

MINER: Mining Intrinsic Mastery for Data-Efficient RL in Large Reasoning Models

Shuyang Jiang^{♣,◇}, Yuhao Wang[♣], Ya Zhang^{♣,◇}, Yanfeng Wang^{♣,◇}, Yu Wang^{♣,◇}

[♣]Fudan University

[♣]Shanghai Jiao Tong University

[◇]Shanghai Artificial Intelligence Laboratory

shuyangjiang23@m.fudan.edu.cn

{colane,ya_zhang,wangyanfeng622,yuwangsjtu}@sjtu.edu.cn

Abstract

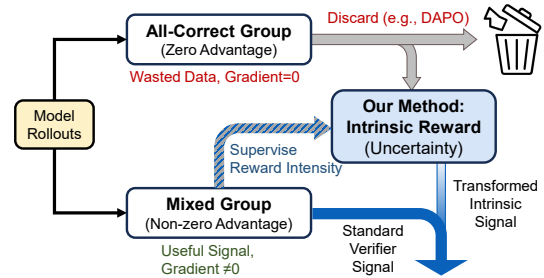
Current critic-free RL methods for large reasoning models suffer from severe inefficiency when training on positive homogeneous prompts (where all rollouts are correct), resulting in waste of rollouts due to zero advantage estimates. We introduce a radically simple yet powerful solution to Mine intrinsic mastery (MINER), that repurposes the policy’s intrinsic uncertainty as a self-supervised reward signal, with no external supervision, auxiliary models, or additional inference cost. Our method pioneers two key innovations: (1) a token-level focal credit assignment mechanism that dynamically amplifies gradients on critical uncertain tokens while suppressing overconfident ones, and (2) adaptive advantage calibration to seamlessly integrate intrinsic and verifiable rewards. Evaluated across six reasoning benchmarks on Qwen3-4B and Qwen3-8B base models, MINER achieves state-of-the-art performance among the other four algorithms, yielding up to **4.58** absolute gains in Pass@1 and **6.66** gains in Pass@K compared to GRPO. Comparison with other methods targeted at exploration enhancement further discloses the superiority of the two newly proposed innovations. This demonstrates that latent uncertainty exploitation is both necessary and sufficient for efficient and scalable RL training of reasoning models. Code is available at <https://github.com/pixas/Miner>.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a central recipe for training large reasoning models (LRMs), enabling substantial reasoning gains from outcome-only supervision without relying on dense reward models or learned critics. Critic-free algorithms such as GRPO (Shao et al., 2024), DAPO (Yu et al., 2025), and REINFORCE++ (Hu et al., 2025) scale favorably by

*Corresponding Author

(a) The GRPO Data Bottleneck: Wasted All-Correct Rollouts



(b) Our Method: Turning Waste Into Treasure

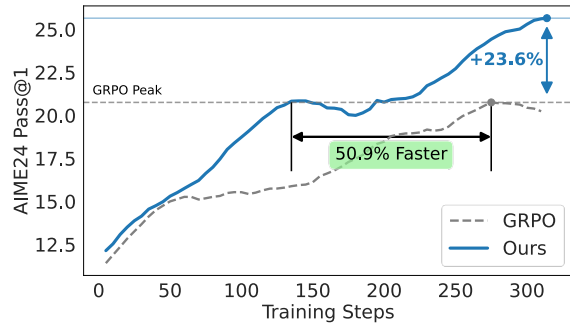


Figure 1: (a) Traditional GRPO algorithms produce a credible number of rollouts that do not contribute to RL updates, due to indistinguishable top rewards. (b) MINER introduces intrinsic rewards to each rollout, injecting beneficial dense reward signals, achieving the same peak performance with only 50% training steps, and up to 23% higher performance on Qwen3-4B-Base.

estimating advantages from multiple rollouts per prompt, making them particularly appealing for large-scale post-training. Yet, this multi-rollout paradigm exposes an increasingly dominant inefficiency as base models strengthen: many prompts yield rollouts that all receive identical verifier rewards. In these cases, GRPO-style relative advantage estimation collapses to (near-)zero, so the corresponding trajectories contribute no learning signal despite incurring full rollout cost (Liu et al., 2025b; Sun et al., 2025; Zhou et al., 2025). Crucially, in the high-accuracy regime, *positive homogeneous* (PH) prompts, where all sampled rollouts

are correct, can occupy a large portion of each batch, rendering a non-trivial fraction of expensive rollouts computationally wasteful (see Fig. 1(a)).

Existing efforts to mitigate homogeneity typically follow two paths, both of which introduce scaling trade-offs. The first seeks to improve data quality through pre-filtering (Xu et al., 2025; Yu et al., 2025; Zheng et al., 2025b), but these strategies inherently incur extra inference costs and their overhead grows as stronger models make more PH prompts. Another line reuses past trajectories through rollout buffers (Sun et al., 2025; Jiang et al., 2025), which introduces off-policy elements and distribution shifts that can complicate stability and large-scale deployment (Xi et al., 2025). This leaves a vital question unanswered: *can we extract learning signals from PH prompts with essentially zero marginal overhead, while preserving the integrity of the primary verifier objective?*

In this work, we provide an affirmative answer by re-examining PH prompts through a simple yet underexploited lens: while PH rollouts are equally *correct* according to the verifier, the underlying “hard” actions are not equally *mastered* by the policy. Many correct solutions are generated via fragile, low-confidence reasoning paths that remain under-optimized if PH prompts are discarded. However, naively rewarding uncertainty is incompatible with the RLVR objective, as injecting intrinsic signals into heterogeneous prompts can blur the correctness boundary or overwhelm the outcome reward (Fig. 3). To address this, we propose MINER, a data-efficient framework that selectively targets PH prompts, transforming intrinsic uncertainty into a *safe, bounded* learning signal.

MINER comprises three tightly coupled designs: (1) **Uncertainty-driven intrinsic rewards for PH only** where we define an intrinsic reward via per-token negative log-likelihood and apply **positive filtering** to exclusively reinforce under-confident but correct trajectories. This prevents the reinforcement of already-mastered modes and focuses optimization on the “fragile” reasoning paths. (2) **Token-level focal credit assignment**, which is a focal reweighting mechanism that concentrates gradients on bottleneck tokens along the reasoning chain, using token probabilities as a discriminative weight to avoid uniformly reinforcing trivial tokens. (3) **Adaptive advantage calibration** where we dynamically scale intrinsic advantages using a reference scale extracted from heterogeneous prompts within the same batch. This ensures a proper signal hier-

archy, prioritizing the optimization of correctness while integrating intrinsic signals at an appropriate magnitude. Notably, MINER requires no additional rollouts, hints, replay buffers, or auxiliary reward models. By reusing quantities already computed during the PPO-style optimization process, it adds negligible overhead while reclaiming the utility of otherwise wasted rollouts.

We evaluate on two base models (Qwen3-4B-Base (Qwen, 2025) and Qwen3-8B-Base) across six diverse reasoning benchmarks. MINER consistently outperforms GRPO variants and other strong baselines, achieving **+4.5** absolute Pass@1 on average and up to > 10 absolute gains in Pass@K on challenging benchmarks. Further analyses of exploration dynamics, calibration stability, and cross-task transferability confirm that PH-targeted uncertainty exploitation is a general and robust strategy for enhancing reasoning models. We summarize our contributions as follows:

1. **Uncertainty-Driven Self-Supervised Reward:** We introduce the first framework to transform a policy’s intrinsic uncertainty into an informative reward signal for homogeneous prompts. By eliminating the need for external hints or auxiliary models, we unlock training signals for approximately 25% of otherwise wasted rollouts at zero marginal cost.
2. **Token-Focal Credit Assignment Mechanism:** We propose a fine-grained focal weighting strategy that dynamically amplifies learning signals on critical, uncertain tokens while suppressing overconfident ones. This level of granularity overcomes sequence-level uniformity and prevents mode collapse, providing a precision entirely unexplored in existing RLVR literature.
3. **State-of-the-Art Empirical Efficiency** Extensive experiments demonstrate that MINER achieves significant gains in both sample efficiency and accuracy over all competitive baselines. With zero additional inference overhead, these results validate that latent uncertainty exploitation is a sufficient and necessary component for scalable RLVR training.

2 Preliminaries

Reinforcement learning with Verifiable Rewards (RLVR) The RL objective for the policy π_θ is to maximize the cumulative rewards r received from

the verifier. Specifically, Policy Gradient (Williams, 1992) gives the following objective function:

$$\nabla \mathcal{J}_\theta = \mathbb{E}_{q \sim \mathcal{D}, \mathbf{o} \sim \pi_\theta(q)} \sum_{j=0}^T \nabla_\theta \log \pi_\theta(o_j | \mathbf{o}_{<j}) A_j, \quad (1)$$

where \mathcal{D} is the training distribution, q is an input prompt, \mathbf{o} is an output sequence consisting of T tokens $\{o_1, o_2, \dots, o_T\}$, and A_j is the advantage of the j -th token given the state $\mathbf{o}_{<j}$. Recently, DeepSeek-R1 (Guo et al., 2025) boosted large language models’ reasoning ability via the Group Relative Policy Optimization (GRPO; Shao et al. (2024)) algorithm. Each rollout is labeled by a verifiable reward $r(\cdot)$ which assigns 1 for correctness and 0 otherwise, and its advantage is estimated using the group average and standard deviation values of rewards from a group of G trajectories $\mathcal{O} = \{\mathbf{o}_i\}_{i=1}^G$ generated based on the same q :

$$A_i = \frac{r(\mathbf{o}_i) - \text{mean}(r(\mathbf{o}_1), \dots, r(\mathbf{o}_G))}{\text{std}(r(\mathbf{o}_1), \dots, r(\mathbf{o}_G))}. \quad (2)$$

GRPO optimizes the policy using the PPO objective (Schulman et al., 2017):

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q \sim \mathcal{D}, \{\mathbf{o}_i\}_{i=1}^G \sim \pi_\theta(\cdot|q)} \left[\sum_{j=1}^{|\mathbf{o}_i|} \min \left(\rho_{i,j} A_i, \text{clip}(\rho_{i,j}, 1 - \epsilon, 1 + \epsilon) A_i \right) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \right], \quad (3)$$

where $\rho_{i,j} = \frac{\pi_\theta(o_{i,j} | \mathbf{o}_{i,<j,q})}{\pi_{\text{old}}(o_{i,j} | \mathbf{o}_{i,<j,q})}$ is the importance sampling ratio, $|\mathbf{o}_i|$ is the sequence length and KL divergence (Kullback and Leibler, 1951) $D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$, serves as a regularizer that encourages the policy π_θ to remain close to the reference policy π_{ref} in distributional space.

