EFFICIENT AND STABLE SCALING OF REINFORCEMENT LEARNING FOR LLMS VIA DYNAMIC ALLOCATION AND GRADIENT MODULATION

Yangyi Fang[1,*], **Jiaye Lin**[1,*,†], **Xiaoliang Fu**[2,*], **Cong Qin**[3,*], **Haolin Shi**[1,*], **Chaowen Hu**[4,*]

[1]Tsinghua University [2]Fudan University [3]Peking University [4]Zhejiang University

ABSTRACT

Post-training Large Language Models (LLMs) via Reinforcement Learning with Verifiable Rewards (RLVR) is a compute-intensive process where efficiency and stability are paramount for scaling. Current methods suffer from suboptimal resource allocation, distributing rollout budgets uniformly regardless of problem difficulty, and token-level optimization instability caused by the softmax policy structure. We propose **DynaMO**, a dual-pronged framework designed to scale RLVR effectively. At the sequence level, we introduce a variance-minimizing dynamic rollout allocation that concentrates compute on high-informativeness problems. At the token level, we develop gradient-aware advantage modulation to compensate for gradient attenuation in high-confidence actions while stabilizing excessive updates. Experiments on Qwen2.5-Math (1.5B, 7B) and Qwen3 (14B) across six benchmarks demonstrate that DynaMO significantly improves performance and training stability, offering a scalable pathway for reasoning optimization.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) has become a cornerstone for advancing reasoning capabilities in Large Language Models (LLMs), as evidenced by models like OpenAI o1 Jaech et al. (2024) and DeepSeek-R1 Guo et al. (2025). However, scaling RLVR presents distinct challenges compared to pre-training. Post-training is increasingly compute-bound by the generation of rollouts (sampling), and the optimization dynamics of policy gradients can be notoriously unstable when scaling to complex reasoning tasks.

Two fundamental inefficiencies persist in current RLVR paradigms (e.g., GRPO Shao et al. (2024)). First, standard methods employ *uniform rollout allocation*, ignoring the heterogeneous gradient informativeness across training samples. This wastes substantial compute on trivial or impossible problems, rather than focusing on the "learning frontier" where gradient variance implies high information gain. Second, the mathematical structure of softmax policies induces a *gradient magnitude imbalance*: high-confidence correct actions yield attenuated gradients Li (2025) (limiting reinforcement), while sudden entropy shifts can trigger exploding updates Luo et al. (2025) that destabilize training at scale.

To address these scaling bottlenecks, we propose **DynaMO** (**Dyna**mic Rollout Allocation and Advantage **M**odulation for Policy **O**ptimization). We theoretically derive an optimal rollout schedule based on gradient variance minimization, implemented via a lightweight Bernoulli proxy. Complementing this, we introduce a token-level modulation mechanism that balances exploration and exploitation by compensating for gradient attenuation and penalizing update instability. Our approach is validated across multiple model scales (up to 14B), showing consistent gains in sample efficiency and stability.

[*] Equal contribution. [†] Corresponding author.

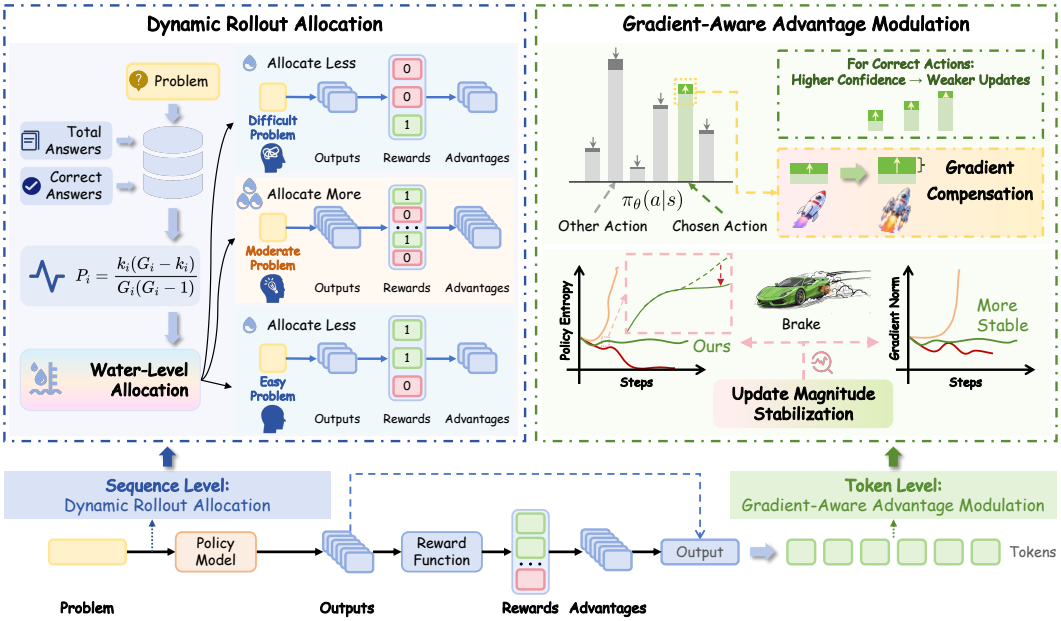


Figure 1: Overview of DynaMO. (i) Left: Dynamic allocation concentrates the rollout budget on high-variance problems to maximize gradient informativeness. (ii) Right: Gradient-aware advantage modulation compensates for attenuated gradients in high-confidence tokens and stabilizes excessive updates indicated by entropy spikes.

