

From Imitation to Discrimination: Toward A Generalized Curriculum Advantage Mechanism Enhancing Cross-Domain Reasoning Tasks

Changpeng Yang^{1,*}, Jinyang Wu^{2,*}, Yuchen Liu¹, Shuai Zhang², Yang Li¹, Qiliang Liang³, Hongzhen Wang¹, Shuai Nie¹, Jiaming Xu^{1,†}, Runyu Shi¹, Ying Huang¹, Guoquan Zhang¹

¹Xiaomi Corporation,

²Tsinghua University,

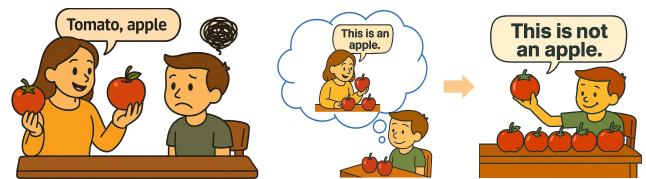
³Peking University

Abstract

Reinforcement learning has emerged as a paradigm for post-training large language models, boosting their reasoning capabilities. Such approaches compute an advantage value for each sample, reflecting better or worse performance than expected, thereby yielding both positive and negative signals for training. However, existing approaches often mix them indiscriminately, especially in the early stages, leading to ambiguous guidance and limited gains. To address this issue, we propose **CAPO** (Curriculum Advantage Policy Optimization), an adaptive curriculum mechanism based on advantage signals. The proposed mechanism bootstraps imitation learning with positive-only advantage samples to establish robust foundations, and subsequently introduces negative signals to cultivate discriminative capabilities, thereby improving generalization across complex scenarios. Compatible with diverse optimization methods including GRPO, PPO, RLOO, and Reinforce++, our method consistently achieves stable and significant improvements in mathematical reasoning tasks, and further generalizes effectively to multimodal Graphical User Interface (GUI) reasoning scenarios, establishing itself as a versatile and robust optimization paradigm.

Introduction

Reinforcement learning (RL) has become a mainstream paradigm for post-training large language models, substantially advancing their reasoning capabilities, as demonstrated by DeepSeek-R1 (Guo et al. 2025) and Kimi-1.5 (Team et al. 2025). A critical component of RL algorithms such as PPO (Schulman et al. 2017) and GRPO (Shao et al. 2024b) is the advantage, which quantifies whether a trajectory performs above or below expectation, providing positive and negative feedback to guide policy updates. Yet, simultaneous training on both positive and negative advantage samples often introduces ambiguity, especially during early optimization, limiting further improvement. This challenge calls for probing the essence of advantage and rethinking its role in shaping training dynamics. Since advantage inherently reflects whether the model’s competence is better



(a) RL: Mixed signals hinder learning.

(b) CAPO: Positive imitation builds stability.

Figure 1: Comparison of RL and CAPO. (a) RL mixes positive and negative signals. (b) CAPO uses staged curriculum: positive imitation builds stability, negative discrimination improves generalization.

or worse than expected, it naturally prompts our central research question: *can advantage itself serve as a guidance indicator, enabling structured integration of positive and negative feedback into a unified, generalizable paradigm?*

Inspired by developmental psychology, we note that children’s learning progresses through gradual stages: they first acquire basic behaviors through *positive imitation*, and only later advance their generalization by incorporating *corrective feedback and punishment* (Bandura, Ross, and Ross 1961; Whitehurst 1969). this staged progression naturally positions advantage as an effective curriculum signal.

Motivated by this perspective, we propose **CAPO** (Curriculum Advantage Policy Optimization), a training paradigm broadly compatible with advantage-based reinforcement learning algorithms. It adopts a two-phase curriculum learning strategy based on advantage signals: in the imitation phase, positive samples foster stable foundations, and in the discriminative phase, positive samples sustain correct behaviors while negative samples refine learning, together improving generalization. As illustrated in Figure 1, panel (a) shows that mixing positive and negative signals from the start leads to early confusion and prevents stable progress. To clarify this phenomenon, panel (b) provides a simple intuition, depicting how separating signals in stages avoids such interference. Together, these panels highlight the necessity of a staged curriculum design in CAPO.

While children’s learning provides intuitive motivation,

*These authors contributed equally.

†Corresponding Author

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

we provide a theoretical basis for this curriculum through the lens of the *variance–bias tradeoff*. The imitation phase reduces gradient variance, fostering stable early training, whereas the discrimination phase restores unbiasedness, enabling effective generalization.

We further highlight CAPO’s unique design: it leverages *advantage* as a dynamic signal that aligns with the model’s evolving competence. Since curriculum learning was first introduced by Bengio et al. (Bengio et al. 2009a), most methods have relied on static heuristics such as sorting tasks from easy to hard. Subsequent adaptive curriculum strategies estimate difficulty through expert annotations or model success rates (Shi et al. 2025; Chen et al. 2025), but they remain external and heuristic, relying on manually defined proxies rather than signals intrinsic to the model’s evolving capability. This limitation motivates our design of CAPO, which leverages advantage estimates as an intrinsic, competence-aware signal to drive dynamic curriculum scheduling.

Extensive experiments show that CAPO consistently enhances mathematical reasoning across diverse advantage-based RL algorithms, including GRPO, PPO, RLOO, and Reinforce++. It further demonstrates strong modality generalization, achieving notable gains on GUI-based reasoning tasks. Together, these results highlight CAPO as a versatile curriculum paradigm, effective across multiple advantage-based RL methods and transferable to multimodal reasoning domains. In summary, our main contributions are threefold:

- We introduce *advantage* as a dynamic curriculum signal and design CAPO, a two-phase mechanism with imitation and discrimination phases adapting to the model’s evolving competence.
- We demonstrate CAPO’s broad generalization, proving effective across diverse advantage-based RL algorithms and transferable to multimodal reasoning tasks on GUI-based environments.
- We validate CAPO with extensive experiments, achieving improvements ranging from +2.4 to +4.0 on 1.5B models and from +1.7 to +3.9 on 7B models across diverse advantage-based RL algorithms, and also +3.81 on multimodal GUI-based reasoning tasks.