Data Efficiency Under the definition of RLVR, we simplify the reward function $r(\cdot)$ as a binary indicator, i.e., its value equals 1 for correct rollouts and 0 otherwise. Under this setting, we classify a prompt q into three categories, i.e., positive homogeneous (PH), negative homogeneous (NH), and heterogeneous (HE), based on the correctness of its G rollouts $\{\mathbf{o}_i\}_{i=1}^G$:

$$q := \begin{cases} q_{ph} & \text{if } \sum_{i=1}^G r(\mathbf{o}_i) = G \\ q_{he} & \text{if } \sum_{i=1}^G r(\mathbf{o}_i) \in (0, G) \\ q_{nh} & \text{if } \sum_{i=1}^G r(\mathbf{o}_i) = 0 \end{cases} \quad (4)$$

Under this definition, it is observed that rollouts would receive zero advantage when generated from PH and NH groups. Given the rapidly evolving rate of LLMs (Kaplan et al., 2020; Xiao et al., 2025), the PH groups increasingly dominate the training batch for future base models, thereby causing substantial useless rollout costs.

Addressing the Diminishing Advantage Issue

Extensive research has explored how to mitigate NH prompts by reducing prompt difficulty, e.g., appending hints (Liu et al., 2025b), incorporating in-context demonstrations (Bamba et al., 2025), or injecting replay buffers (Sun et al., 2025; Jiang et al., 2025). We view these NH-oriented techniques as largely complementary to our goal and thus focus on PH prompts in this work. In contrast, despite the increasing prevalence of PH groups in training batches as base model capabilities rapidly improve (Kaplan et al., 2020; Xiao et al., 2025), strategies for efficiently leveraging positive homogeneous responses remain under-explored. Existing attempts are often suboptimal: DAPO (Yu et al., 2025) adopts over-sampling and filtering, which still wastes rollouts; other approaches introduce denser rewards via implicit process reward models (Yuan et al., 2025; Fei et al., 2025) or cooperation with strong reward models (Tao et al., 2025), but they typically require SFT-tuned models or incur substantial computational overhead, limiting their use in zero-RL and large-scale settings. Therefore, we aim to transform PH responses into heterogeneous ones by leveraging uncertainty-based intrinsic rewards, without incurring additional rollouts or relying on large learned reward models.

3 MINER

We introduce MINER, a data-efficient RLVR framework that *recovers training signals from Positive Homogeneous (PH) prompts*, to mine intrinsic mastery. In standard GRPO, PH prompts yield zero advantage and are often filtered out, despite consuming the same rollout budget. Our key hypothesis is that, although PH rollouts are *equally correct* under the verifier, they are not *equally mastered* by the policy. MINER converts the policy’s intrinsic uncertainty into a dense *mastery* signal, enabling online hard-positive mining to consolidate weak but correct reasoning modes.

MINER consists of three components (Fig. 2): (i) **Uncertainty-Driven Intrinsic Rewards** to select *which* correct trajectories remain weakly mastered

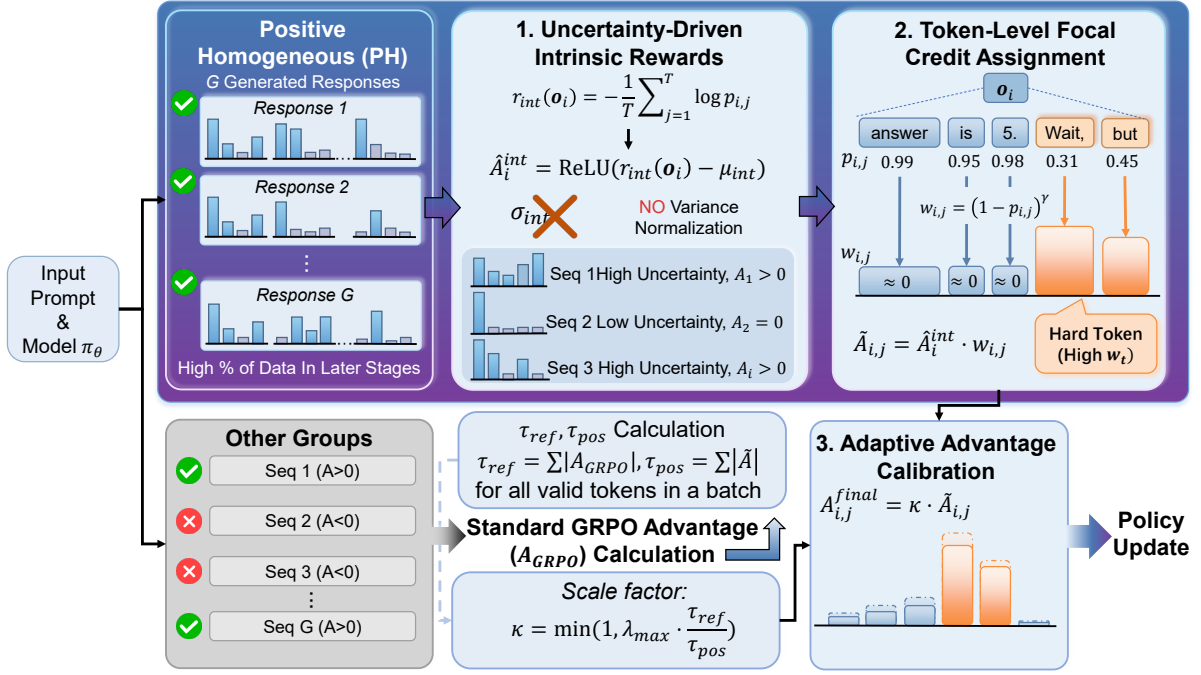


Figure 2: Framework of MINER. We focus on introducing intrinsic rewards to positive homogeneous prompts (PH). **Upper Center:** We use sequence-level uncertainty computed via the old policy π_{old} as the intrinsic rewards, to reinforce correct yet uncertain rollouts, without overfitting to already-mastered sequences; **Upper Right:** Then, we leverage token-level focal credit assignment to specifically rewarding critical tokens, again skipping self-confident tokens; **Lower Right:** Finally, to balance the learning signals from two groups, we calibrate the advantage score to a predefined threshold, significantly enhancing data efficiency without disturbing normal learning progress.

(§3.1); (ii) **Token-Level Focal Credit Assignment** to localize *where* the bottleneck steps are within a trajectory (§3.2); (iii) **Adaptive Advantage Calibration** to ensure the intrinsic mastery signal never overrides the extrinsic verification objective (§3.3).

3.1 Uncertainty-Driven Intrinsic Rewards

For a PH prompt q_{ph} , all rollouts receive identical extrinsic reward $r(o_i) \equiv 1$, providing no learning direction. We therefore define an intrinsic reward that reflects a lack of mastery. Concretely, we use the per-token negative log-likelihood (NLL):

$$r_{\text{int}}(o_i) = -\frac{1}{T_i} \sum_{j=1}^{T_i} \log \pi_{\text{old}}(o_{i,j} | o_{i,<j}, q). \quad (5)$$

A correct response with a high NLL is treated as less mastered and thus more valuable to reinforce. Importantly, we apply this intrinsic signal *only* to PH prompts, avoiding interference with heterogeneous prompts where the verifier-defined reward is distinguishable. We then compute a centered intrinsic advantage using a group-mean baseline:

$$A_i^{\text{int}} = r_{\text{int}}(o_i) - \frac{1}{G} \sum_{k=1}^G r_{\text{int}}(o_k). \quad (6)$$

Unlike GRPO, we omit standard-deviation normalization to preserve absolute uncertainty gaps (mild vs severe), while relying on calibration (§3.3) to control the global scale. To avoid decreasing the probability of already well-mastered trajectories, we adopt **positive filtering**:

$$\hat{A}_i^{\text{int}} = \text{ReLU}(A_i^{\text{int}}), \quad (7)$$

which only pulls up under-confident yet correct modes and circumvents penalizing confident and correct trajectories.

3.2 Token-Level Focal Credit Assignment

Sequence-level advantage assigns identical credit to all tokens, whereas uncertainty in reasoning is often concentrated at a few bottleneck steps. We thus reweight token credits by a focal factor based on token probability $p_{i,j} = \pi_{\text{old}}(o_{i,j} | o_{i,<j}, q)$:

$$w_{i,j} = (1 - p_{i,j})^\gamma, \quad (8)$$

where $\gamma \geq 0$ controls focusing ($\gamma=2$ suggested by Lin et al. (2017)). We treat $w_{i,j}$ as a constant (stop-gradient) to preserve the GRPO update form. The token-level intrinsic advantage is:

$$\tilde{A}_{i,j} = w_{i,j} \cdot \hat{A}_i^{\text{int}}, \quad (9)$$

Model	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB.		Avg.	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
<i>Base Model: Qwen3-4B-Base</i>														
Base	9.51	50.00	6.65	60.00	32.91	97.50	0.99	26.67	58.89	91.00	31.81	65.40	23.46	65.10
GRPO	19.79	63.33	20.34	63.33	61.89	97.50	7.86	33.33	83.71	94.00	52.19	72.63	40.97	70.69
DAPO	21.22	53.33	18.70	50.00	64.51	95.00	6.72	30.00	82.11	94.40	51.17	71.26	40.74	65.67
R++	15.55	56.67	12.68	43.33	58.71	95.00	4.19	36.67	79.90	93.40	46.46	70.40	36.25	65.91
GSPO	16.04	70.00	11.98	46.67	58.09	97.50	4.53	30.00	80.25	93.40	46.21	69.88	36.38	67.91
UCAS	19.09	56.67	15.42	53.33	60.14	100.00	9.19	36.67	82.80	94.00	49.46	70.05	39.35	68.45
MINER	25.86	73.33	22.97	60.00	69.65	97.50	10.81	46.67	86.93	95.80	57.07	76.25	45.55	74.92
<i>Base Model: Qwen3-8B-Base</i>														
Base	9.17	53.33	5.91	40.00	39.14	95.00	1.51	23.33	56.31	91.40	34.51	66.09	24.43	61.53
GRPO	23.25	66.67	19.40	50.00	73.85	97.50	8.36	36.67	87.15	95.40	56.20	76.08	44.70	70.39
DAPO	25.26	70.00	18.02	60.00	67.79	97.50	12.32	46.67	87.74	98.20	56.42	76.94	44.59	74.89
R++	23.67	70.00	20.42	56.67	72.81	97.50	9.27	36.67	87.86	97.00	56.12	77.45	45.03	72.55
GSPO	25.00	70.00	19.68	53.33	70.27	95.00	9.16	46.67	87.56	95.40	56.21	75.90	44.65	72.72
UCAS	22.16	70.00	15.34	53.33	66.00	100.00	8.88	50.00	84.54	94.80	53.36	74.87	41.71	73.83
MINER	27.81	70.00	23.98	66.66	70.41	100.00	12.92	50.00	88.56	97.40	58.75	78.14	47.07	77.03

Table 1: Comprehensive comparison against other critic-free RL algorithms in terms of Pass@1 (P@1) and Pass@K (P@K) scores. ‘‘OlympiadB.’’ refers to the OlympiadBench. Best performance is highlighted with **bold**.

which prioritizes bottleneck tokens and avoids spending gradient budget on trivial connectors. MINER introduces no additional rollouts and incurs negligible overhead, as it reuses token log-probabilities already computed by Eq. (3).

