CDE: Curiosity-Driven Exploration for Efficient Reinforcement Learning in Large Language Models

Runpeng Dai[1,3†], **Linfeng Song**[1†], **Haolin Liu** [1,4], **Zhenwen Liang**[1], **Dian Yu**[1], **Haitao Mi**[1], **Zhaopeng Tu**[2], **Rui Liu**[1,5], **Tong Zheng**[1,5], **Hongtu Zhu**[3], **Dong Yu**[1]

[1]Tencent AI Lab, [2]Tencent Multimodal Department,
[3]University of North Carolina at Chapel Hill,
[4]University of Virginia,
[5]University of Maryland, College Park
† Core contributors
runpeng@unc.edu, lfsong@global.tencent.com

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) is a powerful paradigm for enhancing the reasoning ability of Large Language Models (LLMs). Yet current RLVR methods often explore poorly, leading to premature convergence and entropy collapse. To address this challenge, we introduce **Curiosity-Driven Exploration (CDE)**, a framework that leverages the model's own intrinsic sense of curiosity to guide exploration. We formalize curiosity with signals from both the actor and the critic: for the actor, we use perplexity over its generated response, and for the critic, we use the variance of value estimates from a multi-head architecture. Both signals serve as an exploration bonus within the RLVR framework. Our theoretical analysis shows that the actor-wise bonus inherently penalizes overconfident errors and promotes diversity among correct responses; moreover, we connect the critic-wise bonus to the well-established count-based exploration bonus in RL. Empirically, our method achieves an approximate **+3** point over standard RLVR using GRPO/PPO on AIME benchmarks.

1 Introduction

RLVR is a central technique for advancing the reasoning capabilities of LLMs, demonstrating significant performance on complex reasoning tasks in mathematics and coding (Guo et al., 2025; Lu et al., 2025). Despite the emergence of various training algorithms, such as GRPO (Guo et al., 2024), DAPO (Yu et al., 2025b) and others (Wang et al., 2025; Liu et al., 2025), key issues remain. In particular, problems such as **premature convergence** and phenomena like **entropy collapse** (Cui et al., 2025; Zhuang et al., 2025) have been widely observed during training, posing fundamental challenges to the efficiency of RLVR.

These challenges stem from the classic exploration-exploitation dilemma in reinforcement learning (Sutton & Barto, 2018). Phenomena like entropy collapse reveal a critical flaw in the training process: it is heavily biased towards exploitation, causing models to converge prematurely instead of sufficiently exploring their environment for better solutions. Although the RL literature encompasses a wide range of exploration strategies, these methods exhibit limitations when applied to LLMs. Simple heuristics, including ϵ-greedy policies (Sutton & Barto, 2018) and entropy bonuses (Haarnoja et al., 2018), either inject randomness to the environment or encourage the policy to be more stochastic. Directly applying those approaches often demonstrates debatable effectiveness in complex

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MATH-AI.

environments like Deep RL (Andrychowicz et al., 2021) and LLM-based reasoning (Cui et al., 2025; Shen, 2025).

More principled strategies include count-based and prediction-based approaches. The former, including UCB (Lai, 1987) and LinUCB (Li et al., 2010), incentivizes visits to rarely explored states, while the latter, like ICM (Pathak et al., 2017) and RND (Burda et al., 2018b), reward an agent for reaching hard-to-predict states. Recent work has focused on adapting these methods for tuning LLMs (Bai et al., 2025; Sun et al., 2025; Gao et al., 2025; Yu et al., 2025a; Zhou et al., 2025; Zhuang et al., 2025). However, these strategies often require training auxiliary modules and effective state representations (Burda et al., 2018a). This requirement is particularly challenging for LLMs, where efficiently represent a reasoning path into a fixed-size embedding remains an open problem (Fu et al., 2024), and simplistic approaches such as using the last hidden state are often problematic (Barbero et al., 2024).

In this work, we propose an intuitive approach that leverages the model's intrinsic sense of **curiosity** as a guide for exploration. An LLM, having been trained on vast reasoning corpora, develops a sophisticated internal model of what constitutes a familiar versus a novel reasoning pattern. This parallels early childhood development (Chu & Schulz, 2020), where learning is not driven by a external summary and count of experiences, but is instead propelled by an intrinsic curiosity to explore novel situations. We formalize this principle in our **Curiosity-Driven Exploration (CDE)** framework, which considers curiosity signals from both the actor and the critic. For the actor, perplexity (PPL) over its generated response serves as the curiosity measure. For the critic, we measure curiosity via the variance of its posterior value distribution, which is approximated with a multi-head critic. The curiosity signals serve as an exploration bonus, shaping the reward and advantage functions to guide exploration.

Our theoretical analysis clarifies the calibration behavior of the PPL bonus and formally links critic curiosity to count-based exploration bonuses. Empirically, we observe consistent gains across four standard math-reasoning benchmarks—AIME'25, AIME'24, AMC'23, and MATH. Furthermore, our analysis reveals a phenomenon we term **calibration collapse**: under naive GRPO training, the model's confidence progressively decouples from its correctness, while adding PPL bonus mitigates this miscalibration.

2 Our Approach

In the following sections, we introduce the formulations of exploration Guided by actor and critic curiosity (Section 2.1 and 2.2), and defer the background on GRPO and PPO to Appendix A.

2.1 Exploration Guided by Actor Curiosity

We model actor curiosity as the actor's uncertainty about its own actions. Intuitively, a response that is surprising to the actor—i.e., has a low probability under its current policy—likely resides in an underexplored region of its learned distribution. A natural and computationally efficient measure of this surprise is the perplexity of the actor's generation. Given an actor π, a prompt q and a generated response $o = \{o_1, \ldots, o_T\}$, the perplexity is is defined as: $B_{actor}(q, o) = -\frac{1}{T}\sum_{t=1}^{T} \log \pi(o_t|o_{<t}, q)$. A higher value indicates greater surprise and thus a stronger intrinsic signal for exploration.