Method

We present CAPO, as shown in Figure 2, an advantage-based compatible mechanism. Once samples are generated by the policy model, various algorithms compute their respective advantage estimates, which are then seamlessly unified within our curriculum advantage framework. Leveraging these signals as guidance, CAPO explicitly structures training into an imitation phase and a discrimination phase, thereby reconciling the need for early-stage stability with the demand for enhanced generalization once negative samples are introduced.

Policy Optimization in Reinforcement Learning. In RL for language models, the generation process can be formulated as a policy π_θ that maps input prompts \mathbf{q} to output

sequences \mathbf{o} . The training objective is to maximize the expected reward assigned to the generated sequences:

$$\mathcal{J}(\theta) = \mathbb{E}_{\mathbf{q} \sim p_{\mathcal{Q}}, \mathbf{o} \sim \pi_\theta(\cdot|\mathbf{q})} [R(\mathbf{q}, \mathbf{o})] \quad (1)$$

The policy gradient theorem (Sutton, Barto et al. 1998) allows reinforcement learning objectives to be optimized via gradient ascent:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(\mathbf{o}|\mathbf{q}) \cdot Q^{\pi_\theta}(\mathbf{q}, \mathbf{o})], \quad (2)$$

where $Q^{\pi_\theta}(\mathbf{q}, \mathbf{o})$ denotes the action-value function, representing the expected return for generating sequence \mathbf{o} from prompt \mathbf{q} under policy π_θ .

To reduce variance in gradient estimates, it is common to replace $Q^{\pi_\theta}(\mathbf{q}, \mathbf{o})$ with the advantage function:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(\mathbf{o}|\mathbf{q}) \cdot A^{\pi_\theta}(\mathbf{q}, \mathbf{o})], \quad (3)$$

where $A^{\pi_\theta}(\mathbf{q}, \mathbf{o})$ measures whether a sampled trajectory outperforms the expected baseline under π_θ .

In practice, different policy optimization methods adopt distinct strategies to estimate or construct the advantage. For example, PPO employs generalized advantage estimation (GAE) (Schulman et al. 2015), which balances bias and variance by leveraging a discounted sum of temporal difference residuals. GRPO introduces a group-relative advantage formulation that normalizes advantage values within grouped samples to stabilize training dynamics. Building on these foundations, we treat the advantage as a general-purpose signal that not only drives gradient updates but also dynamically structures the training curriculum.

To stabilize training, modern methods such as PPO optimize a clipped surrogate objective, replacing the raw policy gradient

$$\mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta} \left[\frac{1}{T} \sum_{t=1}^T \min(\rho_t(\theta) A_t, \hat{\rho}_t(\theta) A_t) \right], \quad (4)$$

where $\rho_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the importance sampling ratio, and $\hat{\rho}_t(\theta) = \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon)$ restricts updates to a stable range.

Curriculum Advantage Policy Optimization

We introduce **Curriculum Advantage Policy Optimization (CAPO)**, a progressive training framework that leverages the advantage as both a gradient weight and a curriculum signal. Intuitively, a positive advantage indicates that the model is competent on the sample, whereas a negative advantage suggests insufficient capability. Rather than applying static or heuristic curricula, CAPO dynamically structures learning in two stages.

Phase 1: Imitation Phase with Positive-Only Advantage Samples. Training begins with a positive-only imitation phase ($A_\tau \geq 0$), ensuring that updates are guided by beneficial trajectories. This design consolidates prior knowledge and avoids unstable gradients that might arise from prematurely exposing the model to challenging samples. Formally, building on Eq. 4 and incorporating a KL regularization term

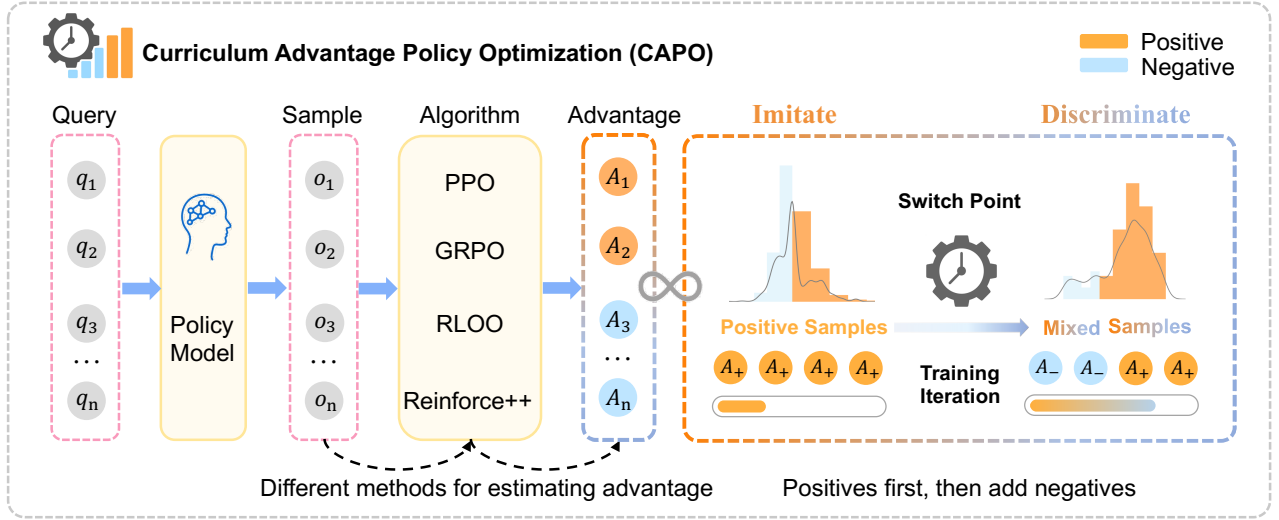


Figure 2: Illustration of the CAPO scheduling mechanism. Each query is processed by the policy model to generate samples, with advantages computed under different optimization algorithms. In Phase 1, only positive-advantage samples are used to ensure stability; after the switch point, Phase 2 incorporates both positive and negative advantages to balance stability and generalization.