3.3 Adaptive Advantage Calibration

Intrinsic mastery rewards and extrinsic verification rewards have different scales. To respect the signal hierarchy, we cap the intrinsic advantage by a reference signal extracted from HE prompts in the same batch. Let \mathcal{B}_{he} and \mathcal{B}_{ph} denote HE and PH prompts in the batch, respectively. We define

$$\begin{aligned} \tau_{\text{ref}} &= \sum_{q \in \mathcal{B}_{he}, i \in \{1..G\}} \sum_{j=1}^{|\mathcal{o}_i^q|} |A_{i,j}^q| \\ \tau_{\text{pos}} &= \sum_{q \in \mathcal{B}_{ph}, i \in \{1..G\}} \sum_{j=1}^{|\mathcal{o}_i^q|} |\tilde{A}_{i,j}^q|, \end{aligned} \quad (10)$$

and use a scale factor λ_{max} to guarantee that the intensity of additional advantages τ_{pos} never exceed the configured signal threshold $\lambda_{\text{max}} \cdot \tau_{\text{ref}}$:

$$\mathcal{A}_{i,j}^{\text{final}} = \tilde{A}_{i,j} \cdot \min \left(1, \frac{\lambda_{\text{max}} \cdot \tau_{\text{ref}}}{\tau_{\text{pos}}} \right) \quad (11)$$

Finally, we optimize Eq. (3) using $\mathcal{A}_{i,j}^{\text{final}}$ for PH prompts and standard GRPO advantages for others.

4 Experiments

4.1 Experiment Setups

Evaluation We adopt MATH500 (Lightman et al., 2023), AMC23 (AI-MO, 2024), OlympiadBench (He et al., 2024), AIME2024 (Mathematical Association of America, 2025a) and AIME2025 (Mathematical Association of America, 2025b), HMMT25 (Balunović et al., 2025) as the evaluation testbeds with diverse complexity. Apart from **GRPO**, we choose **DAPO** (Yu et al., 2025), **GSPO** (Zheng et al., 2025a) and **REINFORCE++** (R++; Hu et al. (2025)) as baselines. To illustrate exploration superiority, we also include UCAS (Xie et al., 2025) as a strong baseline. We set the temperature as 0.7, top_p as 0.95, and use a maximum token limit of 8192. We conduct 128 rollouts for AIME2024, AIME2025, AMC23 and HMMT25, and 16 rollouts for OlympiadBench and MATH500. We adopt Pass@1 and Pass@K (Chen et al., 2021) as evaluation metrics to measure the exploitation and exploration abilities.

Training We adopt DeepScaleR (Luo et al., 2025) as the training set and choose Qwen3-4B-Base (Qwen, 2025) and Qwen3-8B-Base as the base policy. We use veRL (Sheng et al., 2025) as the training framework. We set λ_{max} to $1.5e - 3$ for both models by grid search (elucidated in Appendix D.4). Additional hyperparameters for RL training are presented in Table 4.

4.2 Main Results

As illustrated in Table 1, MINER yields a **4.58** point increase in the Pass@1 metric and a consistent **4.23** point improvement in Pass@K on the Qwen3-4B model. These concurrent gains in both exploitation and exploration performance validate the comprehensive effectiveness of our approach. When applied to the more robust Qwen3-8B backbone, MINER continues to effectively leverage PH problems. Compared to GRPO, the Pass@K improvement rises to 6.66, accompanied by a notable 2.37 gain in Pass@1. Notably, MINER outperforms the strong DAPO baseline—even though DAPO benefits from oracle outcome verifier signals—achieving a +2.48 Pass@1 lead and a 2.14 Pass@K margin. This suggests that the intrinsic rewards for PH prompts are as informative as standard outcome rewards, highlighting a new path for scaling RLVR with diverse data. Furthermore, while hyperparameters were optimized on the 4B model, their seamless transfer to the 8B variant underscores MINER’s robustness and data efficiency when scaling to larger architectures.

4.3 Ablation Study

In this section, we use Qwen3-4B to ablate our method with three variants: (1) MINER without intrinsic reward (w/o IR), which uses a fixed advantage score (0.05) for all positive homogeneous rollouts; (2) MINER without focal weight (w/o FW), which uses a uniform weight 1 for tokens throughout a trajectory; (3) MINER without advantage calibration (w/o AC), which allows for uncapped advantage signals. Results in Fig. 4a and Table 5 demonstrate that rewarding positive homogeneous rollouts with a fixed, undistinguishable advantage is harmful for stable training, verifying that the improvements do come from the beneficial intrinsic rewards. Moreover, focal weighting is extremely useful for improving models’ exploration ability while simultaneously guaranteeing sharpening mastered knowledge, resulting in much higher Pass@K performance compared to the w/o FW variant. Finally, without advantage calibration, the training fails in the middle and suffers from the under-fitting problem, which implies that a simple grid-search on λ_{\max} is sufficient for MINER to achieve fast convergence and stable training simultaneously.

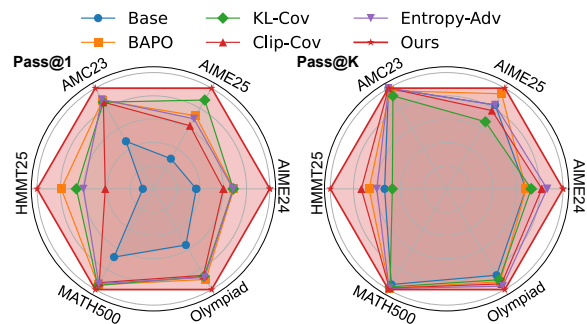


Figure 3: Comparison with other exploration-enhanced algorithms on normalized Pass@1 and Pass@K scores.

5 Analysis

In this section, we discuss the following research questions (RQ) of the MINER algorithm:

- RQ1:** Can MINER mitigate data inefficiency when altering the model backbone and domain task?
- RQ2:** What is the relationship between MINER and entropy-maximization methods?
- RQ3:** Can MINER benefit from larger token budgets?
- RQ4:** How does MINER conduct test-time scaling?
- RQ5:** Can MINER break the RL dilemma to incentivize beyond base capabilities?
- RQ6:** Can MINER stabilize training to further steps and larger base models?

Response to RQ1: MINER is domain- and backbone-agnostic, delivering consistent gains across diverse tasks and architectures. Given replication challenges for RL on Llama-family models in math reasoning (Appendix D.6, (Gandhi et al., 2025; Liu et al., 2025a)), we evaluate MINER on medical reasoning, which is a practical domain feasible for scalable RL training. We test on MedQA (Jin et al., 2021), MedMCQA (Pal et al., 2022), PubMedQA (Jin et al., 2019), MedXpertQA (Zuo et al., 2025), and MMLU-Pro medical subsets (Wang et al., 2024), using an 85K training corpus sub-sampled from MedQA/MedMCQA training partitions (strictly held out; Appendix C.7). With 4 samples per question, statistical variance is minimized across these large-scale benchmarks. To verify data efficiency, we compare against medical-specialized models HuatuoGPT-o1 (Chen et al., 2025) and MedReason (Wu et al., 2025), which require complex data processing and GPT-4o distillation despite sharing the Llama3.1-8B-Instruct backbone (Dubey et al., 2024). Table 2 shows MINER

Model	MedQA		MedMCQA		PubMedQA		MedXpertQA		MMLU-Pro		Avg	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
Llama3.1-8B-Instruct	60.96	81.85	55.41	77.07	75.55	81.20	11.46	25.92	54.95	76.09	51.67	68.43
+GRPO	70.15	84.45	63.64	76.12	77.20	82.00	16.29	29.35	65.57	77.39	58.57	69.86
+HuatuoGPT-o1*	70.20	-	58.20	-	76.10	-	17.30	-	59.90	-	56.34	-
+MedReason*	68.40	-	57.50	-	77.60	-	16.40	-	63.10	-	56.60	-
+MINER	72.03	86.17	64.59	78.29	78.95	82.40	17.53	32.69	66.82	78.96	59.96	71.70

Table 2: Comparison of MINER with other data-centric methods, which outperforms the two representative medical models with consistent gains. “*” denotes the results are sourced from the original paper of Wu et al. (2025).

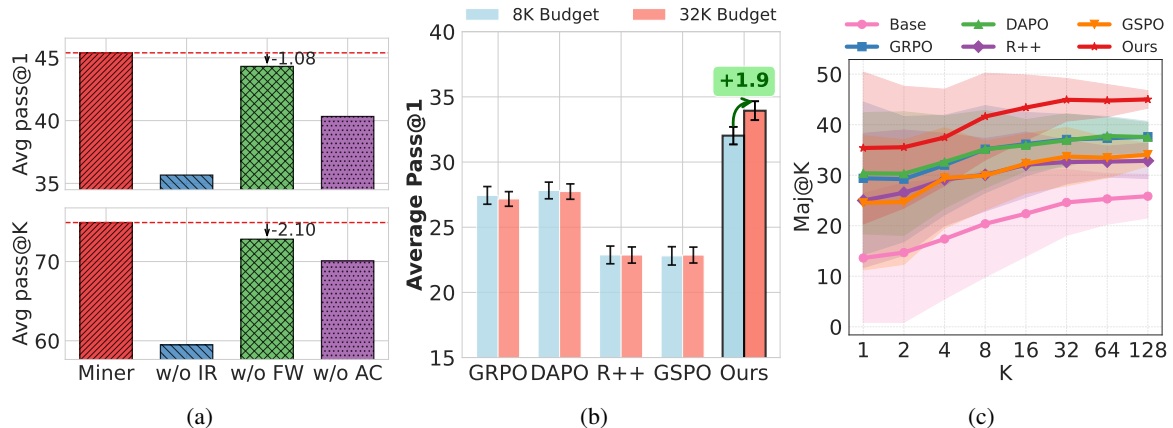


Figure 4: (a) Ablation study with three innovations (Intrinsic Reward (IR), Focal Weighting (FW) and Advantage Calibration (AC)) of MINER on the Qwen3-4B base model. (b) Performance dynamics given sufficient inference budgets. Apart from fluctuation within the error bar, MINER achieves non-negligible and sound improvements. (c) Parallel test-time scaling comparison with other algorithms, where MINER consistently outperforms other baselines with over 5 absolute points. Shaded areas denote ± 1 standard deviation over 10 runs.

surpasses GRPO and both data-intensive baselines across all five medical tasks of varying difficulty. Crucially, training logs (Fig. 11) confirm MINER achieves stable improvement even with $>65\%$ PH prompts per batch. This demonstrates MINER’s dual generalization capability: consistent performance gains across model architectures (backbone-agnostic) and domains (domain-agnostic), particularly valuable in data-scarce fields like medicine.