However, practically simply adding this bonus to the original reward can be unstable and sub-optimal. Unconstrained exploration might incentivize the model to generate high-perplexity but inaccurate responses, or lead to over-exploration where the policy fails to converge to a stable output. To address this, we integrate the bonus using an adaptive clipping mechanism. The total sentence-level reward, $\tilde{r}$, is a combination of the original reward signal $r(q, o)$ and the curiosity bonus $B_{actor}(q, o)$:

$$\tilde{r}(q, o) = r(q, o) + \omega \min(|r(q, o)|/\kappa, \alpha B_{actor}(q, o)),$$

Here, the bonus weight ω is annealed downward over training, enabling aggressive exploration early on and gradually shifting toward exploitation as the policy converges. The clipping ratio κ and the bonus scaling factor α together control the bonus magnitude: by capping it at $|r(q, o)|/\kappa$, the auxiliary term remains a supplement to $r(q, o)$ rather than overwhelming the learning signal.

Intuitions and Theoretical Foundation We analyze responses along two axes—*correctness* and *actor PPL*. Among these four categories, two require particular attention:

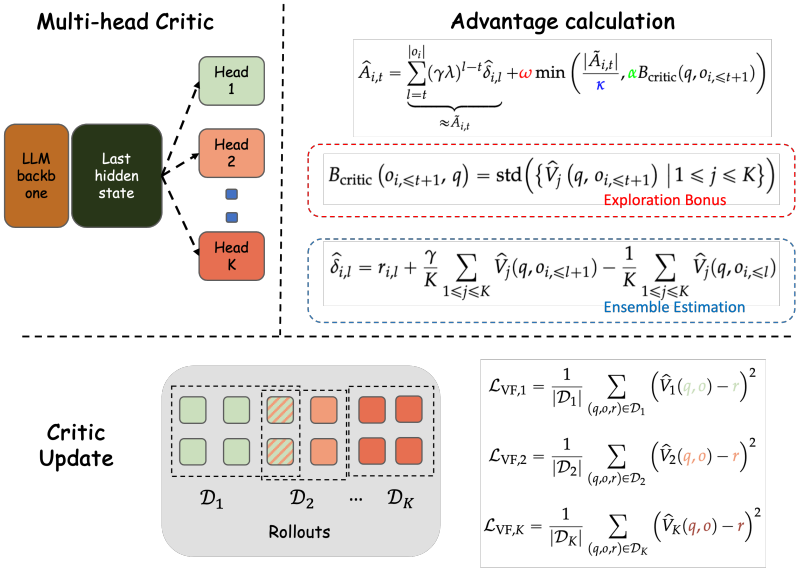


Figure 2: Illustration of the multi-head critic framework.

1. Incorrect responses with low PPL indicate that the model is highly confident in its answer, yet the response is wrong. This reflects overfitting and should be penalized.
2. Correct responses with high PPL suggest that the model is less familiar with such answers, but they turn out to be successful. This reflects effective exploration and should be encouraged.

As illustrated in Figure 1, we find out that the PPL bonus intrinsically penalizes confident mistakes while encouraging novel correct responses. For correct responses, those novel responses (with higher PPL) receive a larger positive reward. For incorrect responses, those confident responses (with lower PPL) receive larger penalty as it receives smaller PPL bonus. Theorem B.1 formalizes this intuition; its statement and proof are deferred to Appendix B.1.

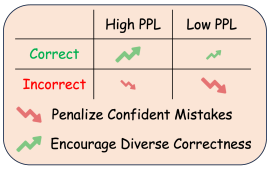


Figure 1: Responses by correctness and avg PPL.

2.2 Exploration Guided by Critic Curiosity

In contrast to critic-free methods such as REINFORCE and GRPO, the critic in actor–critic frameworks provides a higher-level understanding of the prompt–response pair. Its posterior distribution naturally reflects the degree of coverage: regions with dense data yield low-variance posteriors, whereas sparsely sampled regions result in larger-variance. Posterior distributions are a well-established means of quantifying predictive uncertainty in deep learning models (Gal & Ghahramani, 2016).

To approximate the posterior distribution of value estimates, we adopt the classical bootstrap method (Davison & Hinkley, 1997), widely used in statistics and increasingly recognized in the RL community as an effective tool for exploration (Osband et al., 2016; Ciosek et al., 2019; Bai et al., 2021). We implement this idea through a multi-head critic (upper-left subfigure in Figure 2), where K critics share a common LLM backbone. Each head is trained on a resampled subset of the collected trajectories (bottom subfigure in Figure 2), thereby producing an empirical approximation to the posterior distribution. We then use the standard deviation across the K heads as a principled curiosity signal, guiding the policy toward regions of high disagreement. (upper-right subfigure in Figure 2). The full algorithm is leave in Appendix C.

Beyond the practical design, we provide a theoretical link: under standard linear-MDP assumptions, the cross-head standard deviation of bootstrap critics consistently estimates the classical pseudo-count bonus. Assumptions and the theorem appear in Appendix B.2.

3 Experiments

In this paper, we adopt DAPO-17K (Yu et al., 2025b) for training and evaluate the performance of CDE on four challenging mathematical reasoning benchmarks: MATH (Hendrycks et al., 2021), AMC23 (MAA, b), AIME24, and AIME25 (MAA, a). All experiments are implemented with the

Model	MATH	AMC23		AIME24		AIME25		Avg
	Avg@1	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	
Qwen3-4B-Base	23.1	10.9	53.8	1.5	8.4	1.3	8.3	9.2
GRPO based methods								
Qwen3-4B-Base-GRPO	87.3	63.6	89.1	20.8	41.9	21.0	39.2	48.2
↳ **w/ PPL bonus**	87.7	67.8	89.5	23.3	48.5	23.5	42.5	50.6
PPO based methods								
Qwen3-4B-Base-PPO	86.6	64.1	87.2	17.8	36.0	17.5	33.7	46.5
↳ **w/ PPL bonus**	87.9	66.1	88.5	18.3	37.6	18.3	33.5	47.7
↳ **w/ 2 Heads**	83.2	63.6	89.9	19.6	34.8	19.6	36.1	46.6
↳ **w/ 4 Heads**	87.3	63.9	87.9	21.5	35.5	21.5	45.5	48.5
↳ **w/ 8 Heads**	85.1	66.7	86.9	21.7	46.4	19.0	37.1	48.1
↳ **w/ 16 Heads**	88.3	65.0	88.7	20.5	41.9	20.0	38.8	48.6

Table 1: Zero-shot accuracy of different models on the validation datasets. Avg@16 denotes the mean Pass@1 accuracy over 16 sampled generations, while **Avg** column represents the overall average across datasets, computed as Avg@1 for MATH and Avg@16 for the remaining datasets.