to prevent policy collapse as in RLHF (Guo et al. 2025), the objective is defined as:

$$\mathcal{J}_{\text{phase-1}}(\theta) = \mathbb{E}_{\tau} \left[\mathbb{I}_{A(\tau) \geq 0} \left(\frac{1}{T} \sum_{t=1}^T \min(\rho_t A_t, \hat{\rho}_t A_t) - \beta \mathbb{D}_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}) \right) \right]. \quad (5)$$

where the indicator $\mathbb{I}_{A_{\tau} \geq 0}$ filters out negative-advantage samples, and β controls the KL penalty strength. This phase thus encourages the model to reinforce correct reasoning behaviors while remaining close to the reference distribution.

Phase 2: Discriminative Phase with Full Advantage Spectrum. Once a stable foundation is established, CAPO transitions into a discriminative phase that admits the full advantage spectrum. By incorporating negative-advantage samples, the model learns not only to reinforce strong reasoning trajectories but also to suppress suboptimal ones, thereby enhancing generalization. The corresponding objective is: This progressive shift from imitation to discrimination ensures that CAPO first stabilizes learning and then promotes robust reasoning by leveraging both positive and negative feedback.

Curriculum Scheduling Between Phases. To ensure a seamless transition between the imitation and discrimination phases, CAPO adopts a two-stage training strategy with a predefined switch point (e.g., at 10% or 20% of total training steps). We also experimented with gradually introducing negative signals, but found that no such progressive scheme matched the effectiveness of a simple switch point. In practice, a hard switch provides a robust and task-agnostic mechanism that consistently balances early-stage stability with later discriminative learning, without requiring delicate hyperparameter tuning or task-specific monitoring. This

pragmatic design ensures reproducibility across diverse settings while still delivering the intended variance reduction in Phase 1 and generalization gains in Phase 2.

Theoretical Justification

The CAPO algorithm’s two-phase curriculum leverages the variance–bias tradeoff in policy gradient estimation to enhance training stability and convergence. Consider a policy $\pi_{\theta}(a|s)$ with advantage estimate $\hat{A}(s, a) = A^{\pi}(s, a) + \epsilon$, where ϵ is zero-mean noise with variance σ^2 . The true policy gradient g and its stochastic estimate \hat{g} are defined as:

$$g = \mathbb{E}_{\pi_{\theta}} [\nabla_{\theta} \log \pi_{\theta}(a|s) A^{\pi}(s, a)] \quad (6)$$

$$\hat{g} = \nabla_{\theta} \log \pi_{\theta}(a|s) \hat{A}(s, a) \quad (7)$$

The mean squared error (MSE) between \hat{g} and g decomposes into bias and variance:

$$\mathbb{E}[\|\hat{g} - g\|^2] = \|\mathbb{E}[\hat{g}] - g\|^2 + \text{Var}(\hat{g}). \quad (8)$$

Phase 1 (Positive-only Imitation). To suppress early-stage noise, CAPO restricts updates to positive advantages:

$$\hat{g}_{\text{phase-1}} = \mathbb{E} \left[\nabla_{\theta} \log \pi_{\theta}(a|s) \hat{A}(s, a) \mathbb{I}\{\hat{A} > 0\} \right]. \quad (9)$$

Excluding negative outliers reduces $\text{Var}(\hat{g})$, so even with bias, the overall MSE is lowered, ensuring stable improvement.

Phase 2 (Full Discriminative Refinement). As the policy improves and $\text{Var}(\hat{A})$ shrinks, CAPO transitions to the unbiased estimator:

$$\hat{g}_{\text{phase-2}} = \mathbb{E} \left[\nabla_{\theta} \log \pi_{\theta}(a|s) \hat{A}(s, a) \right], \quad (10)$$

restoring $\mathbb{E}[\hat{g}_{\text{phase-2}}] = g$ and enabling generalization.

Proposition. Let $\{\alpha_t\}$ denote the step sizes. Under Robbins–Monro conditions ($\sum_t \alpha_t = \infty$, $\sum_t \alpha_t^2 < \infty$), the CAPO update rule converges almost surely to a local optimum: Phase 1 bounds variance, Phase 2 restores unbiasedness, so the MSE in (8) vanishes asymptotically and the limit point of θ_t is stationary. A detailed proof is provided in Appendix A.

Experimental Setups

We present setups for both *mathematical reasoning* and *GUI-based multimodal* tasks, covering datasets, evaluation, and baselines.

Mathematical Reasoning Tasks

Datasets. Following prior works (Zeng et al. 2025; Liu et al. 2025; Wu et al. 2025), we curate 5.5K Level 3–5 problems from the MATH dataset (Hendrycks et al. 2021).

Evaluation Benchmarks. We evaluate on AIME 2024 (Li et al. 2024a), AMC (Li et al. 2024a), MATH500 (Hendrycks et al. 2021), GSM8K (Cobbe et al. 2021), Minerva (Lewkowycz et al. 2022), OlympiadBench (He et al. 2024) and College Math (Tang et al. 2024). For out-of-domain evaluation, we additionally include ARC-C (Clark et al. 2018) and GPQA-Diamond (Rein et al. 2024). Inference uses greedy decoding with a 3K token budget (Liu et al. 2025; Wu et al. 2025).

Baselines. We evaluate CAPO across four mainstream reinforcement learning algorithms: GRPO, PPO, RLOO, and Reinforce++. For each baseline, we additionally include its CAPO-augmented variant, enabling a direct assessment of CAPO’s effectiveness as a general enhancement. This design allows us to examine both the standalone performance of the baselines and the improvements achieved through CAPO integration.