Response to RQ2: MINER improves the exploration via a healthier manner. We additionally compare MINER with other exploration-enhancement methods that manipulate advantage signals on heterogeneous prompts, to unveil that operating on positive homogeneous prompts results in healthier improvement on exploration. We build on Qwen3-4B and compare with BAPO (Xi et al., 2025), KL-Cov and Clip-Cov (Cui et al., 2025), which enhance exploration by softening the upper clipping bound, as well as Entropy-Adv (Cheng et al., 2025), which explicitly rewards high-entropy tokens by shaping the advantage estimation with the actor entropy value. We do not compare with

entropy regularization, which will result in unbounded entropy collapse for inappropriate setups of the hyperparameter α (Cui et al., 2025). We use the suggested hyperparameters released at their official codebase (details in Appendix C.6). We show the normalized performance in Fig. 3 and full results in Table 16, where these methods could not generalize as perfectly as in their original paper on both metrics. With only one hyperparameter and stable control of shaped advantage estimations, MINER surpasses them by a large margin.

Response to RQ3: Yes. MINER improves with a larger inference budget. To validate whether the intrinsic rewards on positive homogeneous groups would both improve the model’s performance under a less-constrained token limit (Snell et al., 2025; Muennighoff et al., 2025), we evaluate by extending the inference budget to the maximum context limit of 32K. We choose Qwen3-4B as the base backbone and test on four challenging datasets (AIME24, AIME25, AMC23, and HMMT25), as they pose high demands for testing budget (Guo et al., 2025). We plot the comparison in Fig. 4b.

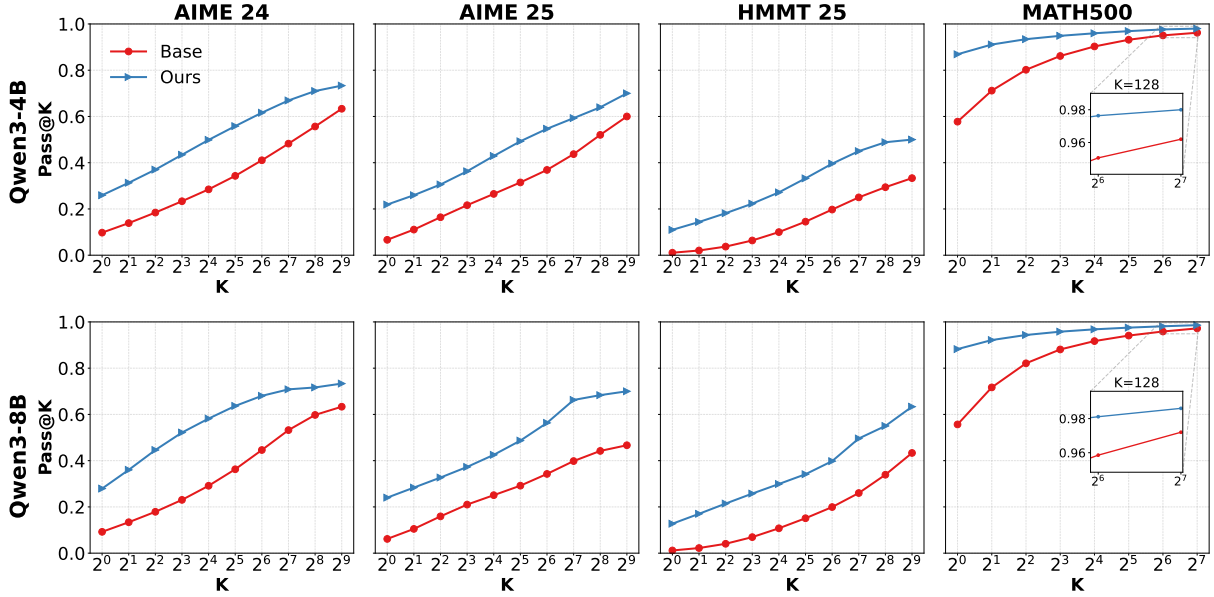


Figure 5: Pass@K scaling of MINER and Base model on Qwen3-4B (*Upper*) and Qwen3-8B (*Lower*) models, where MINER still demonstrates improvements for a sufficiently large K .

We observe that the performance of other methods fluctuates within the margin of error, demonstrating that they fail to achieve consistent gains from increased test-time compute budgets. In contrast, our method exhibits a stable and statistically significant improvement (+1.9 pass@1), indicating that our algorithm not only enhances data efficiency during training but also substantially boosts the model’s scaling potentials during deployment.

Response to RQ4: MINER could consistently improve with more parallel test-time compute.

To illustrate the test-time scaling (Zhang et al., 2025) potential of MINER against other baselines, we adopt self-consistency (SC; Wang et al. (2023)) as the evaluation method under multiple parallel samples. We do not compare sequential scaling as it is empirically verified to be highly inefficient compared to the parallel scaling paradigm (Ghosal et al., 2025). Fig. 4c demonstrates the substantial scaling potentials of MINER compared to other algorithms with 10 repetitions. Most methods suffer from performance staleness, while MINER improves consistently given more samples, surpassing other baselines by 7.38 points. Detailed figures are presented at Table [12,13,14,15].

Response to RQ5: When sampling up to 512 trajectories, the answer is ‘yes’. Yue et al. (2025) reveals that previous RLVR algorithms, e.g., GRPO or DAPO, primarily sharpen the policy distribution but often sacrifice the potential for discovering opti-

mal solutions when given an ample number of trials. In addressing this research question, we rigorously investigate whether MINER can effectively mitigate the mode collapse issue under a sampling regime of 512 trajectories, a threshold considered sufficiently large to comprehensively unveil the model’s underlying reasoning capabilities and behavioral patterns. To ensure a thorough evaluation, we test the models on highly challenging benchmarks such as AIME24, AIME25, and HMMT25; additionally, we include the comparatively accessible MATH500 dataset with $K = 128$ to provide a balanced assessment across difficulty levels. As illustrated in Fig. 5, MINER consistently exceeds the base model’s performance even at large K values, with results derived from both Qwen3-4B and Qwen3-8B architectures. Notably, our approach exhibits a consistent and robust growth trajectory on demanding benchmarks like AIME25 and HMMT25, with no signs of performance plateauing or saturation, thereby indicating sustained exploratory capacity. In essence, MINER substantially alleviates the mode collapse inherent in standard GRPO, achieving an optimal equilibrium between exploration and exploitation that enhances overall solution diversity and reliability.

Response to RQ6: Yes, MINER is consistently stable.

To validate the stability of MINER, we first scale it to Qwen3-14B-Base with the same λ_{\max} value and compare it with GRPO. Experi-

Table 3: Scaling of MINER to a larger Qwen3-14B-base model, with consistent improvement against GRPO.

Model	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB		Avg	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
Qwen3-14B-Base	9.77	53.33	5.55	43.33	37.85	97.50	1.95	26.67	53.22	91.20	34.37	68.16	23.79	63.37
+GRPO	23.15	70.00	20.65	60.00	72.48	97.50	9.32	40.00	87.90	95.80	55.49	76.59	44.83	73.32
+MINER	32.71	73.33	24.90	66.67	73.69	100.00	13.85	53.33	88.90	98.00	59.56	79.69	48.94	78.50

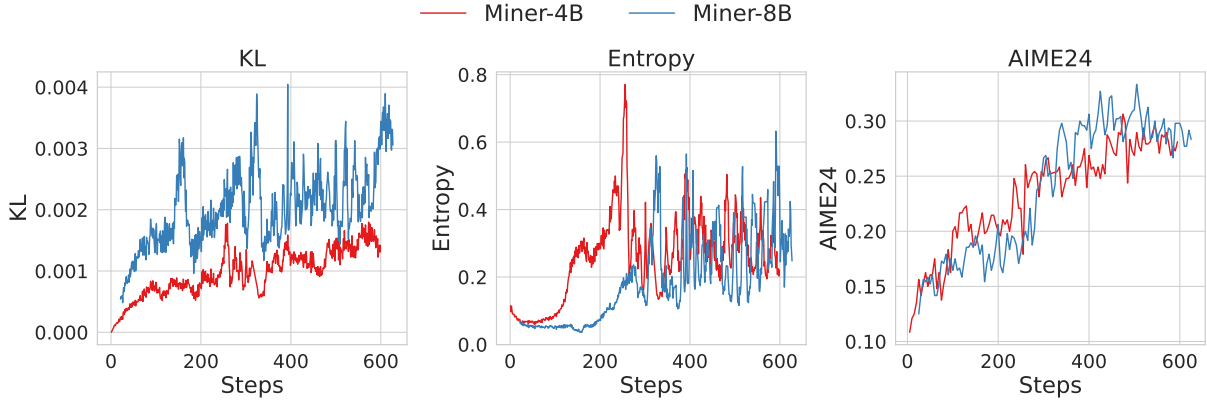


Figure 6: KL, Entropy and AIME24 scores when extending RL training to another 1 epoch on Qwen3-4B and Qwen3-8B backbones.

ment results in Table 3 demonstrate that MINER is still superior to GRPO, where the improvements on Pass@1 and Pass@K remain consistent at 4.11 and 5.18, respectively. This again validates the generality of our method and no requirements for dedicated hyperparameter tuning. Besides capacity scaling, we also extend the training process to another training epoch. We record stability metrics, including KL and entropy, and pass@1 score on the selected dev set, AIME24 and plot the dynamics in Fig. 6. Results reveal that although in the initial epoch the training dynamics fluctuate, both KL and entropy remain stable and the AIME24 score is consistently increasing, which cross-verifies the generalization of MINER and its adaptability in prolonged RL training.

6 Conclusion

We present MINER, a novel reinforcement learning framework that transforms previously wasted positive homogeneous (PH) rollouts into valuable learning signals through uncertainty-driven intrinsic rewards. By introducing sequence-level uncertainty rewards with positive filtering, token-level focal credit assignment, and adaptive advantage calibration, our method effectively converts gradient-desert PH groups into catalysts for knowledge consolidation. Extensive experiments demonstrate that MINER boosts both pass@1 and pass@K, without

additional computation or excessive use of hyperparameters. By turning “solved” prompts into robustness incubators, MINER paves the way for efficient RL training where every rollout counts.

Limitations

This work focuses on unlocking the learning signal from positive homogeneous (PH) prompts, and we validate MINER across three model scales (Qwen3-4B, Qwen3-8B and Qwen3-14B) and a diverse suite of reasoning benchmarks. We did not further scale training to substantially larger backbones (e.g., 32B) due to computational constraints. Nevertheless, MINER introduces negligible additional overhead and only a single new hyperparameter, λ_{\max} ; importantly, the same setting transfers smoothly from 4B to 8B in our experiments, suggesting that scaling primarily requires additional compute rather than methodological changes.

Ethical Considerations

Reinforcement learning from verifiable rewards has become a major part of bootstrapping large language models’ intelligence in the data-scarce world. While effective, a large proportion of rollout data is useless for training, which brings severe training inefficiency. Our work aims to offer a practical solution to resolve the data inefficiency problem

via a robust and computationally friendly manner, fostering innovation and collaboration to accelerate advancements that ultimately benefit society.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576209) and STCSM (No. 2025SHZDZX025G05).