Verl framework using the Qwen3-4B-Base model (Yang et al., 2025). The implementation details are provided in Appendix D, while the full formulations of GRPO and PPO can be found in Appendix A.

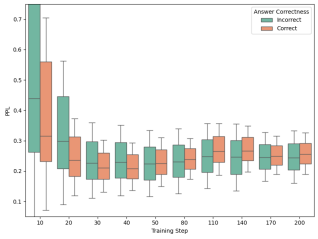
Main Results The main results are presented in Table 1. Here K Heads represents multi-head critic PPO with K head critics. The key observations are as follows:

- Both the PPL bonus and the multi-head critic ($K \geqslant 4$) improve the baselines' reasoning ability, yielding an average gain of $+2.4$ points across datasets and showing consistent superiority. In many cases, we observe over $+6$ points in Pass@16 on AIME benchmarks.
- The performance of multi-head PPO generally increases with the number of heads K and performance increase begin to plateau once $K \geqslant 4$, which suggests that a modest number of heads can capture most of the curiosity signals needed.

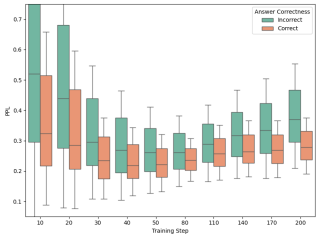
We also performed a series of ablation studies, the results of which are provided in Appendix E.

Analysis of Calibration As shown in Figure 3, we plot the batch-wise mean response perplexity (PPL), stratified by answer correctness. In subfigure (a), we observe a phenomenon we term **calibration collapse**: early in naive GRPO training, correct responses have lower PPL (higher confidence) than incorrect ones, but as training progresses this gap shrinks and ultimately vanishes—confidence no longer tracks correctness. By contrast, with a PPL bonus (subfigure (b)), this separation is sustained throughout training.

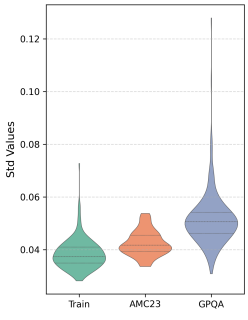
This pattern is explained by Theorem B.1: while both naive GRPO and GRPO with a PPL bonus tend to increase confidence on correct answers, the PPL bonus additionally suppresses confident errors (low-PPL incorrect trajectories), thereby improving calibration.



(a) GRPO without PPL bonus



(b) GRPO with PPL bonus



(c) Distribution of Critic Heads std.

Figure 3: (a)–(b) Mean response PPL over training steps, stratified by correctness. (c) Distribution of the standard deviation across $K = 16$ value heads for each dataset.

Analysis of Multi-Head Curiosity Signal In subfigure (c) of Figure 3, we present a cross-dataset analysis by calculating the average standard deviation of the value estimates across different questions. Specifically, we evaluate three datasets: the training set (DAPO-17K), the in-domain validation set (AMC23), and the out-of-domain validation set GPQA (Rein et al., 2023). We observe that the training set exhibits a smaller standard deviation compared to both the in-domain and out-of-domain

validation sets. This pattern aligns with the intuition that multi-head critics tend to show stronger disagreement on data that is less frequently encountered during training.

4 Conclusion

We have presented Curiosity-Driven Exploration, an efficient technique that enhances exploration by incorporating curiosity signals from both the actor and the critic. Its effectiveness is demonstrated by consistent accuracy improvements over strong baselines on a suite of challenging mathematical reasoning benchmarks, with these empirical results strongly corroborating our underlying theoretical framework and intuition. The **calibration collapse** revealed in our analysis aligns with recent findings on the root causes of LLM hallucination (Kalai et al., 2025), pointing to a promising avenue for future work.

References

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Leonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters for on-policy deep actor-critic methods? a large-scale study. In *International conference on learning representations*, 2021.

Chenjia Bai, Lingxiao Wang, Lei Han, Jianye Hao, Animesh Garg, Peng Liu, and Zhaoran Wang. Principled exploration via optimistic bootstrapping and backward induction. In *International Conference on Machine Learning*, pp. 577–587. PMLR, 2021.

Chenjia Bai, Yang Zhang, Shuang Qiu, Qiaosheng Zhang, Kang Xu, and Xuelong Li. Online preference alignment for language models via count-based exploration. *arXiv preprint arXiv:2501.12735*, 2025.

Federico Barbero, Andrea Banino, Steven Kapturowski, Dharshan Kumaran, João Madeira Araújo, Oleksandr Vitvitskyi, Razvan Pascanu, and Petar Veličković. Transformers need glasses! information over-squashing in language tasks. *Advances in Neural Information Processing Systems*, 37: 98111–98142, 2024.

Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018a.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018b.

Junyi Chu and Laura E Schulz. Play, curiosity, and cognition. *Annual Review of Developmental Psychology*, 2(1):317–343, 2020.

Kamil Ciosek, Quan Vuong, Robert Loftin, and Katja Hofmann. Better exploration with optimistic actor critic. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*. Number 1. Cambridge university press, 1997.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. *arXiv preprint arXiv:2412.11556*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Jingtong Gao, Ling Pan, Yejing Wang, Rui Zhong, Chi Lu, Qingpeng Cai, Peng Jiang, and Xiangyu Zhao. Navigate the unknown: Enhancing llm reasoning with intrinsic motivation guided exploration. *arXiv preprint arXiv:2505.17621*, 2025.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, et al. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*, 2024.

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

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.

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

Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL `https://openai.com/index/why-language-models-hallucinate/`.

Tze Leung Lai. Adaptive treatment allocation and the multi-armed bandit problem. *The annals of statistics*, pp. 1091–1114, 1987.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

Rui Liu, Dian Yu, Tong Zheng, Runpeng Dai, Zongxia Li, Wenhao Yu, Zhenwen Liang, Linfeng Song, Haitao Mi, Pratap Tokekar, et al. Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning. *arXiv preprint arXiv:2510.01444*, 2025.

Yuchen Lu, Run Yang, Yichen Zhang, Shuguang Yu, Runpeng Dai, Ziwei Wang, Jiayi Xiang, Siran Gao, Xinyao Ruan, Yirui Huang, et al. Stateval: A comprehensive benchmark for large language models in statistics. *arXiv preprint arXiv:2510.09517*, 2025.