2 METHODOLOGY

DynaMO operates at two granularities: sequence-level resource allocation and token-level optimization control (Figure 1).

2.1 VARIANCE-DRIVEN DYNAMIC ROLLOUT ALLOCATION

Sample efficiency in RLVR is governed by how effectively the rollout budget B is distributed across N prompts. We formulate this as a gradient variance minimization problem.

Optimal Allocation Principle. For a prompt q_i with n_i rollouts, the gradient estimator variance is $\text{Var}[\hat{g}_i] = \sigma_i^2/n_i$. Minimizing the total variance $\sum_i \text{Var}[\hat{g}_i]$ under the constraint $\sum_i n_i = B$ yields the optimal allocation (derivation in Appendix C):

$$n_i^* = B \cdot \frac{\sigma_i}{\sum_{k=1}^{N} \sigma_k}. \tag{1}$$

This principle follows from standard variance minimization in stochastic optimization Schulman et al. (2017), implying rollouts should be proportional to the gradient standard deviation σ_i.

Bernoulli Variance Proxy. Directly computing σ_i is expensive. We propose a computable proxy based on the observation that gradient variance correlates with reward variance in group-normalized policy gradients Shao et al. (2024). For binary verifiable rewards, we estimate the variance using historical success statistics (k_i correct out of G_i attempts):

$$P_i = \frac{k_i(G_i - k_i)}{G_i(G_i - 1)}. \tag{2}$$

P_i peaks when $k_i \approx G_i/2$, identifying problems at the model's current capability boundary where learning signals are most informative. We employ a "water-filling" algorithm to dynamically update rollout counts G_i for each prompt based on P_i, constrained by $[G_{\min}, G_{\max}]$.

2.2 GRADIENT-AWARE ADVANTAGE MODULATION

To ensure stable scaling, we address the optimization dynamics at the token level. Our analysis (Appendix B) reveals that the expected squared update magnitude is bounded by entropy: $\mathbb{E}[\|\Delta z\|_2^2] \leq \eta^2 \mathbb{E}[A^2](1 - e^{-\mathcal{H}})$. This creates two issues: (1) *Attenuation*: High-confidence actions (low $\mathcal{H}$) have vanishing gradients Li (2025); (2) *Instability*: Sudden entropy increases imply large gradients that can destabilize training Luo et al. (2025).

Gradient Compensation. To counteract attenuation for confident correct actions, we apply a compensation factor $\beta_{i,t}^{\text{comp}}$ that scales inversely with entropy for positive advantages:

$$\beta_{i,t}^{\text{comp}} = \mathbb{I}[A_{i,t} > 0] \cdot \left(1 + \alpha \frac{\mathcal{H}_{\max} - \mathcal{H}_{i,t}}{\mathcal{H}_{\max} - \mathcal{H}_{\min}}\right) + \mathbb{I}[A_{i,t} \leq 0]. \tag{3}$$

Update Magnitude Stabilization. We use the magnitude of entropy change $\Xi_{i,t} = |\Delta\mathcal{H}(\pi_\theta|s_{i,t})|$ as a realtime indicator of optimization instability. We dampen updates for tokens exhibiting excessive entropy shifts using a sigmoid decay:

$$\beta_{i,t}^{\text{stab}} = f\left(\frac{\Xi_{i,t}}{\max_j \Xi_{j,t}}\right), \quad f(x) = \lambda_{\min} + (1 - \lambda_{\min})\sigma(-\gamma(x - \tau)). \tag{4}$$

Integrated Objective. The final advantages are modulated as $A_{i,t}^{\text{final}} = A_{i,t} \cdot \beta_{i,t}^{\text{comp}} \cdot \beta_{i,t}^{\text{stab}}$. These are used in the standard GRPO objective Shao et al. (2024) (see Appendix I for full formulation).

3 EXPERIMENTS

We evaluate DynaMO on Qwen2.5-Math (1.5B, 7B) and Qwen3 (14B) using the VeRL framework Sheng et al. (2025). We use the DAPO-Math-17k dataset Yu et al. (2025) and test on six benchmarks: AIME24/25 MAA (2025), AMC23 MAA (2023), MATH500 Hendrycks et al. (2021), Minerva Lewkowycz et al. (2022), and Olympiad He et al. (2024) (detailed setup in Appendix I and J).

3.1 MAIN RESULTS

Table 1 presents the comprehensive comparison results across six mathematical reasoning benchmarks on Qwen2.5-Math-1.5B and Qwen2.5-Math-7B. DynaMO consistently outperforms all baseline methods by effectively addressing two complementary challenges: the variance-driven rollout allocation concentrates computational budget on problems with balanced success-failure distributions where Bernoulli variance signals high gradient informativeness, while the gradient-aware advantage modulation provides compensation for gradient attenuation in high-confidence correct actions and stabilization against excessive update magnitudes signaled by large entropy changes. Furthermore, a comparison with entropy intervention baselines reveals their critical limitations: coarse-grained clipping and sequence-level reweighting methods like Clip-Higher, Clip-COV, and KL-COV lack fine-grained control over token-level dynamics, while entropy-induced advantage methods such as Entropy Advantages and W-REINFORCE introduce training instability without principled stabilization mechanisms. Results on Qwen3-14B demonstrating effective scaling are presented in Section 3.2.