References

- AI-MO. 2024. [Amc23 dataset](#). Hugging Face Dataset Repository. Accessed: 2025-06-26.
- Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. 2025. [Matharena: Evaluating llms on uncontaminated math competitions](#).
- Udbhav Bamba, Minghao Fang, Yifan Yu, Haizhong Zheng, and Fan Lai. 2025. Xrpo: Pushing the limits of grpo with targeted exploration and exploitation. *arXiv preprint arXiv:2510.06672*.
- Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, and Benyou Wang. 2025. [Towards medical complex reasoning with LLMs through medical verifiable problems](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14552–14573, Vienna, Austria. Association for Computational Linguistics.
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, and 1 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.
- Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with exploration: An entropy perspective. *arXiv preprint arXiv:2506.14758*.
- Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. 2025. [The Entropy Mechanism of Reinforcement Learning for Reasoning Language Models](#). *Preprint*, arXiv:2505.22617.
- Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenefeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025. Beyond binary rewards: Training llms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Wu Fei, Hao Kong, Shuxian Liang, Yang Lin, Yibo Yang, Jing Tang, Lei Chen, and Xiansheng Hua. 2025. Self-guided process reward optimization with masked step advantage for process reinforcement learning. *arXiv preprint arXiv:2507.01551*.
- Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D Goodman. 2025. Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. *arXiv preprint arXiv:2503.01307*.
- Soumya Suvra Ghosal, Souradip Chakraborty, Avinash Reddy, Yifu Lu, Mengdi Wang, Dinesh Manocha, Furong Huang, Mohammad Ghavamzadeh, and Amrit Singh Bedi. 2025. [Does Thinking More always Help? Understanding Test-Time Scaling in Reasoning Models](#). *Preprint*, arXiv:2506.04210.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.
- Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. [OlympiadBench: 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)*, pages 3828–3850, Bangkok, Thailand. Association for Computational Linguistics.
- Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. 2025. Reinforce++: An efficient rlhf algorithm with robustness to both prompt and reward models. *arXiv preprint arXiv:2501.03262*.
- Guochao Jiang, Wenfeng Feng, Guofeng Quan, Chuzhan Hao, Yuwei Zhang, Guohua Liu, and Hao Wang. 2025. Vcrl: Variance-based curriculum reinforcement learning for large language models. *arXiv preprint arXiv:2509.19803*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. [PubMedQA: A dataset for biomedical research question answering](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020.

- Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let’s verify step by step. In *The Twelfth International Conference on Learning Representations*.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025a. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.
- Ziru Liu, Cheng Gong, Xinyu Fu, Yaofang Liu, Ran Chen, Shoubo Hu, Suiyun Zhang, Rui Liu, Qingfu Zhang, and Dandan Tu. 2025b. Ghpo: Adaptive guidance for stable and efficient llm reinforcement learning. *arXiv preprint arXiv:2507.10628*.
- Michael Luo, Sijun Tan, Justin Wong, Xiaoxiang Shi, William Y. Tang, Manan Roongta, Colin Cai, Jeffrey Luo, Li Erran Li, Raluca Ada Popa, and Ion Stoica. 2025. DeepScaler: Surpassing o1-preview with a 1.5b model by scaling rl. <https://pretty-radio-b75.notion.site/DeepScaleR-Surpassing-O1-Preview-with-a-1-5B-Model-by-Scaling-RL-19681902c\1468005bed8ca303013a4e2>. Notion Blog.
- Mathematical Association of America. 2025a. [Aime 2024 dataset](#). Hugging Face Dataset Repository. Accessed: 2025-06-26.
- Mathematical Association of America. 2025b. [Aime 2025 dataset](#). Hugging Face Dataset Repository. Accessed: 2025-06-26.
- Kaixiang Mo, Yuxin Shi, Weiwei Weng, Zhiqiang Zhou, Shuman Liu, Haibo Zhang, and Anxiang Zeng. 2025. Mid-training of large language models: A survey. *arXiv preprint arXiv:2510.06826*.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candes, and Tatsunori Hashimoto. 2025. [s1: Simple test-time scaling](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, Suzhou, China. Association for Computational Linguistics.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.
- Team Qwen. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.
- Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.
- Paul Stangel, David Bani-Harouni, Chantal Pellegrini, Ege Özsoy, Kamelia Zaripova, Matthias Keicher, and Nassir Navab. 2025. Rewarding doubt: A reinforcement learning approach to calibrated confidence expression of large language models. *arXiv preprint arXiv:2503.02623*.
- Yifan Sun, Jingyan Shen, Yibin Wang, Tianyu Chen, Zhendong Wang, Mingyuan Zhou, and Huan Zhang. 2025. Improving data efficiency for LLM reinforcement fine-tuning through difficulty-targeted online data selection and rollout replay. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Leitian Tao, Ilia Kulikov, Swarnadeep Saha, Tianlu Wang, Jing Xu, Sharon Li, Jason E Weston, and Ping Yu. 2025. Hybrid reinforcement: When reward is sparse, it’s better to be dense. *arXiv preprint arXiv:2510.07242*.
- Jinpeng Wang, Chao Li, Ting Ye, Mengyuan Zhang, Wei Liu, and Jian Luan. 2025a. Icpo: Intrinsic confidence-driven group relative preference optimization for efficient reinforcement learning. *arXiv preprint arXiv:2511.21005*.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. [Self-consistency improves chain of thought reasoning in language models](#). In *The Eleventh International Conference on Learning Representations*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang

- Yue, and Wenhu Chen. 2024. [MMLU-pro: A more robust and challenging multi-task language understanding benchmark](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. 2025b. [Octothinker: Mid-training incentivizes reinforcement learning scaling](#). *arXiv preprint arXiv:2506.20512*.
- Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.
- Juncheng Wu, Wenlong Deng, Xingxuan Li, Sheng Liu, Taomian Mi, Yifan Peng, Ziyang Xu, Yi Liu, Hyunjin Cho, Chang-In Choi, Yihan Cao, Hui Ren, Xiang Li, Xiaoxiao Li, and Yuyin Zhou. 2025. [Medreason: Eliciting factual medical reasoning steps in llms via knowledge graphs](#). *Preprint*, arXiv:2504.00993.
- Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, and 1 others. 2025. [Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping](#). *arXiv preprint arXiv:2510.18927*.
- Chaojun Xiao, Jie Cai, Weilin Zhao, Biyuan Lin, Guoyang Zeng, Jie Zhou, Zhi Zheng, Xu Han, Zhiyuan Liu, and Maosong Sun. 2025. [Densing law of llms](#). *Nature Machine Intelligence*, pages 1–11.
- Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. 2025. [Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning](#). *arXiv preprint arXiv:2510.10649*.
- Yixuan Even Xu, Yash Savani, Fei Fang, and J Zico Kolter. 2025. [Not all rollouts are useful: Down-sampling rollouts in llm reinforcement learning](#). *arXiv preprint arXiv:2504.13818*.
- Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, and 16 others. 2025. [DAPO: An Open-Source LLM Reinforcement Learning System at Scale](#). *Preprint*, arXiv:2503.14476.
- Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2025. [Free process rewards without process labels](#). In *Forty-second International Conference on Machine Learning*.
- Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. 2025. [Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?](#) In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, and 1 others. 2025. [A survey on test-time scaling in large language models: What, how, where, and how well?](#) *arXiv preprint arXiv:2503.24235*.
- Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, and 1 others. 2025a. [Group sequence policy optimization](#). *arXiv preprint arXiv:2507.18071*.
- Haizhong Zheng, Yang Zhou, Brian R Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. 2025b. [Act only when it pays: Efficient reinforcement learning for llm reasoning via selective rollouts](#). *arXiv preprint arXiv:2506.02177*.
- Yuzhen Zhou, Jiajun Li, Yusheng Su, Gowtham Ramesh, Zilin Zhu, Xiang Long, Chenyang Zhao, Jin Pan, Xiaodong Yu, Ze Wang, and 1 others. 2025. [April: Active partial rollouts in reinforcement learning to tame long-tail generation](#). *arXiv preprint arXiv:2509.18521*.
- Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. 2025. [The surprising effectiveness of negative reinforcement in LLM reasoning](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. 2025. [MedxpertQA: Benchmarking expert-level medical reasoning and understanding](#). In *Forty-second International Conference on Machine Learning*.

A Reproduction List

In this section, we present a brief reproduction list to implement our method:

1. **Computational Cards:** We use 4xA100 GPUs to train the Qwen3-4B-Base and Llama3.1-8B-Instruct for 4 days. We use 4xA100 GPUs to train the OctoThinker-8B-Hybrid-Base for 7 days. We use 8xA100 GPUs to train the Qwen3-8B-Base for 3 days.
2. **Code:** We attach the implementation code in the supplementary materials.
3. **Data:** All the dataset is officially available through their released links.

B Related Work

This appendix complements the preliminary discussion in §2 by positioning our study in the broader

Table 4: Hyperparameters for MINER training.

Hyperparameter	Qwen3-4B/8B-Base	Llama3.1-8B-Instruct
max response length	8192	8192
batch size	128	128
rollout batch size	128	128
learning rate	2.0e-06	1.0e-06
total training epochs	1	1
rollout number	16	16
PPO clip range ϵ	0.2	0.2
KL coefficient β	0.001	0.001

landscape of (i) data-efficient policy optimization under sparse/binary outcome rewards, (ii) prior attempts to resolve the diminishing-advantage phenomenon induced by homogeneous rollout groups, and (iii) intrinsic-reward designs based on model uncertainty.

Critic-Free Policy Optimization and Group-Based Advantages Policy-gradient methods (Williams, 1992) constitute the foundation of RL-based fine-tuning, while PPO (Schulman et al., 2017) stabilizes updates via clipped importance ratios and KL regularization (Kullback and Leibler, 1951). GRPO (Shao et al., 2024) adapts PPO-style updates to a group sampling scheme by estimating advantages from the relative reward statistics within a set of rollouts generated from the same prompt. This design eliminates a learned value critic and is thus memory-friendly for large-scale LLM training. However, when all rollouts in a group share the same verifiable reward (all-correct or all-wrong), group-relative normalization yields vanishing advantages, directly linking optimization progress to the diversity of outcome rewards.

Data Efficiency in RLVR and Rollout Waste A central challenge in RLVR is data efficiency: each prompt requires sampling multiple rollouts to construct a useful advantage baseline, and sparse outcome rewards can cause many rollouts to contribute negligible learning signals. As LLM capabilities improve (Kaplan et al., 2020; Xiao et al., 2025), an increasing fraction of prompts becomes trivially solvable, making *positive homogeneous* (PH) groups more frequent and amplifying useless rollout costs. DAPO (Yu et al., 2025) targets this inefficiency through over-sampling prompts and filtering zero-advantage groups, which can increase the effective gradient density but still spends compute on rollouts that are later discarded. Our work instead

aims to make the sampled rollouts themselves more informative under PH groups, thereby improving utilization without adding extra rollouts.