Implementation. To ensure fairness across the various baseline algorithms and CAPO variants, we unify the description of all implementation details. Because different methods require distinct parameterizations, the complete training configurations are provided in Appendix B.

GUI-based Multimodal Tasks

Datasets. We adopt GUI-R1-3K (Luo et al. 2025), a dataset derived from OS-Atlas (Wu et al. 2024b), consisting of 3K vision-language-action examples across web and mobile interfaces. The dataset pairs visual states with natural language instructions and action sequences, providing a compact yet diverse testbed for evaluating multimodal reasoning and planning.

Evaluation Benchmarks. *Planning:* GUI-Act-Web (Luo et al. 2025), OmniAct-Web (Kapoor et al. 2024), AndroidControl-Low/High (Li et al. 2024b). *Perception:* Screenspot Pro (Li et al. 2025).

Baselines. We compare GRPO with its CAPO-augmented variant, following the dominant GUI benchmark setup (Wu et al. 2024b). This serves as a supplementary evaluation to

demonstrate the effectiveness of CAPO in multimodal reasoning scenarios, beyond the primary focus on mathematical reasoning tasks.

Implementation. We train QwenVL2.5-3B (Team 2025) in the VERL framework (Sheng et al. 2025). All experiments are conducted on 8×NVIDIA A100-80GB GPUs. To ensure reproducibility, detailed hyperparameters and training configurations are provided in Appendix B.

For evaluating planning ability, we leverage four datasets across web, mobile, and desktop platforms: GUI-Act-Web (Luo et al. 2025), OmniAct-Web, AndroidControl-Low, and AndroidControl-High (Li et al. 2024b; Lu et al. 2024). These benchmarks involve long-horizon action prediction, decision-making under partial observability, and cross-app navigation, covering both short- and long-term planning tasks in real-world interactive environments. We further conduct qualitative case studies on GUI-based multimodal tasks to illustrate the behavioral differences between baselines and CAPO-enhanced models. Representative examples are presented in Appendix D.

Experimental Results

Main Results

Math Reasoning Performance. Following prior works (Liu et al. 2025; Wu et al. 2025), we evaluate CAPO on seven reasoning benchmarks across two model scales and four optimization methods—GRPO (Shao et al. 2024a), PPO (Schulman et al. 2017), RLOO (Ahmadian et al. 2024), and REINFORCE++ (Hu et al. 2025).

As shown in Table 1, CAPO delivers consistent gains of +1.7 to +4.0 points across all methods, confirming its effectiveness as a drop-in enhancement. Notably, it achieves large improvements on competition-level tasks: for instance, AMC improves from 52.5 to 65.0 (+12.5) and AIME24 from 16.7 to 20.0 (+3.3) on the 7B model, with the 1.5B model also achieving absolute improvements of 2.4 4.0. Beyond individual datasets, consistent improvements are observed on GSM8K, OlympiadBench, Minerva, among others, demonstrating robust generalization.

CAPO further scales well with model size: while the 7B model attains the highest performance, the 1.5B variant with CAPO closes much of the gap, approaching the larger baseline.

Overall, CAPO shows (1) broad applicability across optimization methods, (2) scalable effectiveness that benefits both small and large models, and (3) substantial improvements on high-difficulty benchmarks, supporting advantage-based curricula as a principled mechanism to unify stability and exploration.

Multimodal Perception and Planning Capabilities of GUI Agent. We adopt GUI-based tasks as our primary benchmark due to their complexity as cross-domain reasoning challenges in multimodal settings. These tasks require precise language understanding, fine-grained visual perception, and context integration to plan and execute actions effectively, providing a rigorous testbed for CAPO’s robustness and adaptability.

Method	AIME24 \uparrow	AMC \uparrow	MATH500 \uparrow	GSM8K \uparrow	Minerva \uparrow	Olympiad \uparrow	CollegeMath \uparrow	Avg. \uparrow
Qwen2.5-7B-Math								
CoT	13.3	42.5	50.8	77.8	22.5	27.8	42.7	41.3
GRPO	16.7	52.5	75.2	86.5	29.4	36.9	44.8	48.9
GRPO (+Ours)	20.0	65.0	76.8	88.9	33.1	39.7	46.3	52.8$\uparrow^{3.9}$
PPO	26.7	52.5	71.0	80.9	34.2	34.1	41.1	48.6
PPO (+Ours)	30.0	57.5	72.6	85.2	37.9	37.8	41.7	51.8$\uparrow^{3.2}$
RLOO	30.0	55.0	73.8	82.7	35.5	36.0	39.8	50.4
RLOO (+Ours)	33.3	67.5	74.8	84.6	36.0	35.6	41.1	53.3$\uparrow^{2.9}$
Reinforce++	16.7	52.5	72.4	85.6	37.1	37.2	40.3	48.8
Reinforce++ (+Ours)	20.0	55.0	72.0	86.8	40.1	37.2	42.5	50.5 $\uparrow^{1.7}$
Qwen2.5-1.5B-Math								
CoT	10.0	42.5	59.0	74.6	24.3	27.6	39.5	39.6
GRPO	13.3	52.5	71.2	83.2	26.8	30.1	41.2	45.6
GRPO (+Ours)	23.3	62.5	71.8	83.9	32.0	32.9	41.7	49.6$\uparrow^{4.0}$
PPO	13.3	50.0	66.6	74.7	24.6	27.1	37.8	42.0
PPO (+Ours)	13.3	57.5	70.2	78.4	25.4	33.0	40.0	45.4$\uparrow^{3.4}$
RLOO	20.0	50.0	68.0	82.6	28.7	32.0	41.4	46.1
RLOO (+Ours)	23.3	57.5	71.6	83.1	33.8	33.2	41.8	49.2$\uparrow^{3.1}$
Reinforce++	10.0	47.5	70.0	83.2	32.0	31.7	41.3	45.1
Reinforce++ (+Ours)	20.0	50.0	70.8	83.7	34.2	31.6	42.0	47.5$\uparrow^{2.4}$

Table 1: We report results of different LLMs across seven mainstream benchmarks, with all main experiments conducted on Qwen2.5-Math-7B and Qwen2.5-Math-1.5B. Each baseline (GRPO, PPO, Reinforce++, and RLOO) is further evaluated with our CAPO mechanism, denoted as GRPO(+Ours), PPO(+Ours), and RLOO(+Ours), Reinforce++(+Ours), respectively. For clarity, all improvements of the CAPO variants over their corresponding baselines are highlighted in **bold**; identical results remain unmarked.