3.2 SCALING AND STABILITY ANALYSIS

Scalability. Figure 2 validates DynaMO on models from 1.5B to 14B parameters. The performance gap over GRPO widens with model size, demonstrating that variance-driven allocation and gradient-aware modulation scale effectively to larger models where optimization stability becomes more critical.

Training Stability. As shown in Figure 3, DynaMO maintains significantly more stable gradient norms throughout training. Standard GRPO exhibits severe gradient spikes caused by sudden entropy fluctuations during optimization, particularly pronounced in complex reasoning tasks. Our stabilization term β^{stab} effectively mitigates this by dampening updates for tokens with excessive entropy changes.

Table 1: Comparison of benchmark results across Qwen2.5-Math-1.5B and Qwen2.5-Math-7B. Pass@K (%) is abbreviated as P@K. The best results are bold, and the second-best results are underlined, respectively.

Method	AIME24		AIME25		AMC23		MATH500		Minerva		Olympiad		Avg.	
	P@1	P@32	P@1	P@32	P@1	P@32	P@1	P@32	P@1	P@32	P@1	P@32	P@1	P@32
Qwen2.5-Math-1.5B														
GRPO	13.2	32.3	7.6	31.5	56.0	90.0	54.4	79.2	17.2	42.8	25.6	47.0	29.0	53.8
Clip-Higher	12.4	34.7	6.4	30.6	50.6	89.9	56.8	80.2	16.8	41.3	26.4	46.8	28.2	53.9
Entropy Loss	12.6	33.7	5.8	28.4	55.6	86.9	56.3	78.5	17.6	43.6	25.4	46.4	28.9	52.9
Fork Tokens	9.4	32.0	5.9	31.4	52.5	85.6	54.3	74.2	16.6	36.8	25.5	45.2	27.4	50.9
Entropy Advantages	15.7	35.8	8.9	33.4	62.0	86.4	**59.7**	76.2	18.2	43.0	25.9	44.9	31.7	53.3
Clip-COV	13.5	36.4	6.6	**34.4**	59.5	89.7	57.6	75.6	15.8	**44.3**	25.8	47.6	29.8	54.7
KL-COV	12.6	33.9	9.0	33.4	55.8	91.3	54.2	78.1	14.8	40.3	25.4	**48.1**	28.6	54.2
W-REINFORCE	15.3	35.3	8.5	31.7	63.0	85.7	56.7	77.7	18.2	40.3	24.4	46.2	31.0	52.8
DynaMO (Ours)	**17.2**	**37.2**	**9.8**	32.5	**63.6**	**91.9**	58.8	**81.0**	**19.4**	44.0	**27.2**	47.1	**32.7**	**55.6**
Qwen2.5-Math-7B														
GRPO	28.8	52.5	11.7	34.8	68.3	90.8	63.3	75.0	22.6	45.4	28.6	44.7	37.2	57.2
Clip-Higher	27.0	51.9	12.1	39.5	67.8	89.9	64.2	83.6	24.0	46.1	28.1	46.3	37.2	59.6
Entropy Loss	30.6	54.6	13.2	40.6	66.0	87.0	60.6	79.6	23.3	45.9	30.2	41.1	37.3	58.1
Fork Tokens	27.1	52.5	13.4	43.5	71.0	87.3	65.8	79.3	26.1	42.4	30.9	47.3	39.1	58.7
Entropy Advantages	27.5	49.7	9.4	39.2	67.9	85.2	65.3	83.3	23.7	43.7	30.4	47.3	37.4	58.1
Clip-COV	32.2	52.7	13.2	40.4	72.7	89.3	64.3	76.8	25.4	45.9	29.5	44.6	39.5	58.3
KL-COV	32.8	53.3	11.7	36.1	70.6	88.5	64.6	75.3	24.5	39.9	30.2	44.2	39.1	56.2
W-REINFORCE	31.8	55.4	14.3	41.0	72.5	89.8	64.9	**84.0**	26.4	**49.5**	30.9	46.7	40.1	61.1
DynaMO (Ours)	**34.4**	**59.0**	**15.4**	**46.8**	**74.4**	**92.9**	**66.4**	**84.0**	**27.3**	47.2	**31.6**	**50.1**	**41.6**	**63.3**

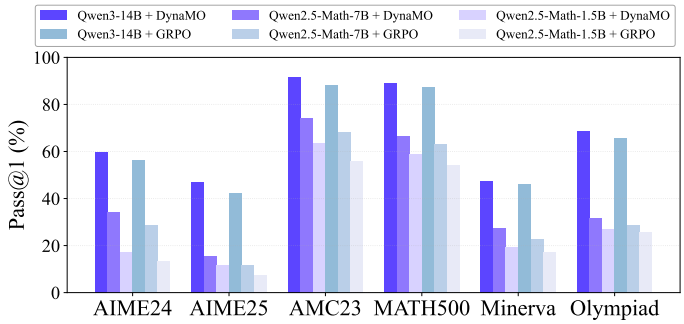


Figure 2: Performance comparison across LLM scales (1.5B/7B/14B). The gap between DynaMO and GRPO widens as model size increases, validating scalability.

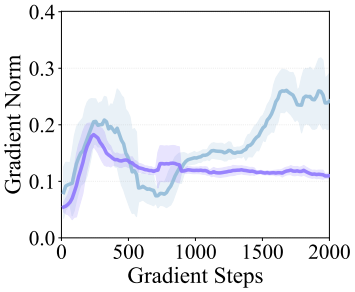


Figure 3: Gradient norms on AIME24. DynaMO (purple) eliminates the instability spikes observed in GRPO (blue).