Mitigating Negative Homogeneous Prompts (NH) A rich literature has explored how to reduce prompt difficulty or reshape the training signal for *negative homogeneous* (NH) groups, where all sampled rollouts are incorrect. Representative strategies include appending hints to prompts (Liu et al., 2025b), adding in-context demonstrations (Bamba et al., 2025), and using replay buffers or replay-style mechanisms (Sun et al., 2025; Jiang et al., 2025). These approaches can convert homogeneous failures into heterogeneous outcomes by making at least some rollouts correct, yielding non-zero advantages. We view NH-oriented techniques as largely complementary to our focus: our method targets the increasingly dominant PH regime and can in principle be combined with NH mitigation when needed.

Leveraging Positive Homogeneous Prompts (PH) In contrast to NH, dedicated treatments for PH groups remain comparatively under-explored, despite their increasing prevalence in modern RLVR pipelines. Existing approaches typically fall into two categories. (1) **Sampling-based heuristics.** DAPO (Yu et al., 2025) filters out zero-advantage groups after over-sampling, improving effective batch quality at the expense of wasted rollouts. (2) **Denser supervision via learned reward signals.** Implicit process reward models (PRMs) (Yuan et al., 2025; Fei et al., 2025) attempt to provide step-wise or token-level guidance beyond the binary outcome reward, but often rely on an SFT-tuned model or additional training/maintenance costs, which can inhibit adoption in strict zero-RL or large-scale settings. Hybrid frameworks that cooperate with strong reward mod-

els (Tao et al., 2025) can further enhance feedback richness, but likewise introduce extra compute and system complexity. Our work targets PH groups while avoiding additional rollouts and dependence on large learned reward models.

Uncertainty Signals in RLVR Several recent RLVR studies incorporate model uncertainty as a lightweight intrinsic signal to enrich learning under sparse (often binary) verifiable rewards: (i) uncertainty-aware advantage shaping methods (Xie et al., 2025) modulate GRPO-style updates using confidence/uncertainty at the response and token levels to improve exploration and credit assignment; (ii) intrinsic confidence-driven variants (Wang et al., 2025a) turn relative confidence among multiple rollouts into a group-relative preference/advantage signal to augment the outcome reward signal; and (iii) calibration-oriented RLVR extensions (Damani et al., 2025; Stangel et al., 2025) augment correctness with proper scoring-rule-based rewards so the model learns to output calibrated confidence alongside answers. However, the first two paradigms overlook the advantage shaping of PH trajectories and fail to utilize them, while calibrated methods destroy the objective of the maximization of correctness, and achieve bad performance compared with pure RLVR baselines. Our algorithm, which calibrates only on PH prompts, operates in a basically orthogonal direction with these methods, and would result in a further superior RLVR method when complementing these algorithms.

C Experimental Details

C.1 Descriptions of Math Testbeds

We present the detailed description of the mathematical evaluation datasets as follows:

1. **AIME2024, AIME2025** (Mathematical Association of America, 2025a,b): These two datasets contain High school Olympiad-level assessment from American Invitational Mathematics Examination in 2024 and 2025. Each dataset contains 30 challenging problems covering Algebra/Geometry/Number theory.
2. **AMC23** (AI-MO, 2024): This dataset is sourced from American Mathematics Competitions system in 2023, which contains 40 problems with hybrid question types.

3. **OlympiadBench** (He et al., 2024): This dataset contains comprehensive math Olympiad problems from various nations. We only select the English version related to Math and keep the problems that require an answer with a number, leaving 581 problems for evaluation in total.
4. **MATH500** (Lightman et al., 2023): This dataset is an advanced mathematics evaluation set curated by OpenAI containing 500 problems with formal mathematical notations.
5. **HMMT25** (Balunović et al., 2025): The original questions were sourced from the HMMT February 2025 competition. 30 questions were extracted, converted to LaTeX and verified.

C.2 Descriptions of Medical Testbeds

We present the detailed description of the medical evaluation datasets as follows:

1. **MedQA** (Jin et al., 2021) is a widely used benchmark for evaluating AI systems in medical question answering, featuring multiple-choice questions from professional medical licensing exams such as the USMLE and exams from China and Taiwan. We adopt its 5-options English version, taking the 1,273 test problems as the evaluation benchmark.
2. **PubmedQA** (Jin et al., 2019) is a specialized benchmark for biomedical question answering, consisting of question-answer pairs derived from PubMed abstracts. It focuses on yes/no/maybe questions that require reasoning over biomedical literature. We use the human-labeled question test set, with 500 problems for evaluation. Note that we include relevant contexts before questions, challenging models' reasoning capability among contexts.
3. **MedMCQA** (Pal et al., 2022) is a large-scale benchmark for medical question answering, featuring over 194,000 multiple-choice questions sourced from Indian medical entrance exams and other educational resources. It spans a wide range of medical topics, including anatomy, pharmacology, and pathology, and is designed to evaluate the reasoning and knowledge application skills of AI systems in a clinical context. The test set contains 4,183 problems.
4. **MMLU-Pro** (Wang et al., 2024) is a challenging multi-task benchmark containing over

Model	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB.		Avg.	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
GRPO	19.79	63.33	20.34	63.33	61.89	97.50	7.86	33.33	83.71	94.00	52.19	72.63	40.97	70.69
MINER	25.86	73.33	22.97	60.00	69.65	97.50	10.81	46.67	86.93	95.80	57.07	76.25	45.55	74.92
w/o IR [†]	15.49	53.33	12.19	33.33	57.36	90.00	2.71	20.00	79.64	91.80	46.59	68.50	35.66	59.49
w/o FW	24.43	66.67	21.09	66.67	68.81	95.00	10.39	36.67	85.93	95.20	55.26	76.76	44.32	72.83
w/o AC [†]	20.23	66.67	19.48	53.33	60.55	97.50	7.89	33.33	82.95	96.00	50.87	73.67	40.33	70.08

Table 5: Ablation study on MINER. We compare (1) without intrinsic reward (w/o IR); (2) without focal weight (w/o FW); and (3) without advantage calibration (w/o AC) to unveil that each design of MINER are beneficial for simultaneously enhanced Pass@1 and Pass@K scores. The experiments marked with a † failed to complete training; we used the checkpoint saved before the crash for testing.

12,000 multiple-choice questions across 14 diverse domains, including subjects in STEM (e.g., math, physics, chemistry), social sciences, law, and humanities. We only maintain health and biology subsets for testing medical reasoning abilities, which includes 1535 problems.

- MedXpertQA** (Zuo et al., 2025) is an expert-level medical benchmark comprising 4,460 questions spanning 17 medical specialties and 11 body systems. It includes two subsets: a text-only version for evaluating textual medical reasoning and a multimodal version (MM) with images, aimed at assessing advanced clinical knowledge comparable to medical licensing exams. We only test models on the text-only subset, which contains 2450 problems.

C.3 Evaluation Prompts

For the mathematical reasoning tasks, we use prompts defined in Fig. 7 to start reasoning. For the medical reasoning tasks, We prompt the LRM with “Please reason step by step and output the final answer as ‘The answer is’ ” and extract the contents after ‘The answer is’ to **exact-match** the ground truth answer.

C.4 Computation of Metrics

Pass@K The pass@K (Chen et al., 2021) scores are computed as below:

$$\text{pass@K} = 1 - \frac{\binom{n-c}{K}}{\binom{n}{K}} \quad (12)$$

where n is the number of samples and c is the number of correct samples. When K is set to 1, this metric is reduced to the average accuracy among the n samples.

C.5 Details of Ablation Study

In this section, we in depth introduce the implementation of three variants of MINER presented in §4.3. For w/o IR, we reward all the responses from homogeneous groups with a fixed advantage score (0.05). This value is taken by referring to Zhu et al. (2025), which rewards positive rollouts with a value less than 0.1. For w/o FW, the weight $w_{i,j}$ for a token $o_{i,j}$ from a rollout \mathbf{o}_i is set to 1 for any token. For w/o AC, the $A_{i,j}^{\text{final}}$ equals to the original $\tilde{A}_{i,j}$ without calibration.

C.6 Details of Experiments of RQ2

These experiments involve many other baselines to enhance the model exploration. The following baselines are fetched directly from their officially released codebase. For BAPO, it controls the clip range in an asymmetric manner, which allows the clip_high argument to be adjusted within the range [1.5, 3.0] in a step of 0.1, and the clip_low argument to be adjusted within [0.5, 0.95] in a step of 0.05. The termination rule for adjustment is that the ratio of positive tokens accounts for 50% of the training batch. For KL-Cov, tokens with the top 0.2% covariance would be augmented with a KL loss by regulating the distribution between the old policy and the current policy. For Clip-Cov, the original clip range [0.8, 1.2] is modified to [0, 2]. Meanwhile, 0.02% tokens whose covariance score is located within [1.0, 5.0] would be sampled randomly from the training batch and clipped from the current training step. For Entropy-Adv, the additional entropy bonus, defined as $\psi(\mathcal{H}_{i,j}) = \alpha \cdot \mathcal{H}_{i,j}$ ($\alpha = 0.4$) is set to be no greater than half of the absolute value of the original advantage score. The shaped advantage function is defined as $A_{i,j}^{\text{shaped}} = A_{i,j} + \psi(\mathcal{H}_{i,j})$.

Training and evaluation prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The assistant thinks deeply and output the final answer within `\boxed{}`.

User: {prompt}

Assistant:

Figure 7: Training and evaluation prompt

C.7 Details of Experiments of RQ5

The training data is constructed as follows. As the original training set of MedMCQA contains 182K data, which includes many low-quality questions. Therefore, we use a simple filtering rule, which we prompt Llama3.1-8B-Instruct (Dubey et al., 2024) to conduct **greedy** decoding on the whole dataset, and filter questions that are judged to be correct. We do not conduct a similar filtering process on the MedQA training set, as its data is more challenging than that of MedMCQA. Finally, the number of training set from MedMCQA reaches 73K; the combination of the MedQA training set includes 85K high-quality training data.

D Additional Experiments

D.1 Pass@K Scaling

We compare MINER with the base model and GRPO variants, on the representative exploration metric, i.e., Pass@K, to unveil the improved exploration potentials. Specifically, we select $k = [1, 2, 4, 8, 16, 32, 64, 128]$, and show the detailed improvements on the four challenging benchmarks (AIME24, AIME25, AMC23, and HMMT25) in Table-[8,9,10,11] with 10 runs. The visualized result is shown in Fig. 8a. We observe that MINER consistently outperforms other methods in most of the challenging benchmarks under various sampling candidates, especially on the extremely challenging HMMT25 set, which reflects the superb potential of MINER for breaking the capability boundary.

