tang, Kexian Tang, Shixiang Tang, Jian Tong, Aoran Wang, Bin Wang, Dong Wang, Lintao Wang, Rui Wang, Weiyun Wang, Wenhai Wang, Yi Wang, Ziyi Wang, Ling-I Wu, Wen Wu, Yue Wu, Zijian Wu, Linchen Xiao, Shuhao Xing, Chao Xu, Huihui Xu, Jun Xu, Ruiliang Xu, Wanghan Xu, GanLin Yang, Yuming Yang, Haochen Ye, Jin Ye, Shenglong Ye, Jia Yu, Jiashuo Yu, Jing Yu, Fei Yuan, Bo Zhang, Chao Zhang, Chen Zhang, Hongjie Zhang, Jin Zhang, Qiaosheng Zhang, Qiuyinzhe Zhang, Songyang Zhang, Taolin Zhang, Wenlong Zhang, Wenwei Zhang, Yechen Zhang, Ziyang Zhang, Haiteng Zhao, Qian Zhao, Xiangyu Zhao, Xiangyu Zhao, Bowen Zhou, Dongzhan Zhou, Peiheng Zhou, Yuhao Zhou, Yunhua Zhou, Dongsheng Zhu, Lin Zhu, and Yicheng Zou. Intern-s1: A scientific multimodal foundation model, 2025. URL https://arxiv.org/abs/2508.15763.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research, 2025. URL https://arxiv.org/abs/2503.13399.

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

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30377–30385, 2026.

LMDeploy Contributors. Lmdeploy: A toolkit for compressing, deploying, and serving llm. https://github.com/InternLM/lmdeploy, 2023a.

XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023b.

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.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving, 2025. URL https://arxiv.org/abs/2503.21821.

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

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

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, 2024. URL https://arxiv.org/abs/2403.07974.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts, 2024. URL https://arxiv.org/abs/2310.02255.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartqapro: A more diverse and challenging benchmark for chart question answering, 2025. URL https://arxiv.org/abs/2504.05506.

Valentina Pyatkin, Saumya Malik, Victoria Graf, Hamish Ivison, Shengyi Huang, Pradeep Dasigi, Nathan Lambert, and Hannaneh Hajishirzi. Generalizing verifiable instruction following, 2025. URL https://arxiv.org/abs/2507.02833.

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

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.

Weizhou Shen, Ziyi Yang, Chenliang Li, Zhiyuan Lu, Miao Peng, Huashan Sun, Yingcheng Shi, Shengyi Liao, Shaopeng Lai, Bo Zhang, et al. Qwenlong-l1. 5: Post-training recipe for long-context reasoning and memory management. *arXiv preprint arXiv:2512.12967*, 2025.

Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

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

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

Zhongwei Wan, Yun Shen, Zhihao Dou, Donghao Zhou, Yu Zhang, Xin Wang, Hui Shen, Jing Xiong, Chaofan Tao, Zixuan Zhong, et al. Dsdr: Dual-scale diversity regularization for exploration in llm reasoning. *arXiv preprint arXiv:2602.19895*, 2026.

Shumin Wang, Yuexiang Xie, Wenhao Zhang, Yuchang Sun, Yanxi Chen, Yaliang Li, and Yanyong Zhang. On the entropy dynamics in reinforcement fine-tuning of large language models. *arXiv preprint arXiv:2602.03392*, 2026.

Weida Wang, Dongchen Huang, Jiatong Li, Tengchao Yang, Ziyang Zheng, Di Zhang, Dong Han, Benteng Chen, Binzhao Luo, Zhiyu Liu, Kunling Liu, Zhiyuan Gao, Shiqi Geng, Wei Ma, Jiaming Su, Xin Li, Shuchen Pu, Yuhan Shui, Qianjia Cheng, Zhihao Dou, Dongfei Cui, Changyong He, Jin Zeng, Zeke Xie, Mao Su, Dongzhan Zhou, Yuqiang Li, Wanli Ouyang, Yunqi Cai, Xi Dai, Shufei Zhang, Lei Bai, Jinguang Cheng, Zhong Fang, and Hongming Weng. Cmphysbench: A benchmark for evaluating large language models in condensed matter physics, 2025. URL https://arxiv.org/abs/2508.18124.