Models	GUI-Act-Web			OmniAct-Web			AndroidControl-Low			AndroidControl-High			Overall
	Type	GR	SR	Type	GR	SR	Type	GR	SR	Type	GR	SR	
Os-Atlas-4B	79.22	58.57	42.62	46.74	49.24	22.99	64.58	71.19	40.62	49.01	49.51	22.77	49.75
QwenVL2.5-3B	56.10	64.28	55.61	50.63	46.89	47.02	62.03	74.07	59.32	47.81	46.51	38.90	54.09
GRPO	85.10	82.36	70.23	79.02	71.10	70.76	82.13	80.15	63.87	60.10	58.25	46.81	70.79
Ours	87.73	85.85	85.85	87.24	74.02	74.16	82.29	81.19	61.41	65.91	61.47	47.71	74.60
$\Delta(\uparrow)$	\uparrow 2.63	\uparrow 3.49	\uparrow 15.62	\uparrow 8.22	\uparrow 2.92	\uparrow 3.40	\uparrow 1.16	\uparrow 1.04	\downarrow 2.46	\uparrow 5.81	\uparrow 3.22	\uparrow 0.90	\uparrow 3.81

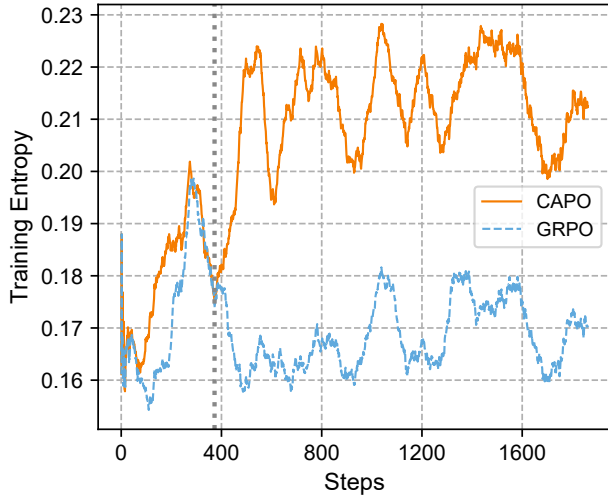
Table 2: Performance comparison of GUI reasoning tasks across both low-level and high-level settings on GUI-Act-Web, OmniAct-Web, AndroidControl-Low, and AndroidControl-High. We report Type (action type prediction), GR (grounding accuracy), and SR (step success rate) under a unified zero-shot prompt for fair comparison. CAPO consistently improves over the GRPO baseline, yielding an average gain of +3.81 points across planning benchmarks, further demonstrating its effectiveness in multimodal reasoning scenarios.

Following the setup in (Luo et al. 2025), which employs GRPO, we evaluate performance across perception, low-level, and high-level planning tasks. As shown in Table 2, CAPO yields an overall gain of +3.81 on planning tasks. For completeness, we also report its GUI grounding results in perception tasks, detailed in Appendix C.

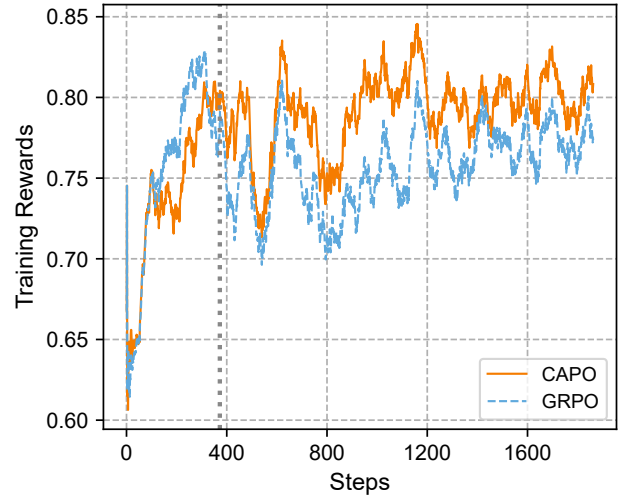
These results highlight CAPO’s strong generalization beyond mathematical reasoning, boosting multimodal perception and control by leveraging advantage as a curriculum signal that enables consistent generalization across diverse modalities.

Detailed Analysis

Training Dynamics Analysis. In Figure 3, we present the training reward and entropy dynamics of the 7B model under CAPO and GRPO. Before the phase transition, both methods exhibit comparable reward growth; however, CAPO becomes consistently superior once imitation has stabilized. The gray vertical line marks the transition from the imitation phase to the discrimination phase. After this point, the entropy trajectory of CAPO exhibits a *steady climb*, in contrast to the plateau observed in GRPO, while rewards continue to improve. This indicates that CAPO not only secures



(a) Entropy dynamics



(b) Reward dynamics

Figure 3: Comparison of reward and entropy curves between GRPO and CAPO on the 7B model. The gray vertical line marks the switch from imitation to discrimination. CAPO first relies on positive-only training to establish robust foundations. After the switch, negative samples lead to a steady increase in entropy and rewards, demonstrating enhanced generalization.