MAA. American invitational mathematics examination (AIME). Mathematics Competition Series, n.d.a. URL `https://maa.org/math-competitions/aime`.

MAA. American mathematics competitions (AMC 10/12). Mathematics Competition Series, n.d.b. URL `https://maa.org/math-competitions/amc`.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

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. *arXiv preprint arXiv:2311.12022*, 2023.

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

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

Han Shen. On entropy control in llm-rl algorithms. *arXiv preprint arXiv:2509.03493*, 2025.

Haoran Sun, Yekun Chai, Shuohuan Wang, Yu Sun, Hua Wu, and Haifeng Wang. Curiosity-driven reinforcement learning from human feedback. *arXiv preprint arXiv:2501.11463*, 2025.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Xiangqi Wang, Yue Huang, Yujun Zhou, Xiaonan Luo, Kehan Guo, and Xiangliang Zhang. Causally-enhanced reinforcement policy optimization. *arXiv preprint arXiv:2509.23095*, 2025.

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

Fei Yu, Yingru Li, and Benyou Wang. Uncertainty-aware search and value models: Mitigating search scaling flaws in llms. *arXiv preprint arXiv:2502.11155*, 2025a.

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

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*, 2025.

Haomin Zhuang, Yujun Zhou, Taicheng Guo, Yue Huang, Fangxu Liu, Kai Song, and Xiangliang Zhang. Exploring multi-temperature strategies for token-and rollout-level control in rlvr. *arXiv preprint arXiv:2510.08892*, 2025.

A Background on RLVR Algorithms: GRPO and PPO

We formulate the language generation process of LLMs as a sequential decision-making problem (Yu et al., 2025b; Yue et al., 2025). Specifically, we consider two reinforcement learning algorithms: *Group Relative Policy Optimization* (GRPO), a critic-free method, and *Proximal Policy Optimization* (PPO), a canonical actor–critic method. We adopt the training paradigm of Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al., 2025; Lambert et al., 2024) and utilize a rule-based verifier to compare the generated response with the ground truth to judge its correctness.

A.1 Group Relative Policy Optimization (GRPO, Shao et al. 2024)

GRPO is an REINFORCE-style optimization algorithm. Let π_θ denote the LLM policy with parameters θ. At each training step, given a prompt q sampled from the dataset $\mathcal{D}$, the current policy π_θ generates a group of G candidate outputs $\{o_1, o_2, \ldots, o_G\}$. For each candidate o_i, we compute its total reward $r_i = r(o_i, q)$.

The advantage for each output is computed by normalizing its reward with respect to the group's rewards:

$$A_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G) + \delta},$$

where δ is a small constant for numerical stability. The same advantage A_i is applied to all tokens in o_i. Let $\pi_{\theta_{\text{old}}}$ be the policy from the previous step and π_{ref} the original pre-trained model. GRPO maximizes:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G} \sum_{i=1}^{G} \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \mathcal{L}_\theta(\tilde{r}_{i,t}, A_i) \right] - \beta D_{\text{KL}} \left(\pi_\theta \| \pi_{\text{ref}}\right),$$

where the clipped objective is

$$\mathcal{L}_\theta(\tilde{r}_{i,t}, A_i) = \min\left(\tilde{r}_{i,t} A_i, \text{clip}(\tilde{r}_{i,t}, 1-\varepsilon, 1+\varepsilon) A_i\right), \quad \tilde{r}_{i,t} = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}.$$

Here, ε and β control the ratio clipping threshold and the KL-penalty strength, respectively. The clipping mitigates large, unstable policy updates, while the KL term constrains deviation from π_{ref}.

A.2 Proximal Policy Optimization (PPO, Schulman et al. 2017)

PPO is an actor–critic algorithm that maintains both a policy (actor) π_θ and a value function (critic) V_ϕ with parameters ϕ, estimating the expected total reward from a given state (prompt and sequence prefix). The advantage function in PPO leverages the critic to reduce variance. Specifically, **Generalized Advantage Estimation (GAE)** is applied to compute token-level advantages. For an output o_i with sentence-level reward r_i, the GAE at token t is:

$$A_{i,t} = \sum_{l=t}^{|o_i|} (\gamma\lambda)^{l-t} \delta_{i,l},$$

where

$$\delta_{i,l} = r_{i,l} + \gamma V_\phi(q, o_{i,\leqslant l+1}) - V_\phi(q, o_{i,\leqslant l}),$$

and in our setting $r_{i,l} = 0$ for all non-terminal tokens, with $r_{i,|o_i|} = r_i$. The hyperparameters γ and λ are the discount factor and GAE trace-decay, respectively. The PPO objective is:

$$\mathcal{L}_{\text{PPO}}(\theta, \phi) = \mathbb{E}_{q\sim\mathcal{D},\{o_i\}\sim\pi_{\theta_{\text{old}}}} \Big[\frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \left[\mathcal{L}_\theta(\tilde{r}_{i,t}, A_{i,t}) - c_1 \mathcal{L}_\phi(q, o_{i,<t}, r_i) \right] \Big] - \beta D_{\text{KL}} \left(\pi_\theta \| \pi_{\text{ref}}\right),$$

where $\mathcal{L}_\theta$ is as in GRPO but with per-token $A_{i,t}$, and the value loss is:

$$\mathcal{L}_\phi(\phi) = \left(V_\phi(q, o_{i,<t}) - r_i\right)^2.$$

In practice, we alternate optimization of the actor (θ) and the critic (ϕ).

B Theoretically Results

B.1 Calibration Effect of PPL Bonus

Theorem B.1. *Let π_t denote the policy at training step t. With PPL bonus in Equation* (**??**)*, the update to π_{t+1} calibrates the policy's confidence as follows:*

(i) Among correct responses, trajectories with higher perplexity receive a larger relative probability increase.

(ii) Among incorrect responses, trajectories with lower perplexity receive a larger relative probability decrease.