4 Conclusion

In this paper, we propose **DynaMO**, a theoretically-grounded dual-pronged framework designed to systematically address fundamental RLVR challenges. *At the sequence level*, we prove that uniform allocation is suboptimal and subsequently derive a variance-minimizing allocation strategy to concentrate computational resources on high-informativeness problems. *At the token level*, we establish the gradient-entropy relationship, enabling an integrated advantage modulation mechanism that compensates for gradient attenuation while stabilizing excessive updates. Extensive experiments

conducted across multiple reasoning benchmarks and varying LLM scales demonstrate consistent improvements over strong baselines, with comprehensive ablation studies validating the independent contributions from both mechanisms.

REFERENCES

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *International Conference on Machine Learning (ICML)*, 2009.

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

Xiangxiang Chu, Hailang Huang, Xiao Zhang, Fei Wei, and Yong Wang. Gpg: A simple and strong reinforcement learning baseline for model reasoning. *arXiv preprint arXiv:2504.02546*, 2025.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

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 (ICML)*, 2018.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, et al. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems. *arXiv preprint arXiv:2402.14008*, 2024.

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.

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.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Yingru Li. Logit dynamics in softmax policy gradient methods. *arXiv preprint arXiv:2506.12912*, 2025.

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

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

MAA. American mathematics competitions - amc, 2023. URL `https://maa.org/`.

MAA. American invitational mathematics examination - aime, 2025. URL `https://maa.org/`.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2016.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

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, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *European Conference on Computer Systems (EuroSys)*, 2025.

Hongze Tan and Jianfei Pan. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy. *arXiv preprint arXiv:2508.04349*, 2025.

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.

Yuxuan Tong, Xiwen Zhang, Rui Wang, Ruidong Wu, and Junxian He. Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

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

Shihui Yang, Chengfeng Dou, Peidong Guo, Kai Lu, Qiang Ju, Fei Deng, and Rihui Xin. Dcpo: Dynamic clipping policy optimization. *arXiv preprint arXiv:2509.02333*, 2025.

Jiarui Yao, Yifan Hao, Hanning Zhang, Hanze Dong, Wei Xiong, Nan Jiang, and Tong Zhang. Optimizing chain-of-thought reasoners via gradient variance minimization in rejection sampling and rl. *arXiv preprint arXiv:2505.02391*, 2025.

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

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *arXiv preprint arXiv:2506.01347*, 2025.

A Entropy Change Derivation

A.1 Definitions

Definition 1 (Centered Log-Probability)**.**

$$\Lambda_\theta(a|s) := \log \pi_\theta(a|s) + \mathcal{H}(\pi_\theta|s). \tag{5}$$

Definition 2 (Centered Logit Change)**.**

$$\delta z_{s,a} := z_{s,a}^{k+1} - z_{s,a}^{k} - \mathbb{E}_{a' \sim \pi_\theta^k(\cdot|s)}[z_{s,a'}^{k+1} - z_{s,a'}^{k}]. \tag{6}$$

A.2 Entropy Gradient

First-order Taylor expansion:

$$\mathcal{H}(\pi_\theta^{k+1} \mid s) \approx \mathcal{H}(\pi_\theta^{k} \mid s) + \langle \nabla \mathcal{H}(\pi_\theta^{k} \mid s), (z^{k+1} - z^{k}) \rangle. \tag{7}$$

Entropy gradient:

$$\nabla_\theta \mathcal{H}(\pi_\theta \mid s) = -\mathbb{E}_{a \sim \pi_\theta(\cdot|s)}[\log \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s)], \tag{8}$$

since $\mathbb{E}_{a \sim \pi_\theta}[\nabla_\theta \log \pi_\theta(a \mid s)] = 0$.

Softmax derivative:

$$\frac{\partial \log \pi_\theta(a \mid s)}{\partial z_{s,a'}} = \mathbf{1}\{a = a'\} - \pi_\theta(a' \mid s). \tag{9}$$

Substituting:

$$\langle \nabla_\theta \mathcal{H}(\theta^k \mid s), (z^{k+1} - z^k) \rangle = -\mathbb{E}_{a \sim \pi_\theta^k}[\log \pi_\theta^k(a \mid s) \cdot \delta z_{s,a}]. \tag{10}$$

Using $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] + \mathrm{Cov}(X, Y)$ and $-\mathbb{E}[\log \pi_\theta] = \mathcal{H}(\pi_\theta|s)$:

$$\langle \nabla_\theta \mathcal{H}(\theta^k \mid s), (z^{k+1} - z^k) \rangle = -\mathbb{E}_{a \sim \pi_\theta^k}[\Lambda_\theta^k(a|s) \cdot \delta z_{s,a}]. \tag{11}$$

A.3 GRPO Update

Define composite update coefficient:

$$\xi_{i,t}(a) := \mathbb{I}_{\text{clip}} \cdot r_{i,t} \cdot A_{i,t}. \tag{12}$$

Lemma 1 (GRPO Logit Update)**.**

$$z_{s,k}^{k+1} - z_{s,k}^{k} = \eta\, \xi_{i,t}\, (\mathbf{1}\{a = k\} - \pi_\theta(k \mid s)). \tag{13}$$

Theorem 1 (Factorized Entropy Change)**.**

$$\boxed{\Delta \mathcal{H}(\pi_\theta^k|s) \approx -\eta \sum_a \pi_\theta^k(a|s)^2 \cdot \Lambda_\theta^k(a|s) \cdot \xi_{i,t}(a)} \tag{14}$$