Kai Yang, Xin Xu, Yangkun Chen, Weijie Liu, Jiafei Lyu, Zichuan Lin, Deheng Ye, and Saiyong Yang. Entropic: Towards stable long-term training of llms via entropy stabilization with proportional-integral control. *arXiv preprint arXiv:2511.15248*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. URL https://arxiv.org/abs/2409.02813.

Yuhao Zhou, Yiheng Wang, Xuming He, Ao Shen, Ruoyao Xiao, Zhiwei Li, Qiantai Feng, Zijie Guo, Yuejin Yang, Hao Wu, Wenxuan Huang, Jiaqi Wei, Dan Si, Xiuqi Yao, Jia Bu, Haiwen Huang, Manning Wang, Tianfan Fu, Shixiang Tang, Ben Fei, Dongzhan Zhou, Fenghua Ling, Yan Lu, Siqi Sun, Chenhui Li, Guanjie Zheng, Jiancheng Lv, Wenlong Zhang, and Lei Bai. Scientists' first exam: Probing cognitive abilities of mllm via perception, understanding, and reasoning, 2025. URL https://arxiv.org/abs/2506.10521.

A Advantage Asymmetry under Per-Group Normalization

We formalize how per-group normalization produces structurally different advantage distributions depending on the group accuracy, and quantify the resulting signal concentration effect described in Section 3.2.

Setup. Consider a prompt group with K responses and binary rewards $r_i \in \{0, 1\}$, where n_+ responses are correct ($r_i = 1$) and $n_- = K - n_+$ are incorrect ($r_i = 0$). (Since the normalized advantage $A_i = (r_i - \bar{r})/\sigma_r$ is invariant to affine transformations of the reward, this binary formulation applies without loss of generality to any two-valued reward scheme, *e.g.*, $r_i \in \{-1, +1\}$.) The group accuracy is $p = n_+/K$. GRPO's per-group normalization yields advantages:

$$A_i = \frac{r_i - \bar{r}}{\sigma_r}, \qquad \text{where } \bar{r} = p, \quad \sigma_r = \sqrt{p(1-p)}. \tag{10}$$

Substituting, correct responses receive $A_+ = \frac{1-p}{\sqrt{p(1-p)}} = \sqrt{\frac{1-p}{p}}$, and incorrect responses receive $A_- = \frac{-p}{\sqrt{p(1-p)}} = -\sqrt{\frac{p}{1-p}}$.

Signal concentration. The total positive signal equals $n_+ \cdot A_+ = K\sqrt{p(1-p)}$, which is symmetric and equals the magnitude of the total negative signal $n_- \cdot |A_-|$ by the zero-mean constraint. However, the *per-response* signal magnitudes are asymmetric:

$$|A_+| = \sqrt{\frac{1-p}{p}}, \qquad |A_-| = \sqrt{\frac{p}{1-p}}. \tag{11}$$

As $p \to 0$, $|A_+| \to \infty$ while $|A_-| \to 0$: the positive signal concentrates on a small number of correct responses, whereas the negative signal is distributed across many incorrect responses with small per-response magnitude. In our multi-task rollouts, such low-accuracy groups are empirically associated with higher group-entropy regimes.

Skewness. The skewness of the advantage distribution in a single group with accuracy p is:

$$\text{skew}(A) = \frac{1-2p}{\sqrt{p(1-p)}}. \tag{12}$$

This is zero when $p = 0.5$ (perfectly balanced), positive when $p < 0.5$ (more incorrect than correct), and grows without bound as $p \to 0$. For our observed task accuracies: Physics ($p \approx 0.48$) yields

skewness ≈ 0.08, Math ($p \approx 0.39$) yields ≈ 0.45, and Instruction Following ($p \approx 0.16$) yields ≈ 1.86.

Remark: pooled vs. single-group skewness. Equation 12 gives the skewness for a *single* group with accuracy exactly p. The empirical skewness values reported in Observation 2 of the main text (Physics ≈ 0, Math ≈ 0.75, IF ≈ 2.32) are instead computed by pooling advantages across *all* groups of a given task. Since different prompts within the same task have different accuracies p_j, the pooled skewness equals $\frac{1}{N}\sum_j \text{skew}(p_j)$ rather than $\text{skew}(\bar{p})$. Because $\text{skew}(p) = (1-2p)/\sqrt{p(1-p)}$ is *convex* on $(0, 0.5)$, Jensen's inequality gives

$$\frac{1}{N}\sum_{j=1}^{N} \text{skew}(p_j) \;\geq\; \text{skew}\left(\frac{1}{N}\sum_{j=1}^{N} p_j\right),$$

with equality only when all groups share the same accuracy. This accounts for the quantitative gap between the formula predictions and the empirical values, and implies that heterogeneous group accuracies *amplify* the advantage asymmetry beyond what the single-group analysis predicts.