Method	AIME24 \uparrow	AMC \uparrow	MATH500 \uparrow	GSM8K \uparrow	Minerva \uparrow	Olympiad \uparrow	Avg. \uparrow
CoT	13.3	42.5	50.8	77.8	22.5	27.8	39.1
ADARFT [20]	15.8	55.0	74.4	91.0	25.4	24.9	47.8
GRPO	16.7	52.5	75.2	86.5	29.4	36.9	49.5
GRPO(+SC)	16.7	65.0	75.0	86.3	29.8	38.1	51.8
GRPO(+Ours)	20.0	65.0	76.8	88.9	33.1	39.7	53.9

Table 3: Performance comparison on Qwen2.5-7B-Math across six reasoning benchmarks. We compare five settings: CoT, ADARFT, vanilla GRPO, GRPO with a static curriculum (GRPO(+SC)), and our advantage curriculum method GRPO(+Ours). Our approach achieves the strongest overall performance among all variants.

stronger reward gains in the later stage but also maintains higher entropy, a property often linked to more diverse and exploratory reasoning paths. Moreover, the smooth rise in entropy suggests that CAPO avoids the sharp entropy collapses typically triggered by prematurely mixing negative samples. By deferring their incorporation, CAPO stabilizes the imitation stage and later exploits negative feedback more effectively, enabling better generalization across tasks.

Effect of Switch Stage in CAPO. We analyze how the switch stage influences the effectiveness of CAPO. Figure 4 reports results on two representative benchmarks (AIME24 and AMC23). Performance peaks when the switch occurs around 20%–30% of training, suggesting that after a short period of imitation-style learning, introducing negative advantages early enough encourages more discriminative reasoning and leads to more robust learning dynamics. Complete benchmark results for both are presented in Appendix B.1 for reference.

Comparing Static and Dynamic Curriculum Strategies. We compare the conventional static curriculum method with

our dynamic advantage-based training strategy. In the static curriculum setting, we estimate sample difficulty by performing pass@16 evaluation for each sample. The dataset is then sorted based on this metric, and the model is trained following this fixed order. In contrast, our dynamic approach focuses on adjusting the advantage signal progressively during training, without the need for manually reordering the data. As shown in Table 3, both ADARFT and the static curriculum GRPO(+SC) offer only limited and inconsistent gains over vanilla GRPO. In contrast, our dynamic method GRPO(+Ours) achieves the strongest improvements across most tasks. This suggests that predefined difficulty heuristics are insufficient, and dynamically adjusting the advantage signal provides a more effective curriculum.

Further Discussion on Generalization. Recently some work has underscored the vulnerability of LLMs to distributional shifts (Yuan et al. 2023; Wang et al. 2024), with models often exhibiting strong in-distribution (ID) performance but significant degradation on out-of-distribution (OOD) domains (Berglund et al. 2024; Yang et al. 2024). To evaluate

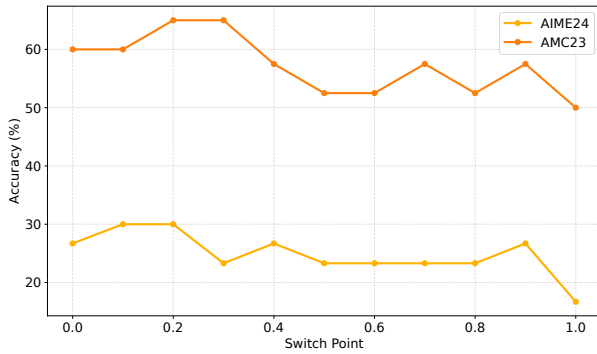


Figure 4: Switch Point Sensitivity: Results on AIME24 and AMC23. While we evaluate CAPO across 8 benchmarks in total, here we present the two most representative results. The curves show that introducing the switch point in the early phase (around 0.2–0.3) yields the best performance, aligning with the theoretical expectation that early positive-only training stabilizes learning before timely inclusion of negative signals enhances generalization.

CAPO’s generalization under distributional shift, we benchmark it on two representative reasoning datasets: ARC-C and GPQA-Diamond. Since all models are trained exclusively on mathematical data, this setting naturally provides a robust OOD evaluation. As shown in Figure 5, CAPO achieves an average accuracy of 52.8, outperforming GRPO by +3.8, with gains observed consistently across all benchmarks (+1.4 on ARC-C and +6.2 on GPQA-D). These results demonstrate the effectiveness of CAPO’s progressive learning strategy in enhancing OOD generalization by integrating both imitation and discrimination phases. These results confirm CAPO’s robustness under OOD conditions.

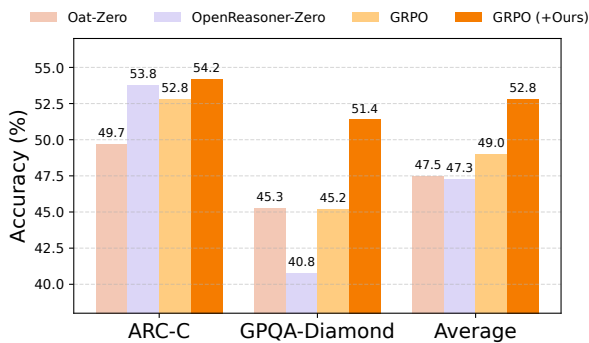


Figure 5: Results on two representative out-of-distribution benchmarks (Qwen2.5-Math-7B-Base). CAPO achieves an average accuracy of 52.8, outperforming GRPO by +6.5%, demonstrating improved robustness under distributional shifts.

Related Work

Reinforcement Learning for Large Reasoning Model. Recent advancements in both LLMs and MLLMs have in-

creasingly focused on enabling models to simulate reasoning processes. Inspired by powerful reasoning models like DeepSeek-R1 (Guo et al. 2025), and Kimi-k1.5 (Team et al. 2025), research focus has been drawn to reinforcement learning with verifiable rewards (RLVR) (Wei et al. 2022; Wang et al. 2023; Wu et al. 2024a; Hua et al. 2025; Wu et al. 2025), which combines both positive and negative feedback. Recent works have begun to explore the sperate function of themZhu et al. (2025) leverage them primarily to balance diversity, yet fall short of fully unlocking their optimization potential. Xu et al. (2025b) incorporate both positive and negative samples into a DPO-inspired loss, but still rely on a fine-tuning-centric, two-stage which lacks a principled mechanism to integrate with advantage-based reinforcement learning methods. In contrast, our method first leverages positive advantage samples to establish stable behavioral priors, then introduces negative samples to improve generalization. This staged scheduling avoids mixing signals prematurely and aligns naturally with advantage-based reinforcement learning frameworks.