Proof. Substitute Lemma 1 and expand expectation:

$$\begin{aligned} \Delta \mathcal{H}(\pi_\theta^k|s) &\approx -\eta\, \mathbb{E}_{a \sim \pi_\theta^k}[\pi_\theta^k(a \mid s) \Lambda_\theta^k(a|s) \xi_{i,t}(a)] \\ &= -\eta \sum_a \pi_\theta^k(a|s)^2 \Lambda_\theta^k(a|s) \xi_{i,t}(a). \end{aligned} \tag{15}$$

□

B GRADIENT-ENTROPY RELATIONSHIP

For softmax policy $\pi_k = \exp(z_k)/\sum_j \exp(z_j)$ with $\frac{\partial \log \pi_k}{\partial z_i} = \delta_{ik} - \pi_i$:

$$\|\nabla_z \log \pi_k\|^2 = 1 - 2\pi_k + \sum_{j=1}^{|\mathcal{V}|} \pi_j^2. \tag{16}$$

Taking expectation:

$$\mathbb{E}_{a_k \sim \pi_\theta}\left[\|\nabla_z \log \pi_k\|^2\right] = 1 - \sum_{k=1}^{|\mathcal{V}|} \pi_k^2. \tag{17}$$

By Jensen's inequality ($x \mapsto \log x$ concave):

$$\sum_{k=1}^{|\mathcal{V}|} \pi_k^2 \geq e^{-\mathcal{H}(\pi_\theta|s)}. \tag{18}$$

For advantage-weighted updates $\Delta z(s) = \eta A(\mathbf{e}_k - \pi_\theta(\cdot|s))$:

$$\mathbb{E}_{a_k \sim \pi_\theta}\left[\|\Delta z(s)\|_2^2\right] \leq \eta^2 \mathbb{E}[A^2]\left(1 - e^{-\mathcal{H}(\pi_\theta|s)}\right). \tag{19}$$

C GRADIENT VARIANCE MINIMIZATION

For gradient estimator $\hat{g}_i = \frac{1}{n_i}\sum_{k=1}^{n_i} g_{i,k}$ with $\text{Var}[\hat{g}_i] = \sigma_i^2/n_i$, minimizing $\sum_{i=1}^N \sigma_i^2/n_i$ subject to $\sum_i n_i = B$:

Lagrangian: $\mathcal{L} = \sum_i \sigma_i^2/n_i + \lambda(\sum_i n_i - B)$.

FOC: $-\sigma_i^2/n_i^2 + \lambda = 0 \Rightarrow n_i = \sigma_i/\sqrt{\lambda}$.

Substituting: $\sum_i \sigma_i/\sqrt{\lambda} = B \Rightarrow \sqrt{\lambda} = (\sum_k \sigma_k)/B$.

$$\boxed{n_i^* = B \cdot \frac{\sigma_i}{\sum_{k=1}^N \sigma_k}} \tag{20}$$

For GRPO, assuming gradient orthogonality:

$$\mathbb{E}[\|g_{i,k}\|_2^2] \propto \text{Var}(R) \cdot (1 - \bar{C}_i), \tag{21}$$

where $\bar{C}_i = \mathbb{E}[\frac{1}{|o|}\sum_t \sum_k \pi_k^2]$ is average collision probability.

For binary rewards, Bernoulli variance estimator:

$$P_i = \frac{k_i(G_i - k_i)}{G_i(G_i - 1)}, \quad \mathbb{E}[P_i] = p_i(1 - p_i). \tag{22}$$

Use P_i as proxy for σ_i: $n_i^* \propto P_i$ with constraints $[G_{\min}, G_{\max}]$.

Variance reduction by Cauchy-Schwarz:

$$\frac{\text{Var}^*}{\text{Var}_{\text{uniform}}} = \frac{(\sum_i \sigma_i)^2}{N\sum_i \sigma_i^2} \leq 1. \tag{23}$$

D DYNAMIC ROLLOUT ALLOCATION

Water-level algorithm: Initialize $G_i^{\text{new}} = G_{\min}$, $B_{\text{rem}} = B - NG_{\min}$. Iteratively allocate residual budget proportional to P_i until $\sum_i G_i^{\text{new}} = B$ or all prompts reach $G_{\max}$. Update statistics: $G_i \leftarrow G_i + G_i^{\text{new}}$, $k_i \leftarrow k_i + k_i^{\text{new}}$ after each iteration.

E Related Works

E.1 Reinforcement Learning for LLMs

Reinforcement learning has emerged as a dominant paradigm for LLM post-training, with RLHF and RLVR demonstrating significant success Ouyang et al. (2022); Bai et al. (2022); Schulman et al. (2017). Recent breakthrough models, including DeepSeek-R1 Guo et al. (2025), DeepSeekMath Shao et al. (2024), OpenAI o1 Jaech et al. (2024), and Kimi k1.5 Team et al. (2025), further demonstrate the effectiveness of RLVR on reasoning tasks with verifiable rewards. While subsequent works have introduced algorithmic refinements Liu et al. (2025); Yu et al. (2025); Chu et al. (2025); Hu et al. (2025), fundamental challenges in computational efficiency and optimization stability still persist.