Proof. Define $\tilde{r}_t(q,o) = r(q,o) + b_t(q,o)$ where $b_t(q,o) = \omega \min\{\kappa | r(q,o)|, -\frac{\alpha}{T_o}\log \pi_t(o|q)\}$ is a bonus function where T_o is the length of response o. Note that ω is a redundant variable in theory because we can write $b_t(q,o) = \min\{\kappa' | r(q,o)|, -\frac{\alpha}{T_o}\log \pi_t(o|q)\}$ with $\kappa' = \omega\kappa$ and $\alpha' = \omega\alpha$. Given that $r(x,y) \in \{1,-1\}$, it suffices to consider $b_t(q,o) = \min\{\kappa, -\frac{\alpha}{T_o}\log \pi_t(o|q)\}$. Thus, as long as we use $\kappa < 1$, we have $\text{sign}(\tilde{r}_t(q,o)) = \text{sign}(r(q,o))$. The introduce of bonus does not change the sign of the original correctness reward.

Consider single step policy optimization

$$\pi_{t+1}(\cdot|q) = \arg\max_{\pi} \left\{ \sum_o \pi(o|q)\tilde{r}_t(q,o) - \frac{1}{\eta}\text{KL}\left(\pi(\cdot|q) \| \pi_t(\cdot|q)\right) \right\},$$

which has closed-form solution

$$\pi_{t+1}(o|q) = \frac{\pi_t(o|q)\exp\left(\eta\tilde{r}_t(q,o)\right)}{\sum_{o'}\pi_t(o'|q)\exp\left(\eta\tilde{r}_t(q,o')\right)}.$$

For any question q and response o. Define $Z(q) = \sum_{o'}\pi_t(o'|q)\exp\left(\eta\tilde{r}_t(q,o')\right)$, we have

$$\log \pi_{t+1}(o|q) = \log \pi_t(o|q) + \eta\tilde{r}_t(q,o) - \log\left(Z(q)\right).$$

Define $\Delta_t(o|q) = \log \pi_{t+1}(o|q) - \log \pi_t(o|q)$ as the change of likelihood of response o under question q at update step t. For two correct response o_1^+ and o_2^+ with length $T_{o_1^+}$ and $T_{o_2^+}$, and $-\frac{\alpha}{T_{o_1^+}}\log \pi_t(o_1^+|q) \geqslant -\frac{\alpha}{T_{o_2^+}}\log \pi_t(o_2^+|q)$ (i.e. o_1^+ has larger perplexity), we have

$$\begin{aligned}
&\Delta_t(o_1^+|q) - \Delta_t(o_2^+|q) \\
&= \tilde{r}_t(q,o_1^+) - \tilde{r}_t(q,o_2^+) \\
&= b_t(q,o_1^+) - b_t(q,o_2^+) \\
&= \min\{\kappa, -\frac{\alpha}{T_{o_1^+}}\log \pi_t(o_1^+|q)\} - \min\{\kappa, -\frac{\alpha}{T_{o_2^+}}\log \pi_t(o_2^+|q)\} \\
&\geqslant 0
\end{aligned}$$

Similarly, for two incorrect response o_1^- and o_2^- with $-\frac{\alpha}{T_{o_1^-}}\log \pi_t(o_1^-|q) \geqslant -\frac{\alpha}{T_{o_2^-}}\log \pi_t(o_2^-|q)$ (i.e. o_1^- has larger perplexity), we have $\Delta_t(o_1^-|q) - \Delta_t(o_2^-|q) \geqslant 0$.

Specifically, given a question q, for any response (o_1, o_2) that has the same correctness label and $-\frac{\alpha}{T_{o_1}}\log \pi_t(o_1|q) \geqslant -\frac{\alpha}{T_{o_2}}\log \pi_t(o_2|q)$, we have

- If $\tilde{r}_t(q,o_1) \geqslant \frac{1}{\eta}\log\left(Z(q)\right)$ and $\tilde{r}_t(q,o_2) \geqslant \frac{1}{\eta}\log\left(Z(q)\right)$, then $\Delta_t(o_1|q) \geqslant 0$ and $\Delta_t(o_2|q) \geqslant 0$ but o_1 has more likelihood increase.
- If $\tilde{r}_t(q,o_1) \geqslant \frac{1}{\eta}\log\left(Z(q)\right)$ and $\tilde{r}_t(q,o_2) < \frac{1}{\eta}\log\left(Z(q)\right)$, then $\Delta_t(o_1|q) \geqslant 0$ and $\Delta_t(o_2|q) < 0$ where o_1's likelihood increase but o_2's likelihood decrease.
- If $\tilde{r}_t(q,o_1) < \frac{1}{\eta}\log\left(Z(q)\right)$ and $\tilde{r}_t(q,o_2) < \frac{1}{\eta}\log\left(Z(q)\right)$, then $\Delta_t(o_1|q) < 0$ and $\Delta_t(o_1|q) < 0$ but o_1 has less likelihood decrease.

- It is impossible that $\tilde{r}_t(q, o_1) < \frac{1}{\eta}\log(Z(q))$ and $\tilde{r}_t(q, o_2) \geqslant \frac{1}{\eta}\log(Z(q))$ given that (o_1, o_2) has the same correctness label and $-\frac{\alpha}{T_{o_1}}\log\pi_t(o_1|q) \geqslant -\frac{\alpha}{T_{o_2}}\log\pi_t(o_2|q)$.

□

B.2 Consistency of Multi-head Critic Curiosity

Linear MDP and Assumptions

Assumption B.2 (Linear MDP)**.** *We consider finite horizon $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, H)$ with horizon H, state space $\mathcal{S}$, action space $\mathcal{A}$, reward function $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and transition $P : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ such that there exists a known feature $\phi \in \mathbb{R}^d$ and unknown features $\theta, \psi \in \mathbb{R}^d$ to ensure*

$$R(s, a) = \phi(s, a)^\top\theta \qquad P(s'|s, a) = \phi(s, a)^\top\psi(s').$$

Without loss of generality, we assume $\|\phi(s, a)\| \leqslant 1$ for all (s, a), and $\|\psi(s')\| \leqslant \sqrt{d}$, $\|\theta\|_2 \leqslant \sqrt{d}$.

Lemma B.3 (Proposition 2.3 in Jin et al. (2020))**.** *For linear MDPs that satisfy Assumption B.2, there exists $w_h^\star \in \mathbb{R}^d$ such that*

$$Q_h^\pi(s, a) := \mathbb{E}\Big[\sum_{t=h}^{H} r_t \big| s_h = s, a_h = a\Big] = \phi(s, a)^\top w_h^\star.$$

The linearity of Q-functions enables using regression technique to solve it. Consider a dataset with n observations $\mathcal{D} = \{s_{i,h}, a_{i,h}, G_{i,h}\}_{i=1}^n$ where $G_{i,h}$ is the Monte-Carlo return. Let $\phi_{i,h} = \phi(s_{i,h}, a_{i,h})$ and denote the regression noise as $\varepsilon_{i,h} = G_{i,h} - \phi_{i,h}^\top w_h^\star$. We impose the following assumptions.