Curriculum Learning. Curriculum learning (CL), originally proposed by Bengio et al. (2009b), has long adhered to a data-centric paradigm, and has been widely adopted in LLM development (Parashar et al. 2025; Xu et al. 2025a). Recent methods like Speed-RL (Zhang et al. 2025) and Kim and Lee (2024) still employ data-centric strategies that progress from simple to complex tasks using external criteria such as task difficulty to determine sample ordering. Similarly, LBS3 (Luo et al. 2024) guides models through progressive training with easy-to-hard proxy queries. However, these approaches fundamentally misalign with effective curriculum design by relying on static, externally defined difficulty metrics rather than the model’s evolving capabilities. Truly effective curriculum learning should be competence-aware, dynamically adapting to the model’s competence.

CAPO leverages advantage estimates as an intrinsic, competence-aware signal, avoiding the early mixing of positive and negative feedback that often destabilizes RL training. It employs a staged curriculum: positive advantages first establish stable behavioral priors, and negative ones are later introduced to enhance generalization. Unlike curriculum learning methods that depend on static, externally defined difficulty measures, our framework adapts dynamically to the policy’s evolving competence, thereby unifying the stability of early reinforcement learning with the adaptability of curriculum scheduling.

Conclusion

We propose CAPO (Curriculum Advantage Policy Optimization), which treats advantage as an intrinsic learning signal to construct adaptive curricula for reasoning model training. CAPO adopts a two-phase strategy that progresses from positive-only imitation learning to discrimination learning with negative signals, stabilizing optimization and mitigating premature mixed-signal interference. Extensive experiments show that CAPO consistently outperforms strong baselines such as GRPO across benchmarks and model scales, exhibiting cross-domain generalization.

References

- Ahmadian, A.; Cremer, C.; Gallé, M.; Fadaee, M.; Kreutzer, J.; Pietquin, O.; Üstün, A.; and Hooker, S. 2024. Back to Basics: Revisiting REINFORCE Style Optimization for Learning from Human Feedback in LLMs. *arXiv:2402.14740*.
- Bandura, A.; Ross, D.; and Ross, S. A. 1961. Transmission of aggression through imitation of aggressive models. *The Journal of Abnormal and Social Psychology*, 63(3): 575.
- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009a. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.
- Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009b. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*, 41–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.
- Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A. C.; Korbak, T.; and Evans, O. 2024. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”. In *The Twelfth International Conference on Learning Representations*.
- Chen, X.; Lu, J.; Kim, M.; Zhang, D.; Tang, J.; Piché, A.; Gontier, N.; Bengio, Y.; and Kamaloo, E. 2025. Self-Evolving Curriculum for LLM Reasoning. *arXiv:2505.14970*.
- Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457v1*.
- Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- He, C.; Luo, R.; Bai, Y.; Hu, S.; Thai, Z.; Shen, J.; Hu, J.; Han, X.; Huang, Y.; Zhang, Y.; et al. 2024. Olympiad-Bench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3828–3850.
- Hendrycks, D.; Burns, C.; Kadavath, S.; Arora, A.; Basart, S.; Tang, E.; Song, D.; and Steinhardt, J. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Hu, J.; Liu, J. K.; Xu, H.; and Shen, W. 2025. REINFORCE++: An Efficient RLHF Algorithm with Robustness to Both Prompt and Reward Models. *arXiv:2501.03262*.
- Hua, C.; Xu, Q.; Yang, Z.; Wang, Z.; Bao, S.; and Huang, Q. 2025. OpenworldAUC: Towards Unified Evaluation and Optimization for Open-world Prompt Tuning. *CoRR*, abs/2505.05180.
- Kapoor, R.; Butala, Y. P.; Russak, M.; Koh, J. Y.; Kamble, K.; Alshikh, W.; and Salakhutdinov, R. 2024. OmniACT: A Dataset and Benchmark for Enabling Multimodal Generalist Autonomous Agents for Desktop and Web. *arXiv:2402.17553*.
- Kim, J.; and Lee, J. 2024. Strategic Data Ordering: Enhancing Large Language Model Performance through Curriculum Learning. *arXiv preprint arXiv:2405.07490*.
- Lewkowycz, A.; Andreassen, A.; Dohan, D.; Dyer, E.; Michalewski, H.; Ramasesh, V.; Slone, A.; Anil, C.; Schlag, I.; Gutman-Solo, T.; et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35: 3843–3857.
- Li, J.; Beeching, E.; Tunstall, L.; Lipkin, B.; Soletskyi, R.; Huang, S.; Rasul, K.; Yu, L.; Jiang, A. Q.; Shen, Z.; et al. 2024a. Numinamath: The largest public dataset in AI4Maths with 860k pairs of competition math problems and solutions. <https://huggingface.co/datasets/Numinamath>. Hugging Face repository, 13:9.
- Li, K.; Meng, Z.; Lin, H.; Luo, Z.; Tian, Y.; Ma, J.; Huang, Z.; and Chua, T.-S. 2025. ScreenSpot-Pro: GUI Grounding for Professional High-Resolution Computer Use. *arXiv:2504.07981*.
- Li, W.; Bishop, W.; Li, A.; Rawles, C.; Campbell-Ajala, F.; Tyamagundlu, D.; and Riva, O. 2024b. On the Effects of Data Scale on UI Control Agents. *arXiv:2406.03679*.
- Li, X.; Xu, H.; Zhang, J.; and Chang, H.-h. 2023. Deep reinforcement learning for adaptive learning systems. *Journal of Educational and Behavioral Statistics*, 48(2): 220–243.
- Liu, Z.; Chen, C.; Li, W.; Qi, P.; Pang, T.; Du, C.; Lee, W. S.; and Lin, M. 2025. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Lu, Q.; Shao, W.; Liu, Z.; Meng, F.; Li, B.; Chen, B.; Huang, S.; Zhang, K.; Qiao, Y.; and Luo, P. 2024. GUI Odyssey: A Comprehensive Dataset for Cross-App GUI Navigation on Mobile Devices. *arXiv:2406.08451*.
- Luo, K.; Ding, Z.; Weng, Z.; Qiao, L.; Zhao, M.; Li, X.; Yin, D.; and Shu, J. 2024. Let’s Be Self-generated via Step by Step: A Curriculum Learning Approach to Automated Reasoning with Large Language Models. *arXiv preprint arXiv:2410.21728*.
- Luo, R.; Wang, L.; He, W.; and Xia, X. 2025. GUI-R1 : A Generalist R1-Style Vision-Language Action Model For GUI Agents. *arXiv:2504.10458*.
- Parashar, S.; Gui, S.; Li, X.; Ling, H.; Vemuri, S.; Olson, B.; Li, E.; Zhang, Y.; Caverlee, J.; Kalathil, D.; and Ji, S. 2025. Curriculum Reinforcement Learning from Easy to Hard Tasks Improves LLM Reasoning. *arXiv preprint arXiv:2506.06632*.
- Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*.
- Schulman, J.; Moritz, P.; Levine, S.; Jordan, M.; and Abbeel, P. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.

- Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv:1707.06347*.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024a. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024b. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300*.
- Sheng, G.; Zhang, C.; Ye, Z.; Wu, X.; Zhang, W.; Zhang, R.; Peng, Y.; Lin, H.; and Wu, C. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems, EuroSys '25*, 1279–1297. ACM.
- Shi, T.; Wu, Y.; Song, L.; Zhou, T.; and Zhao, J. 2025. Efficient Reinforcement Finetuning via Adaptive Curriculum Learning. *arXiv:2504.05520*.
- Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.
- Tang, Z.; Zhang, X.; Wang, B.; and Wei, F. 2024. Mathscales: Scaling instruction tuning for mathematical reasoning. *arXiv preprint arXiv:2403.02884*.
- Team, K.; Du, A.; Gao, B.; Xing, B.; Jiang, C.; Chen, C.; Li, C.; Xiao, C.; Du, C.; Liao, C.; et al. 2025. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.
- Team, Q. 2025. Qwen2.5-VL.
- Wang, J.; Hu, X.; Hou, W.; Chen, H.; Zheng, R.; Wang, Y.; Yang, L.; Ye, W.; Huang, H.; Geng, X.; Jiao, B.; Zhang, Y.; and Xie, X. 2024. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. *IEEE Data Eng. Bull.*, 47(1): 48–62.
- Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.
- Whitehurst, G. J. 1969. Discrimination learning in children as a function of reinforcement condition, task complexity, and chronological age. *Journal of Experimental Child Psychology*, 7(2): 314–325.
- Wu, J.; Feng, M.; Zhang, S.; Che, F.; Wen, Z.; and Tao, J. 2024a. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts. *arXiv preprint arXiv:2411.18478*.
- Wu, J.; Liao, C.; Feng, M.; Zhang, S.; Wen, Z.; Shao, P.; Xu, H.; and Tao, J. 2025. Thought-Augmented Policy Optimization: Bridging External Guidance and Internal Capabilities. *arXiv preprint arXiv:2505.15692*.
- Wu, Z.; Wu, Z.; Xu, F.; Wang, Y.; Sun, Q.; Jia, C.; Cheng, K.; Ding, Z.; Chen, L.; Liang, P. P.; and Qiao, Y. 2024b. OS-ATLAS: A Foundation Action Model for Generalist GUI Agents. *arXiv:2410.23218*.
- Xu, F.; Hao, Q.; Zong, Z.; Wang, J.; Zhang, Y.; Wang, J.; Lan, X.; Gong, J.; Ouyang, T.; Meng, F.; et al. 2025a. Towards Large Reasoning Models: A Survey of Reinforced Reasoning with Large Language Models. *arXiv preprint arXiv:2501.09686*.
- Xu, S.; Peng, C.; Long, J.; Xu, W.; Chu, W.; and Qi, Y. 2025b. Harnessing Negative Signals: Reinforcement Distillation from Teacher Data for LLM Reasoning. *arXiv preprint arXiv:2505.24850*.
- Yang, H.; Zhang, Y.; Xu, J.; Lu, H.; Heng, P.-A.; and Lam, W. 2024. Unveiling the Generalization Power of Fine-Tuned Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 884–899. Mexico City, Mexico: Association for Computational Linguistics.
- Yuan, L.; Chen, Y.; Cui, G.; Gao, H.; Zou, F.; Cheng, X.; Ji, H.; Liu, Z.; and Sun, M. 2023. Revisiting Out-of-distribution Robustness in NLP: Benchmarks, Analysis, and LLMs Evaluations. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zeng, W.; Huang, Y.; Liu, Q.; Liu, W.; He, K.; Ma, Z.; and He, J. 2025. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *arXiv preprint arXiv:2503.18892*.
- Zhang, R.; Arora, D.; Mei, S.; and Zanette, A. 2025. SPEED-RL: Faster Training of Reasoning Models via Online Curriculum Learning. *arXiv preprint arXiv:2506.09016*.
- Zhu, X.; Xia, M.; Wei, Z.; Chen, W.-L.; Chen, D.; and Meng, Y. 2025. The Surprising Effectiveness of Negative Reinforcement in LLM Reasoning. *arXiv preprint arXiv:2506.01347*.