E.2 Entropy Dynamics in Policy Optimization

Entropy regularization balances exploration and exploitation Haarnoja et al. (2018); Mnih et al. (2016), yet its role in LLM training remains contentious Ouyang et al. (2022); Shao et al. (2024); Yu et al. (2025); Chu et al. (2025). A central challenge is entropy collapse Luo et al. (2025), motivating mitigation strategies such as ratio clipping Yu et al. (2025); Yang et al. (2025), sample reweighting Zhu et al. (2025), or entropy-induced advantages Cheng et al. (2025); Tan & Pan (2025); Wang et al. (2025). Beyond entropy control, Li (2025) demonstrates that high-confidence actions yield attenuated gradient magnitudes, inducing asymmetric learning dynamics. However, existing methods lack unified theoretical grounding: coarse-grained interventions may amplify fluctuations, while the interplay between gradient attenuation and entropy dynamics remains underexplored. Our work addresses these challenges through gradient-aware policy update control grounded in the gradient-entropy relationship, where entropy serves as a computable indicator of update magnitude rather than a direct optimization target.

E.3 Sample Efficiency

Sample efficiency is critical for RLVR training, where generating multiple rollouts per problem incurs substantial cost. Standard methods employ uniform rollout budgets Shao et al. (2024), overlooking heterogeneous gradient informativeness. Recent works explore adaptive strategies from different perspectives: curriculum learning Bengio et al. (2009) and prioritized experience replay Schaul et al. (2015) strategically choose which problems to train on, but emphasize sample ordering rather than resource allocation. Offline budget allocation methods Tong et al. (2024) repeatedly sample until obtaining a fixed number of correct responses per prompt, lacking dynamic scheduling for iterative online training. EM-based methods Dong et al. (2023); Gulcehre et al. (2023); Yao et al. (2025) enhance efficiency via iterative rejection sampling with dynamic allocation, yet require gradient norm computations and target the EM framework rather than policy gradient methods. In contrast, we derive optimal rollout allocation by minimizing gradient variance specifically for policy gradient methods, establishing theoretical convergence guarantees with a lightweight, gradient-free proxy that relies on historical success statistics.

F Ablation Study

Table 2 demonstrates that all three components contribute independently to DynaMO's performance:

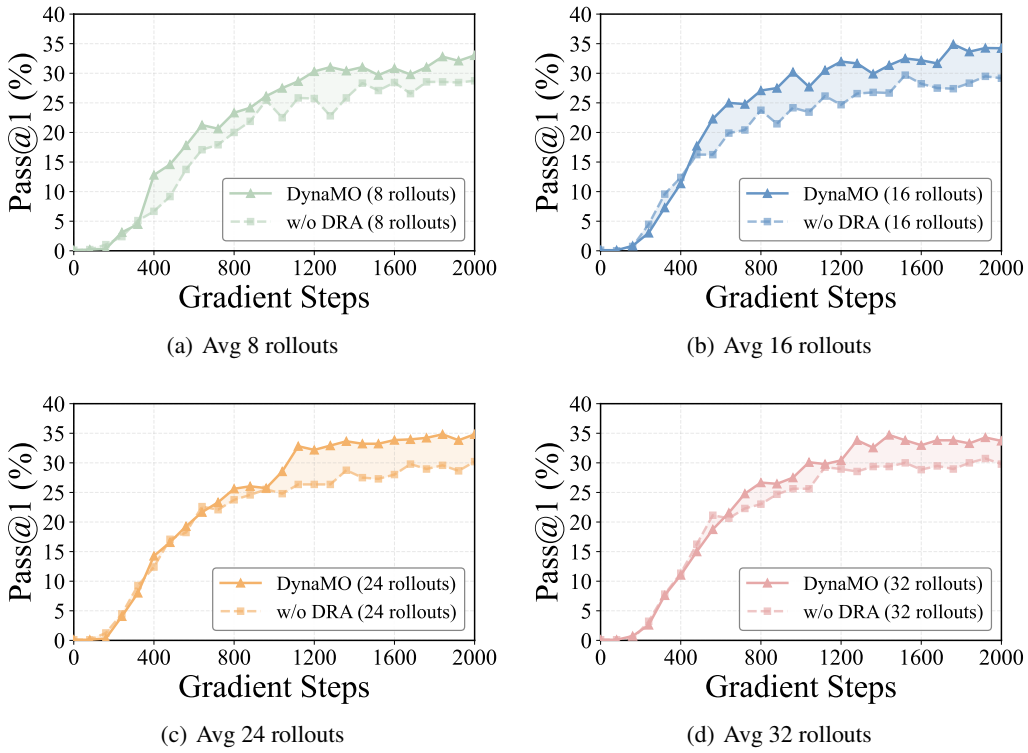
DRA shows the largest impact on Minerva (4.3% drop when removed), which contains problems with highly varied difficulty where adaptive allocation is most beneficial. GC most affects Olympiad (2.3% drop), where complex reasoning chains require stable gradients for high-confidence tokens. UMS provides consistent gains across all benchmarks (0.7-3.0% improvement), validating the importance of stabilization.

G Budget Sensitivity Analysis

Figure 4 evaluates DRA across varying computational budgets (8 to 32 rollouts per problem on average). Key observations:

Table 2: Complete ablation study on Qwen2.5-Math-7B with Pass@1 (%). DRA: Dynamic Rollout Allocation, UMS: Update Magnitude Stabilization, GC: Gradient Compensation.