D.2 Performance Across Diverse Difficulties

In this section, we examine whether MINER improves the model comprehensively across diverse difficulty levels. We compute pass@1 and pass@K scores on the MATH500 and AIME2024 datasets, which provide self-contained difficulty gradients

across six levels. As shown in Fig. 8b, MINER maintains the same mastery as GRPO on easy problems, again verifying that MINER does not sacrifice exploitation for exploration. And the performance leap enlarges with the increase of problem complexity, showing MINER enhances LRM’s performance in a promising manner for breaking more knowledge boundaries.

D.3 Computational Overhead

In this section, we compare MINER with GRPO in terms of the timing cost of the auxiliary advantage computation. We derive the ratio of time for computing advantages against that for completing a training step during a whole training epoch. Results in Fig. 8c demonstrate that the additional advantage incurs less than 0.05% more timing cost than the normal GRPO baseline. This consolidates our claim that MINER adds near-zero computation while being sufficiently effective across models and tasks.

D.4 Sensitivity Analysis

In this section, we study the sensitivity of the only hyperparameter λ_{max} using by MINER. We use grid-search using three different λ_{max} values: $[1e-3, 1.5e-3, 2e-3]$ on Qwen3-4B-Base model using the same training configurations used in §4.1. The results in Table 6 demonstrate that $1e-3$ is stable but results in a slow convergent rate, while the auxiliary learning signal given $\lambda_{max} = 2e-3$ pathologically impacts the major optimization objective. In contrast, the value $1.5e-3$ is modest and suitable for both a stable training procedure and fast convergence, presenting significant improvements on Pass@1 and Pass@K with only ~ 300 RL updates. Meanwhile, this hyperparameter is extendable to a larger model, Qwen3-8B-Base and even Llama with a different backbone and intelligence, demonstrating that our method is not sensitive.

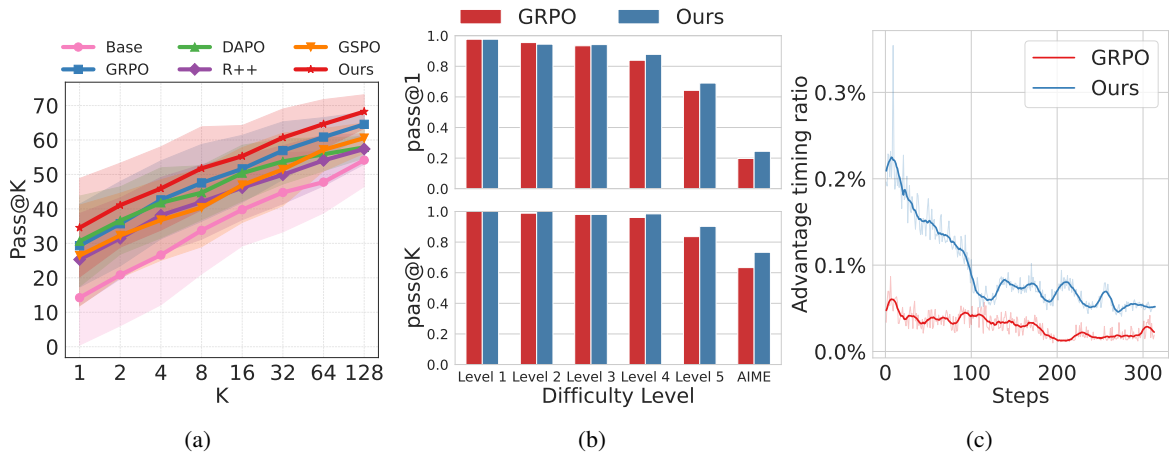


Figure 8: (a) Pass@K scaling comparison of MINER against other GRPO variants; (b) Maintain performance on easy queries and breaking boundaries on challenging problems when evaluating MINER on six difficulty levels sourced from MATH500 and AIME2024; (c) Negligible extra computational overhead compared to GRPO.

λ_{\max}	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB.		Avg.	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
0 (GRPO)	19.79	63.33	20.34	63.33	61.89	97.50	7.86	33.33	83.71	94.00	52.19	72.63	40.96	70.69
$1e-3$	21.74	63.33	19.22	56.67	65.98	95.00	9.74	33.33	84.89	94.80	53.03	73.32	42.43	69.41
$1.5e-3$	25.86	73.33	22.97	60.00	69.65	97.50	10.81	46.67	86.93	95.80	57.07	76.25	45.55	74.92
$2e-3$	20.21	66.67	19.19	60.00	61.88	97.50	10.13	43.33	83.80	96.60	52.02	75.04	41.20	73.19

Table 6: Hyperparameter sensitivity analysis on λ_{\max} . $1e-3$ renders slow convergence given limited data, while $2e-3$ would interfere but not benefit the objective of maximizing correctness.

D.5 Training Logs

We present the training logs of mathematical reasoning, by taking Qwen3-4B-Base and Qwen3-8B-Base in Fig. 9 and Fig. 10, respectively. For the medical reasoning, we present training logs with Llama3.1-8B-Instruct as the backbone in Fig. 11. Here, “master ratio” is the ratio of PH prompts within a batch.

D.6 Trial of RL on Math with Llama

It is difficult to apply MINER to Llama, for the following two reasons: (i) Due to lack of necessary pre-training data, Llama models fail to incentivize reasoning abilities as Qwen; (ii) Even if Llama models are injected required corpus via mid-training (Mo et al., 2025), its reasoning ability is much lower than models using the Qwen backbone, resulting in a much lower portion of PH prompts among the batch. The extreme case where the portion of PH prompts decreases to 0 degrades MINER to normal GRPO. These two reasons lead to not as significant improvement gains as applying MINER to Qwen architectures when optimizing math reasoning. To illustrate, we choose OctoThinker-8B-Hybrid-Base (Wang et al., 2025b),

which undergoes a fine-grained mid-training procedure to enable incentivization of necessary reasoning capabilities like Qwen. Results in Table 7 demonstrate that MINER still outperforms the other algorithms, but with a relatively small margin as expected. However, MINER still makes **+2.18** gains on Pass@1 and **+1.75** gains on Pass@K compared to GRPO, and even outperform the DAPO baseline, with 1.64 absolute gains in Pass@1 and 6.09 pass@K gains in Pass@K, whose rollouts receive non-zero advantages. Given that MINER’s strong performance in such a disadvantageous scenario, it is sufficiently generalizable to other modern models with even higher intelligence.

D.7 Training Stability

Due to severe computational resource constraints, all methods including our approach and the other baselines were trained for a fixed duration of 314 steps (equivalent to one full epoch of data collection). We acknowledge that this may raise concerns about whether the algorithms have fully converged. However, three key observations support the robustness of our conclusions:

First, the training dynamics of our method ex-

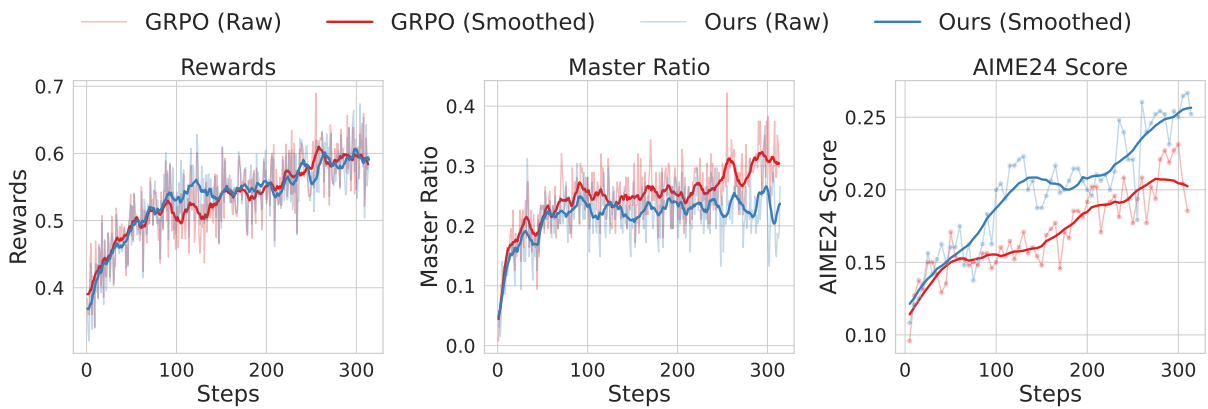


Figure 9: Training rewards, master ratio (PH ratio) and AIME24 dev set score of GRPO and our method trained with Qwen3-4B-Base on the mathematical reasoning task.

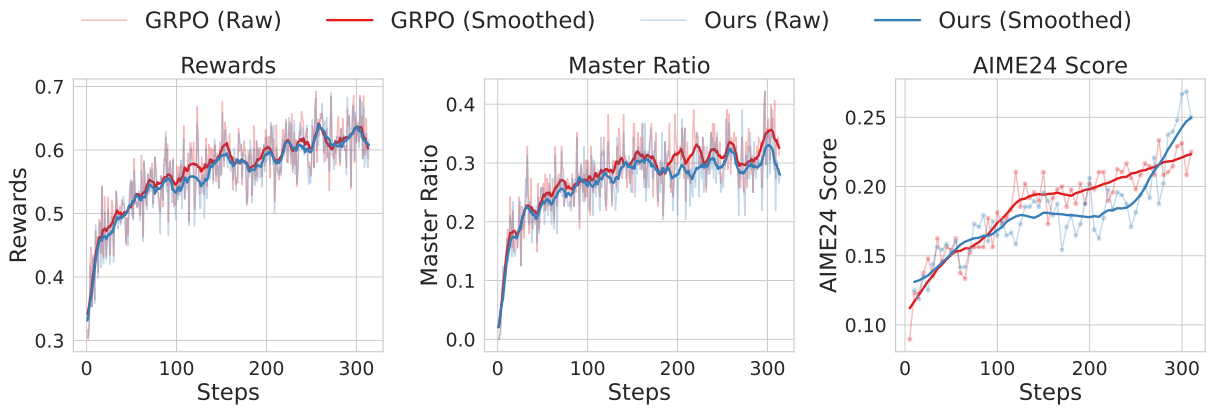


Figure 10: Training rewards, master ratio (PH ratio) and AIME24 dev set score of GRPO and our method trained with Qwen3-8B-Base on the mathematical reasoning task.