(A1) $\mathbb{E}[\varepsilon_{i,h} \mid \phi_{i,h}] = 0$ and $\{(\varepsilon_{i,h})\}_{i=1}^n$ are i.i.d. σ^2–sub-Gaussian for each fixed h;

(A2) $\frac{1}{n}\sum_{i=1}^n \phi_{i,h}\phi_{i,h}^\top \xrightarrow{\mathbb{P}} \Sigma_t \succ 0$

Jin et al. (2020) shows that doing value iteration on optimistically estimated Q function can achieve near-optimal regret for linear MDP, where the optimistic Q function is the combination of linear regression estimation and exploration bonus $b_{n,h} = \beta\sqrt{\phi_{n,h}^\top\Lambda_{n,h}^{-1}\phi_{n,h}}$, where $\Lambda_{n,h} = \lambda I + \sum_{i=1}^n \phi_{i,h}\phi_{i,h}^\top$ and β is some constant. Below we will formally connect our bootstrapped bonus with this term.

Formulation of the bootstrap multi-head critic

We accommodate the bootstrap multi-head into the linear-MDP setting. For any time step h, we sample K mini-batches $\{S_k \subset [n]\}_{k=1}^K$ of size $m = \zeta n$ uniformly without replacement from $\mathcal{D}$ and construct the ridge estimator as follows

$$\widehat{w}_{n,h}^{(k)} = \arg\min_w \sum_{r \in S_k} (G_{r,h} - \phi_{r,h}^\top w)^2 + \zeta\lambda\|w\|^2.$$

For any feature $\phi \in \mathbb{R}^d$, we define the bootstrap multi-head bonus as

$$b_{h,K}^{\text{boot}}(\phi) = \text{std}\left(\{\phi^\top\widehat{w}_{n,h}^{(k)} | 1 \leqslant k \leqslant K\}\right).$$

Elliptical ("count-based") bonus in (Jin et al., 2020). The ridge estimator is constructed using all data across n trajectories as follows

$$\widehat{w}_{n,h} = \arg\min_w \sum_{i=1}^{n} (G_{i,h} - \phi_{i,h}^\top w)^2 + \zeta\lambda\|w\|^2.$$

For any query feature $\phi \in \mathbb{R}^d$, the bonus term is $b_h^{\text{cnt}}(\phi) = \sqrt{\phi^\top\Lambda_{n,h}^{-1}\phi}$.

Theorem B.4. *Under Assumption B.2 and assumptions (A1)–(A2), for any fixed time-step h and query $\phi \in \mathbb{R}^d$,*

$$b_{h,K}^{boot}(\phi) \xrightarrow[K\to\infty,\, n\to\infty]{\mathbb{P}} \beta\sqrt{\phi^\top \Lambda_{n,h}^{-1}\phi},$$

where β is some constant.

Proof. For any time-step h and $S_k \subset [n]$, we have the explicit solution of the ridge regression

$$\widehat{w}_{n,h} = \Lambda_{n,h}^{-1} \sum_{i=1}^{n} \phi_{i,h} G_{i,h}.$$

Conditioning on $X_h = [\phi_{1,h}^\top; \ldots; \phi_{n,h}^\top]$, the conditional variance of the estimator is

$$\mathrm{Var}\big(\phi^\top \widehat{w}_{n,h} \mid X_h\big) = \sigma^2 \phi^\top \Big(\Lambda_{n,h}^{-1} - \lambda\Lambda_{n,h}^{-2}\Big)\phi.$$

From Assumption (A2), we have $\|\Lambda_{n,h}^{-1}\|_{\mathrm{op}} = O_p(1/n)$, therefore

$$\phi^\top \Lambda_{n,h}^{-1}\phi \leqslant \|\phi\|^2 \, \|\Lambda_{n,h}^{-1}\|_{\mathrm{op}} = O_p(1/n) \;\text{ and }\; \phi^\top \Lambda_{n,h}^{-2}\phi = O_p(1/n^2),$$

and

$$n\phi^\top \Big(\Lambda_{n,h}^{-1} - \lambda\Lambda_{n,h}^{-2}\Big)\phi - n\phi^\top \Lambda_{n,h}^{-1}\phi = -n\lambda\phi^\top \Lambda_{n,h}^{-2}\phi \xrightarrow{\mathbb{P}} 0.$$

Therefore, we have

$$\phi^\top \Big(\Lambda_{n,h}^{-1} - \lambda\Lambda_{n,h}^{-2}\Big)\phi \xrightarrow{\mathbb{P}} \phi^\top \Lambda_{n,h}^{-1}\phi.$$

Before moving to $b_{h,K}^{\mathrm{boot}}(\phi)$, we define the following quantities

$$\Delta\Sigma = \frac{1}{\zeta}\sum_{r\in S_k} \phi_{r,h}\phi_{r,h}^\top - \sum_{i=1}^{n} \phi_{i,h}\phi_{i,h}^\top, \quad b = \sum_{i=1}^{n} \phi_{i,h} G_{i,h}, \quad b_s = \frac{1}{\zeta}\sum_{r\in S_k} \phi_{r,h} G_{r,h}, \quad \Delta b = b_s - b.$$

Since $\Sigma_t \succ 0$, matrix Bernstein for sampling without replacement yields $\|\Delta\Sigma\|_{\mathrm{op}} = O_p(\sqrt{n})$. Use the expansion

$$(\Lambda_{n,h} + \Delta\Sigma)^{-1} = \Lambda_{n,h}^{-1} - \Lambda_{n,h}^{-1}\Delta\Sigma\Lambda_{n,h}^{-1} + R_\Sigma, \quad \|R_\Sigma\|_{\mathrm{op}} = O_p(\|\Lambda_{n,h}^{-1}\|_{\mathrm{op}}^3 \|\Delta\Sigma\|_{\mathrm{op}}^2) = O_p(1/n^2).$$