B Structural Bias from Entropy Heterogeneity

This appendix provides the formal setup and complete proofs for Propositions 3.1 and 3.2 stated in the main text, as well as a finite-sample error decomposition that clarifies the distinct sources of estimation error in group-based advantage computation.

B.1 Formal Setup

For a prompt x, let $\mathcal{A}_x$ denote the finite response space induced by the vocabulary and maximum generation length, with $|\mathcal{A}_x| = N_x$. We define

$$\pi_x(a) := \pi_\theta(a \mid x), \qquad \mu_x(a) := \mathbb{E}[R \mid x, a].$$

During training, group responses are sampled independently from the policy:

$$a_1, \ldots, a_K \overset{\text{i.i.d.}}{\sim} \pi_x.$$

We introduce a reference measure ν_x over $\mathcal{A}_x$. When explicitly stated, we take $\nu_x(a) = 1/N_x$. The corresponding reward CDFs are

$$F_x^\nu(t) := \nu_x(\{a : \mu_x(a) \leq t\}), \qquad F_x^\pi(t) := \pi_x(\{a : \mu_x(a) \leq t\}).$$

We define two reward cumulative distribution functions:

- **Reference reward CDF:** $F_x^\nu(t) := \nu_x(\{a : \mu_x(a) \leq t\})$
- **Policy-weighted reward CDF:** $F_x^\pi(t) := \pi_x(\{a : \mu_x(a) \leq t\})$

Standard group-based methods observe samples from F_x^π, not F_x^ν. The discrepancy between the two is governed by how far the policy deviates from the reference measure.

B.2 Proof of Proposition 3.1 (Distribution Distortion Bound)

Theorem B.1 (Restatement)**.** *For any prompt x,*

$$\sup_t |F_x^\pi(t) - F_x^\nu(t)| \;\leq\; \text{TV}(\pi_x, \nu_x).$$

If ν_x is the uniform distribution over $\mathcal{A}_x$, then

$$\text{TV}(\pi_x, \nu_x) \;\leq\; \sqrt{\frac{1}{2}(\log N_x - H(x))}.$$

Proof. For any threshold t, define the level set $B_t := \{a \in \mathcal{A}_x : \mu_x(a) \le t\}$. Then

$$F_x^\pi(t) - F_x^\nu(t) = \pi_x(B_t) - \nu_x(B_t).$$

Taking the supremum over all t (equivalently, over all level sets B_t):

$$\sup_t |F_x^\pi(t) - F_x^\nu(t)| \;\le\; \sup_{B \subseteq \mathcal{A}_x} |\pi_x(B) - \nu_x(B)| \;=\; \mathrm{TV}(\pi_x, \nu_x).$$

The first inequality holds because level sets $\{B_t\}$ form a subset of all measurable sets.

For the second part, when ν_x is uniform:

$$\mathrm{KL}(\pi_x \| \nu_x) = \sum_a \pi_x(a) \log \frac{\pi_x(a)}{1/N_x} = \sum_a \pi_x(a) \log \pi_x(a) + \log N_x = \log N_x - H(x).$$

Applying Pinsker's inequality (Tsybakov (2009)) $\mathrm{TV}(P, Q) \le \sqrt{\frac{1}{2}\mathrm{KL}(P\|Q)}$:

$$\mathrm{TV}(\pi_x, \nu_x) \;\le\; \sqrt{\frac{1}{2}(\log N_x - H(x))}.$$

□

B.3 Proof of Proposition 3.2 (Advantage Bias Bound)