Method	AIME24	AIME25	AMC23	MATH500	Minerva	Olympiad	Avg.
DynaMO	**34.4**	**15.4**	**74.4**	**66.4**	**27.3**	**31.6**	**41.6**
w/o GC	33.8	15.0	71.9	65.0	26.7	29.3	40.3
w/o UMS	33.2	14.7	71.9	64.9	25.9	30.2	40.1
w/o DRA	31.9	15.2	73.4	65.7	23.0	30.4	39.9
w/o GC & DRA	31.9	14.5	69.0	64.4	23.7	29.7	38.9
w/o GC & UMS	30.5	14.3	70.2	63.5	22.2	29.4	38.4
w/o UMS & DRA	30.0	13.8	70.1	61.6	19.9	30.2	37.6
GRPO (w/o ALL)	28.8	11.7	68.3	63.3	22.6	28.6	37.2



(a) Avg 8 rollouts (b) Avg 16 rollouts

(c) Avg 24 rollouts (d) Avg 32 rollouts

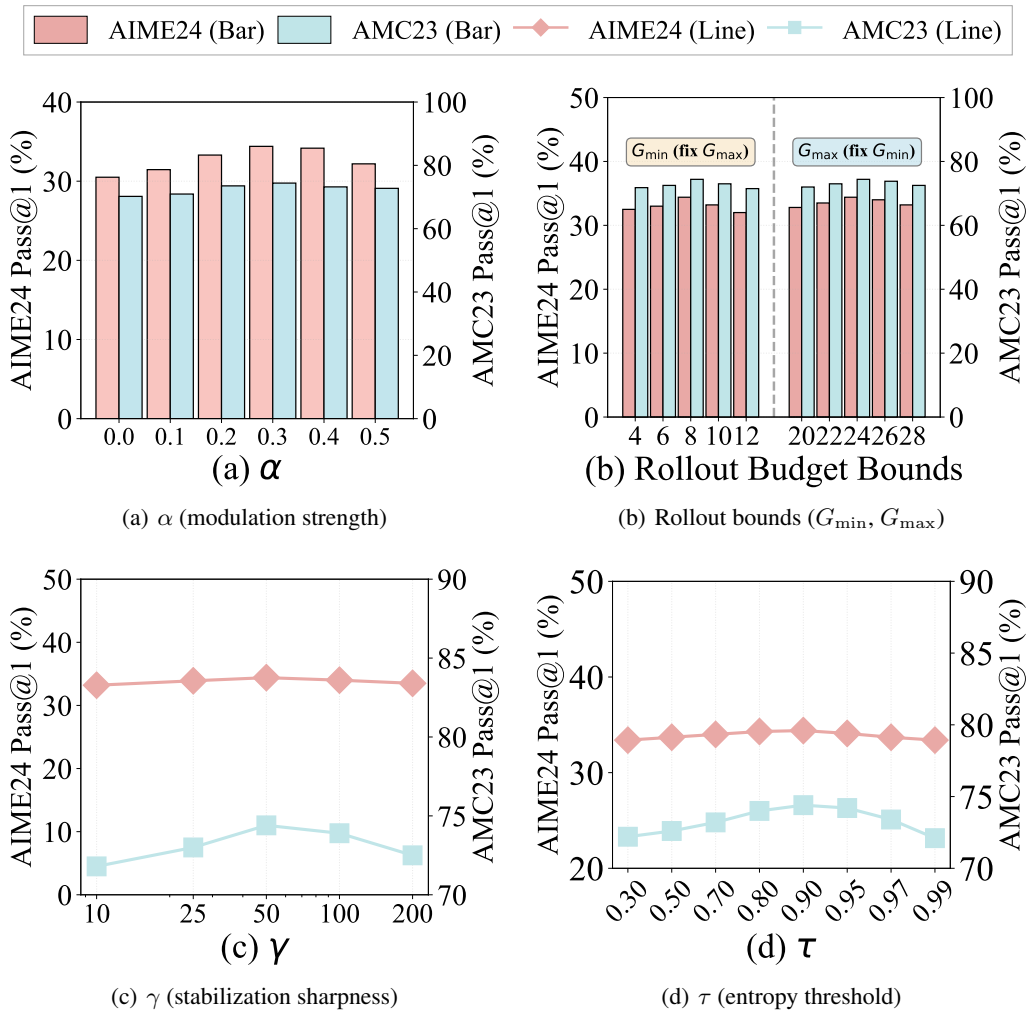
Figure 4: Impact of dynamic rollout allocation across different computational budgets on AIME24 with Pass@1 (%). Solid lines denote full DynaMO; dashed lines denote DynaMO w/o DRA (uniform allocation).

Consistent Gains Across Budgets. DRA provides stable performance improvements across all budget levels, with gains ranging from 1.5% (at 8 rollouts) to 2.3% (at 24 rollouts). This demonstrates that variance-driven allocation is beneficial regardless of absolute resource availability.

Learning Dynamics. During early training (steps 0-500), limited historical statistics result in near-uniform allocations, causing both variants to exhibit similar performance. As variance data accumulates (steps 500-2000), DRA progressively concentrates budget on problems with balanced success-failure distributions (maximizing Bernoulli variance P_i). These problems reside within the model's capability gap where both positive and negative learning signals are actively generated, maximizing gradient informativeness per computational unit.

Resource Efficiency. The performance gap between DRA and uniform allocation widens during mid-training (steps 500-1500), then stabilizes as the model approaches convergence. This pattern validates that DRA effectively identifies and exploits the "learning frontier" throughout training, whereas uniform allocation continues wasting resources on trivially-solved or currently-inaccessible problems.