The k-th bootstrap ridge solution is

$$\widehat{w}_{n,h}^{(k)} = (\Lambda_{n,h} + \Delta\Sigma)^{-1} b_s.$$

Subtracting $\widehat{w}_{n,h} = \Lambda_{n,h}^{-1} b$ and inserting the expansion,

$$\widehat{w}_{n,h}^{(k)} - \widehat{w}_{n,h} = \underbrace{\Lambda_{n,h}^{-1}\Delta b - \Lambda_{n,h}^{-1}\Delta\Sigma\, \widehat{w}_{n,h}}_{\text{first order}} + \underbrace{\big(-\Lambda_{n,h}^{-1}\Delta\Sigma\Lambda_{n,h}^{-1}\Delta b + R_\Sigma b_s\big)}_{=:r_n}.$$

Since $G_{i,h} = \phi_{i,h}^\top w_h + \epsilon_{i,h}$, for any ϕ we have

$$\phi^\top(\widehat{w}_{n,h}^{(k)} - \widehat{w}_{n,h}) = \phi^\top \Lambda_{n,h}^{-1}\Big(\frac{1}{\zeta}\sum_{r\in S_k} \phi_{r,h}\epsilon_{r,h} - \sum_{i=1}^{n} \phi_{i,h}\epsilon_{i,h}\Big) + \phi^\top \Lambda_{n,h}^{-1}\Delta\Sigma(w_h^\star - \widehat{w}_{n,h}) + \phi^\top r_n. \quad (1)$$

From standard results for ridge regression, we have $\|w_h^\star - \widehat{w}_{n,h}\|_2 = O_{\mathbb{P}}(1/\sqrt{n})$, thus we have the second term $\phi^\top \Lambda_{n,h}^{-1}\Delta\Sigma(w_h^\star - \widehat{w}_{n,h}) = O_{\mathbb{P}}(1/n)$. Similarly, for the last term we have

$$\phi^\top r_n \leqslant \|\phi\| \Big(\|\Lambda_{n,h}^{-1}\|_{\mathrm{op}}^2 \|\Delta\Sigma\|_{\mathrm{op}} \|\Delta b\|_{\mathrm{op}} + \|\Delta\Sigma\|_{\mathrm{op}} \|b_s\|_{\mathrm{op}}\Big) = O_{\mathbb{P}}(1/n).$$

Therefore, both terms are negligible at the $\sqrt{\cdot}$ scale. Condition on $(X_h, \{\epsilon_{i,h}\}_{i=1}^n)$ the only randomness comes from S. By finite-population sampling theory,

$$\mathrm{Var}^*\Big(\frac{1}{\zeta}\sum_{r\in S} \phi_{r,h}\epsilon_{r,h}\Big) = \frac{1-\zeta}{\zeta}\sum_{i=1}^{n} \phi_{i,h}\phi_{i,h}^\top \sigma^2.$$

Therefore,

$$\begin{aligned}\text{Var}^*\Big(\phi^\top(\widehat{w}_{n,h}^{(k)} - \widehat{w}_{n,h})\Big) &= \frac{1-\zeta}{\zeta}\sigma^2\phi^\top\Lambda_{n,h}^{-1}\Big(\sum_{i=1}^{n}\phi_{i,h}\phi_{i,h}^\top\Big)\Lambda_{n,h}^{-1}\phi + o_{\mathbb{P}}\big(1/n\big) \\ &= \frac{1-\zeta}{\zeta}\sigma^2\phi^\top\Big(\Lambda_{n,h}^{-1} - \lambda\Lambda_{n,h}^{-2}\Big)\phi + o_{\mathbb{P}}\big(1/n\big) \\ &= \frac{1-\zeta}{\zeta}\sigma^2\phi^\top\Lambda_{n,h}^{-1}\phi + o_{\mathbb{P}}\big(1/n\big)\end{aligned}$$

Finally, by the conditional strong law of large numbers, we have

$$b_{h,K}^{\text{boot}}(\phi) = \text{std}\left(\{\phi^\top\widehat{w}_{n,h}^{(k)} | 1 \leqslant k \leqslant K\}\right) \to_{\text{a.s.}} \sqrt{\text{Var}^*(\phi^\top\widehat{w}_{n,h}^{(k)})} \xrightarrow{\mathbb{P}} \sqrt{\frac{1-\zeta}{\zeta}}\sigma\sqrt{\phi^\top\Lambda_{n,h}^{-1}\phi}.$$

□

C The Algorithmic Framework of the Multi-head Critic PPO

In this section, we describe the training procedure of the multi-head PPO algorithm, which follows the standard stages of vanilla PPO: (i) generating trajectories with the actor, (ii) updating the actor, and (iii) updating the critic. The key distinction is that we incorporate the multi-head variance as an exploration bonus, encouraging the policy to visit under-explored regions.

- **Actor roll-out:** Given a prompt q, the actor generates a set of responses $\{o_1, \ldots, o_n\}$. Each response is denoted as $o_i = \{o_{i,1}, \ldots, o_{i,|o_i|}\}$. Correspondingly, we associate each response with a verifiable reward r_i. For clarity, we focus on the case of a single prompt q.
- **Actor update:** In this step, the advantage is estimated as

$$\widehat{A}_{i,t} = \underbrace{\sum_{l=t}^{|o_i|}(\gamma\lambda)^{l-t}\widehat{\delta}_{i,l}}_{\approx\tilde{A}_{i,t}} + \omega\min\left(\frac{|\tilde{A}_{i,t}|}{\kappa}, \alpha B_{\text{critic}}(q, o_{i,\leqslant t+1})\right). \tag{2}$$

  The advantage consists of two components. The first term, $\tilde{A}_{i,t}$, largely follows the standard advantage estimation in PPO, except that we exploit bootstrap estimators by using an *ensemble* of value functions rather than a single point estimate:

$$\widehat{\delta}_{i,l} = r_{i,l} + \frac{\gamma}{K}\sum_{j=1}^{K}\widehat{V}_j(q, o_{i,\leqslant l+1}) - \frac{1}{K}\sum_{j=1}^{K}\widehat{V}_j(q, o_{i,\leqslant l}).$$