Define the policy-weighted and reference moments:

$$m_x^\pi := \mathbb{E}_{\pi_x}[\mu_x(a)], \quad (\sigma_x^\pi)^2 := \mathrm{Var}_{\pi_x}(\mu_x(a)), \quad m_x^\nu := \mathbb{E}_{\nu_x}[\mu_x(a)], \quad (\sigma_x^\nu)^2 := \mathrm{Var}_{\nu_x}(\mu_x(a)),$$

and the corresponding standardized advantages:

$$Z_x^\pi(a) = \frac{\mu_x(a) - m_x^\pi}{\sigma_x^\pi}, \qquad Z_x^\nu(a) = \frac{\mu_x(a) - m_x^\nu}{\sigma_x^\nu}.$$

Theorem B.2 (Restatement)**.** *Under the assumptions $|\mu_x(a)| \le M$ and $\sigma_x^\pi, \sigma_x^\nu \ge \sigma_0 > 0$, there exists $C = C(M, \sigma_0) > 0$ such that*

$$\sup_{a \in \mathcal{A}_x} |Z_x^\pi(a) - Z_x^\nu(a)| \;\le\; C\,\mathrm{TV}(\pi_x, \nu_x).$$

Proof. **Step 1: Mean difference.**

$$|m_x^\pi - m_x^\nu| = \left| \sum_a (\pi_x(a) - \nu_x(a)) \mu_x(a) \right| \le 2M \cdot \mathrm{TV}(\pi_x, \nu_x),$$

where we used $\sum_a |p(a) - q(a)| = 2\,\mathrm{TV}(p, q)$ and $|\mu_x(a)| \le M$.

Step 2: Second moment difference. Let $u_x^\pi := \mathbb{E}_{\pi_x}[\mu_x(a)^2]$ and $u_x^\nu := \mathbb{E}_{\nu_x}[\mu_x(a)^2]$. Similarly:

$$|u_x^\pi - u_x^\nu| \le 2M^2 \cdot \mathrm{TV}(\pi_x, \nu_x).$$

Step 3: Variance difference. Since $(\sigma_x^\pi)^2 = u_x^\pi - (m_x^\pi)^2$ and $(\sigma_x^\nu)^2 = u_x^\nu - (m_x^\nu)^2$:

$$\begin{aligned} |(\sigma_x^\pi)^2 - (\sigma_x^\nu)^2| &\le |u_x^\pi - u_x^\nu| + |(m_x^\pi)^2 - (m_x^\nu)^2| \\ &\le 2M^2 \cdot \mathrm{TV} + |m_x^\pi + m_x^\nu| \cdot |m_x^\pi - m_x^\nu| \\ &\le 2M^2 \cdot \mathrm{TV} + 2M \cdot 2M \cdot \mathrm{TV} \\ &= 6M^2 \cdot \mathrm{TV}(\pi_x, \nu_x). \end{aligned}$$

Using $|a^2 - b^2| = |a - b||a + b| \ge |a - b| \cdot \sigma_0$ (since $\sigma_x^\pi + \sigma_x^\nu \ge \sigma_0$):

$$|\sigma_x^\pi - \sigma_x^\nu| \le \frac{6M^2}{\sigma_0} \cdot \mathrm{TV}(\pi_x, \nu_x) =: C_1 \cdot \mathrm{TV}.$$

Step 4: Advantage difference. For any $a \in \mathcal{A}_x$:

$$
\begin{aligned}
|Z_x^\pi(a) - Z_x^\nu(a)| &= \left| \frac{\mu_x(a) - m_x^\pi}{\sigma_x^\pi} - \frac{\mu_x(a) - m_x^\nu}{\sigma_x^\nu} \right| \\
&\leq \frac{|m_x^\pi - m_x^\nu|}{\sigma_x^\pi} + |\mu_x(a) - m_x^\nu| \cdot \left| \frac{1}{\sigma_x^\pi} - \frac{1}{\sigma_x^\nu} \right| \\
&\leq \frac{2M \cdot \mathrm{TV}}{\sigma_0} + 2M \cdot \frac{|\sigma_x^\pi - \sigma_x^\nu|}{\sigma_0^2} \\
&\leq \frac{2M}{\sigma_0} \cdot \mathrm{TV} + \frac{2M \cdot C_1}{\sigma_0^2} \cdot \mathrm{TV} \\
&= \underbrace{\left(\frac{2M}{\sigma_0} + \frac{12M^3}{\sigma_0^3} \right)}_{=: C} \cdot \mathrm{TV}(\pi_x, \nu_x).
\end{aligned}
$$

□