H Hyperparameter Sensitivity Analysis



(a) α (modulation strength)

(b) Rollout bounds ($G_{\min}$, $G_{\max}$)

(c) γ (stabilization sharpness)

(d) τ (entropy threshold)

Figure 5: Hyperparameter sensitivity analysis on Qwen2.5-Math-7B across AIME24 and AMC23.

Figure 5 demonstrates robustness of DynaMO across hyperparameter ranges:

Modulation Strength (α). The unified modulation parameter exhibits an inverted-U pattern: performance degrades without modulation ($\alpha = 0$), peaks at moderate values ($\alpha = 0.3$), and declines at excessive settings ($\alpha \geq 0.5$). This validates that insufficient modulation fails to address gradient attenuation, while over-intervention constrains learning. The broad optimal plateau ($\alpha \in [0.2, 0.4]$) simplifies tuning across different model scales.

Allocation Bounds ($G_{\min}$, $G_{\max}$). Both minimum and maximum rollout constraints show stable performance across wide ranges. For $G_{\min} \in [4, 12]$ and $G_{\max} \in [16, 32]$, performance varies by less than 1%, confirming that variance-driven allocation naturally avoids pathological under/over-sampling. This robustness eliminates the need for careful bound tuning.

Stabilization Sharpness (γ). The sigmoid transition parameter exhibits a flat plateau around optimal values ($\gamma \in [3, 7]$), balancing selective intervention against learning flexibility. Values too low ($\gamma < 2$) apply stabilization too broadly, dampening useful updates; values too high ($\gamma > 10$) create abrupt transitions that may miss borderline unstable tokens.

Entropy Threshold (τ). The threshold determining when stabilization activates displays smooth variation with a broad optimal region ($\tau \in [0.4, 0.6]$). This demonstrates reliability in identifying problematic tokens: lower thresholds ($\tau < 0.3$) over-stabilize, while higher thresholds ($\tau > 0.7$) under-stabilize.

I Training Configuration

Table 3: Complete training hyperparameters.

Parameter	Value	Parameter	Value
Gen Batch Size	512	Clip Ratio	0.2
Update Batch Size	32	Entropy Coef	0
PPO Mini-batch	32	KL Coef	0.0
Avg Rollouts	16	Temperature	1.0
Rollout Range	[8, 24]	Top-p	1.0
LR (Actor)	$1e-6$	Max Prompt Len	2048
LR (Critic)	$1e-5$	Max Response Len	8192
Weight Decay	0.1/0.01	Rollout Engine	vLLM
Grad Clip	1.0	TP Size	2
Warmup Steps	10	GPU Memory Util	0.8

Hardware: 8× A100 80GB per node, bfloat16, FSDP. Answer verification: Math-Verify. Training prompts left-truncated to 2048 tokens if exceeding max length.

J Benchmark Details

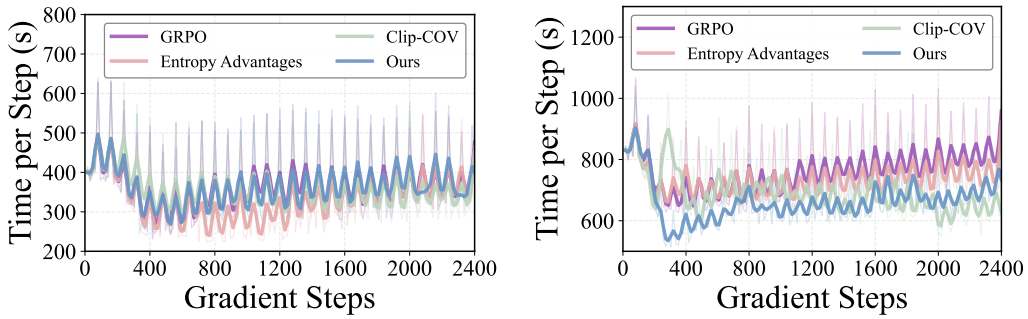
Table 4: Detailed benchmark characteristics.

Benchmark	Size	Format	Description
AIME24	30	Integer (0-999)	American Invitational Mathematics Examination 2024; high school competition
AIME25	30	Integer (0-999)	American Invitational Mathematics Examination 2025; algebra, geometry, number theory
AMC23	75	Multiple choice	American Mathematics Competitions 2023 (AMC 8/10/12); curriculum-aligned
MATH-500	500	Open-ended	Curated from MATH dataset; seven domains; five difficulty levels
Minerva	272	Open-ended	STEM problems (undergrad-grad level); symbolic manipulation
Olympiad	412	Open-ended	Bilingual International Mathematics Olympiad problems; theorem-proving

K Computational Efficiency

Figure 6 demonstrates that DynaMO maintains comparable training efficiency to baseline methods:

Overhead Analysis. DynaMO adds minimal overhead: $< 3\%$ for 1.5B model (8.2s vs 8.0s per step) and $< 2\%$ for 7B model (32.1s vs 31.5s per step). This negligible cost comes from:



(a) Qwen2.5-Math-1.5B (b) Qwen2.5-Math-7B

Figure 6: Per-step training time comparison. Bold lines: smoothed measurements; light lines: raw data.

- Bernoulli variance computation: $O(N)$ arithmetic operations
- Water-level allocation: $O(N \log N)$ sorting, amortized over batch
- Token-level modulation: simple arithmetic on existing logits

All operations are lightweight compared to model inference (forward/backward passes), confirming practical viability for production deployment.