  The second term of Equation 2 introduces the *multi-head critic bonus* (B_{critic}), governed by the bonus weight ω, clipping ratio κ, and scaling factor α (see discussion following Equation (**??**) for interpretation). Specifically, B_{critic} is defined as the standard deviation across the K value heads, encouraging exploration by assigning higher bonus to actions leading to uncertain/less-visited regions:

$$B_{\text{critic}}\,(q, o_{i,\leqslant t+1}) = \text{std}\left(\{\widehat{V}_j(q, o_{i,\leqslant t+1}) | 1 \leqslant j \leqslant K\}\right). \tag{3}$$

- **Critic update:** We use the collected roll-outs to update the critic. For notational convenience, let the dataset be

$$\mathcal{D} = \{(q, o_{i,\leqslant t}, r_i) | i \in [n], t \in [|o_i|]\}, \tag{4}$$

  consisting of (prompt, partial response, reward) triplets. For each critic head j, we sample without replacement a subset $\mathcal{D}_j \subset \mathcal{D}$ of size $|\mathcal{D}_j| = \zeta|\mathcal{D}|$, where the hyperparameter $\zeta \in (0, 1]$ controls the fraction of data assigned per head. Smaller ζ increases head diversity, while larger ζ improves sample efficiency. The multi-head critic is then updated with the following bootstrap loss:

$$\mathcal{L}_\phi = \frac{1}{\zeta K|\mathcal{D}|}\sum_{j=1}^{K}\sum_{(q,o,r)\in\mathcal{D}_j}\Big(\widehat{V}_j(q, o) - r\Big)^2.$$

D Training Details

We use `verl` as the training framework[1]. Configurations for training CDE and baseline models are listed in Table 2.

Config	**GRPO**	**PPO**
actor-lr	1e-6	1e-6
critic-lr	-	1e-5
critic-warmup	-	10
kl_coef	0.0	0.0
max_prompt_length	2K	2K
max_response_length	3K	3K
train_batch_size	256	512
ppo_mini_batch_size	256	256
clip_ratio	0.20	0.20
sample temperature	1.0	1.0
rollout.n	8	4
total_training_steps	300	300

(a)

Config	**PPL**	**2,4 Heads**	**8,16 Heads**
κ	3	3	3
α	1	0.5	0.5
ω	Staircase	No decay	No decay
ζ	-	1	0.5

(b)

Table 2: (a) Baseline training configurations. The **GRPO** setup is shared across all GRPO-based methods (e.g., "Qwen3-4B-Base-GRPO" and "w/PPL bonus" in Table 1); likewise, the **PPO** setup is shared across all PPO-based methods. (b) **CDE**-specific configurations. The **PPL** settings are identical for both the GRPO "w/PPL bonus" and PPO "w/PPL bonus" variants.

Solve the following math problem step by step. The last line of your response should be of the form Answer: $Answer (without quotes) where $Answer is the answer to the problem.
{Problem}
Remember to put your answer on its own line after "Answer:".

Figure 4: The prompt for RLVR training.

E Ablation Studies

Bonus weight decay is crucial We compare four schedules for the bonus weight ω—*No decay*, *Linear*, *Cosine*, and *Staircase*—as illustrated in Figure 5, with the performance of models trained under each schedule summarized in Table 3. Briefly, the *No decay* schedule maintains strong exploration throughout training, while the *Staircase* schedule reduces ω abruptly, enabling strong exploration in the early phase and then removing the bonus for final convergence. The *Linear* and *Cosine* schedules provide intermediate behaviors.

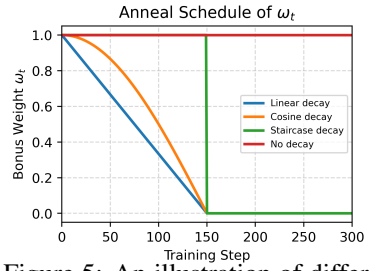


Figure 5: An illustration of different weight anneal schedules.

The results in Table 3 underscore two insights: First, decay of the bonus weight is necessary, as all decay schedules outperform the no-decay baseline by enabling a gradual shift from exploration to exploitation. Second, strong exploration in the early phase is crucial, with the staircase scheme proving most effective by sustaining high exploration initially to broaden state–action coverage and then removing the bonus abruptly to allow stable convergence, whereas the gentler cosine and linear decays weaken the signal too soon and thus yield smaller gains.

[1] `https://github.com/volcengine/verl`

Model	MATH	AMC23		AIME24		AIME25		Avg
	Avg@1	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	
Bonus Weight Decay Schedules								
Qwen3-4B-Base-GRPO	87.3	63.6	91.1	21.0	41.9	20.8	39.2	48.2
↳ ω **No decay**	85.1	64.5	84.6	20.8	39.0	22.3	36.2	48.2
↳ ω **Linear decay**	85.4	66.1	91.9	23.3	40.4	20.0	40.4	48.7
↳ ω **Cosine decay**	86.7	68.1	90.0	22.5	44.9	21.5	40.7	49.7
↳ ω **Staircase decay**	87.7	67.8	89.2	23.5	48.5	23.3	40.3	50.6

Table 3: Zero-shot accuracy of GRPO models under different PPL bonus weight decay schedules. The schedules follow those illustrated in Figure 5.

Analysis of sub-sample fraction ζ during critic update Additionally, we examine the sensitivity of the critic update to the hyperparameter ζ (sub-sample fraction). We vary ζ under two configurations—critics with 16 heads and with 4 heads—and compare $\zeta \in \{0.5, 1\}$. As shown in Table 4, while a larger number of heads benefits from a larger sub-sample fraction, the overall performance is stable across settings. The model demonstrates robustness to the masking fraction ζ, achieving similar results for both values tested (0.5 and 1.0).

Model	MATH	AMC23		AIME24		AIME25		Avg
	Avg@1	Avg@16	Pass@16	Avg@16	Pass@16	Avg@16	Pass@16	
Mask fraction								
16 Heads ; $\zeta = 0.5$	88.3	65.0	88.7	20.5	41.9	20.0	38.8	48.6
16 Heads ; $\zeta = 1$	85.4	65.3	85.3	21.0	39.2	21.7	43.2	48.4
4 Heads ; $\zeta = 0.5$	86.1	66.4	85.8	18.1	36.7	23.1	39.1	48.4
4 Heads ; $\zeta = 1$	87.3	63.9	87.9	21.5	35.5	21.5	45.5	48.5

Table 4: Ablation study on sub-sample fraction ζ.