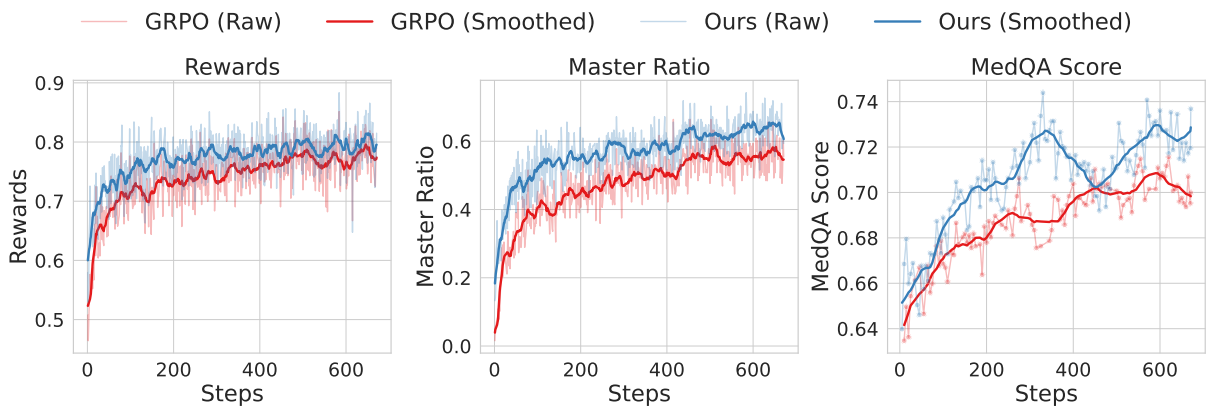


Figure 11: Training rewards, master ratio (PH ratio) and MedQA dev set score of GRPO and our method trained with Llama3.1-8B-Instruct on the medical reasoning task.

Model	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB.		Avg.	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
<i>Base Model: OctoThinker-8B-Hybrid-Base (Llama Arch.)</i>														
Base	0.96	26.67	0.49	23.33	18.95	92.50	0.18	13.33	32.63	82.00	18.18	51.29	11.90	48.19
GRPO	11.95	43.33	12.01	50.00	47.70	92.50	4.64	40.00	76.06	93.00	42.96	67.47	32.55	64.38
DAPO	12.03	46.67	10.05	43.33	52.05	92.50	4.90	20.00	75.93	91.80	43.58	65.92	33.09	60.04
R++	12.53	36.67	9.32	46.67	48.42	92.50	3.49	26.67	75.98	93.40	41.91	67.64	31.94	60.59
GSPO	8.36	43.33	7.50	36.67	41.43	85.00	3.91	23.33	66.51	87.40	33.87	58.52	26.93	55.71
MINER	14.11	50.00	13.28	50.00	52.73	97.50	5.26	36.67	77.16	93.40	45.80	69.19	34.73	66.13

Table 7: Comprehensive comparison against other critic-free RL algorithms in terms of Pass@1 (P@1) and Pass@K (P@K) scores. ‘‘OlympiadB.’’ refers to the OlympiadBench. Best performance is highlighted with **bold**.

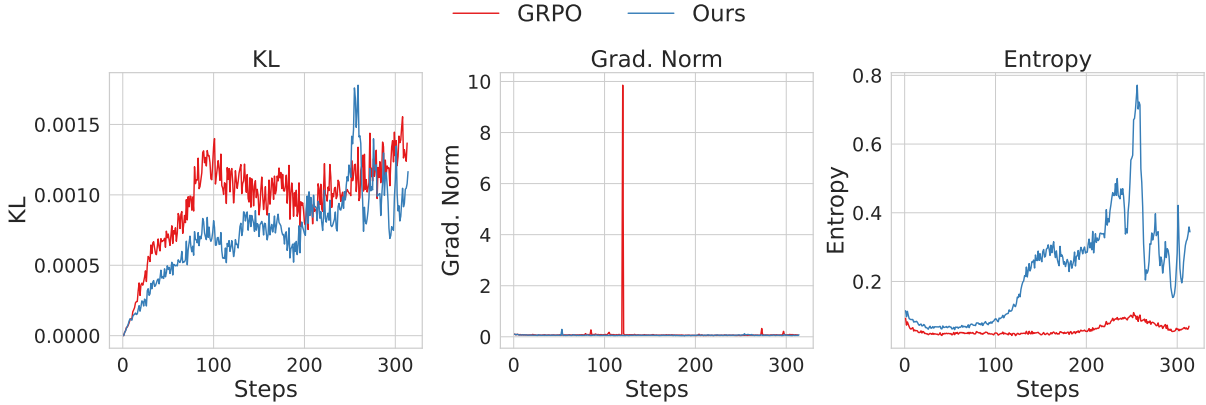


Figure 12: The KL loss, gradient norm and entropy dynamics of applying MINER and GRPO algorithms built on Qwen3-4B. After undergoing a long range of exploration, with steadily increasing policy entropy values, our method quickly transforms the exploration to exploitation signals, accompanied by a rapid fall back of entropy signals and a surge in performances of downstream benchmarks (see Fig. 9).

hibit exceptional stability. As shown in Fig. [12,13], the KL divergence remained consistently low, and the gradient norm stayed within an even narrower range, indicating no signs of divergence or oscillation. This stability implies that extending training would likely preserve our method’s performance gains rather than erode them.

Second, the late-stage performance surge (observed in the final 30 steps) is not an artifact of under-training but reflects our method’s deliberate exploration-exploitation trade-off. Specifically, the algorithm prioritizes extensive exploration of the policy space in early stages (evidenced by steady entropy enhancement), enabling it to discover high-reward regions that GRPO overlooks. Once a promising mode is identified (around step 280), rapid policy refinement occurs, causing the sharp performance lift. This behavior, common in entropy-regularized RL algorithms, is a feature of MINER; it ensures thorough exploration before committing to exploitation. This behavior helps to achieve a great trade-off between pass@1 and

pass@K, avoiding mode collapse which is a known drawback of GRPO.

Consequently, while longer training was infeasible for all baselines under our constraints, the combination of stable convergence indicators and the intrinsic exploration dynamics suggests our method would maintain its lead if trained to full convergence. We will extend the training to guarantee a sound conclusion in the later stage.

E Clarification of MINER on Trivial Prompts

In this section, we clarify that MINER naturally skips *truly trivial* problems (e.g., 1+1) while selectively reinforcing *fragile-correct* PH prompts where correctness is unstable—a distinction critical to data efficiency.

1. **Automatic filtering of trivial problems.** By omitting std-normalization in Eq. 6 (unlike GRPO Eq. 2), MINER preserves absolute uncertainty scales. For trivial prompts like

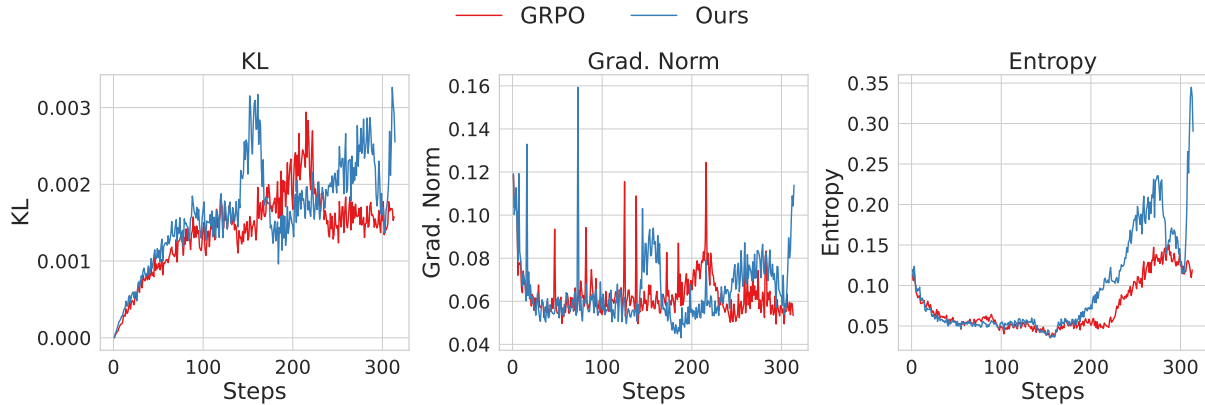


Figure 13: The KL loss, gradient norm and entropy dynamics of applying MINER and GRPO algorithms built on Qwen3-8B. After undergoing a long range of exploration, with steadily increasing policy entropy values, our method quickly transforms the exploration to exploitation signals, accompanied by a rapid fall back of entropy signals and a surge in performances of downstream benchmarks (see Fig. 10).

1+1, all rollouts exhibit near-identical low NLL, which implies that $A_i^{\text{int}} \approx 0$ after mean-centering and the rectified value $\text{ReLU}(A_i^{\text{int}}) \approx 0$. Token-level focal weighting (Eq. 8) further suppresses gradients on deterministic tokens. Thus, MINER automatically allocates zero gradient budget to already-mastered prompts, requiring no manual filtering.

2. **Value of “fragile-correct” PH prompts.** 100% accuracy within 16 trials is unequal to true mastery. As shown in Fig. 8b, MINER gains 0 improvements on Level 1–3 (truly mastered) but +5.2% on Level 4–6 (fragile-correct) compared to GRPO, confirming it targets unstable correctness instead of already mastery. Crucially, training curves (Fig. 9–10) reveal a late-stage surge ($>+5$ Pass@1 in the final 30 steps), evidencing that consolidating fragile paths yields delayed but high-value signals instead of trivial memorization.
3. **Generalization beyond training domain.** MINER’s gains transfer to medical reasoning (Table 2) and scale with test-time compute (Fig. 4c, 5), demonstrating that PH consolidation enhances robust reasoning capability rather than overfitting. In Fig. 11, the policy answers 60% of the problems with all 16 rollouts. This is similar to a scenario where input problems are easy enough for the policy. However, from Table 2, MINER consistently improves over GRPO by 1.39 pass@1 and 1.84 pass@K on average, which demonstrates

that with the proposed three key components of Miner, it can learn only useful information from PH prompts and boost the final performance in terms of exploration and exploitation even trained on a large amount of easy problems.

4. **Connection with prior work.** MINER extracts additional signals from the policy’s intrinsic uncertainty apart from the outcome verifier, without external synthesis or auxiliary models. Such intrinsic learning signals impart valuable learning signals and the policy could learn from positive homogeneous problems, i.e., problems whose all rollouts are incorrect.

F License

License We plan to release our training and evaluation code under the MIT License (or Apache-2.0). Model checkpoints will be distributed for research use only and will comply with the license terms of the underlying base model.

Intended use and compatibility with upstream terms. We use existing artifacts (e.g., base models, datasets, and toolchains) in a manner consistent with their stated intended use and license/terms when specified. In particular, we only use resources available for research and comply with any restrictions on redistribution and derivative works. We will release our code under the [MIT/Apache-2.0] license. We will release our model checkpoints for research use only, and their use and redistribution are subject to the licenses/terms of the underlying base model(s) and dataset(s). We do not authorize

Token Advantage Visualization

Sequence Advantage: 0.0610

Prompt

A conversation between User and Assistant. The user asks a question, and the Assistant solves it. The assistant first thinks about the reasoning process in the mind and then provides the user with the answer. The assistant thinks deeply and output the final answer within `\boxed{}`.
 User: I have a bag with yellow and blue marbles in it. At the moment, the ratio of blue marbles to yellow marbles is 4:3. If I add 5 blue marbles and remove 3 yellow marbles, the ratio will be 7:3. How many blue marbles were in the bag before I added more? Let's think step by step and output the final answer within `\boxed{}`.
 Assistant:

Token-level Advantages

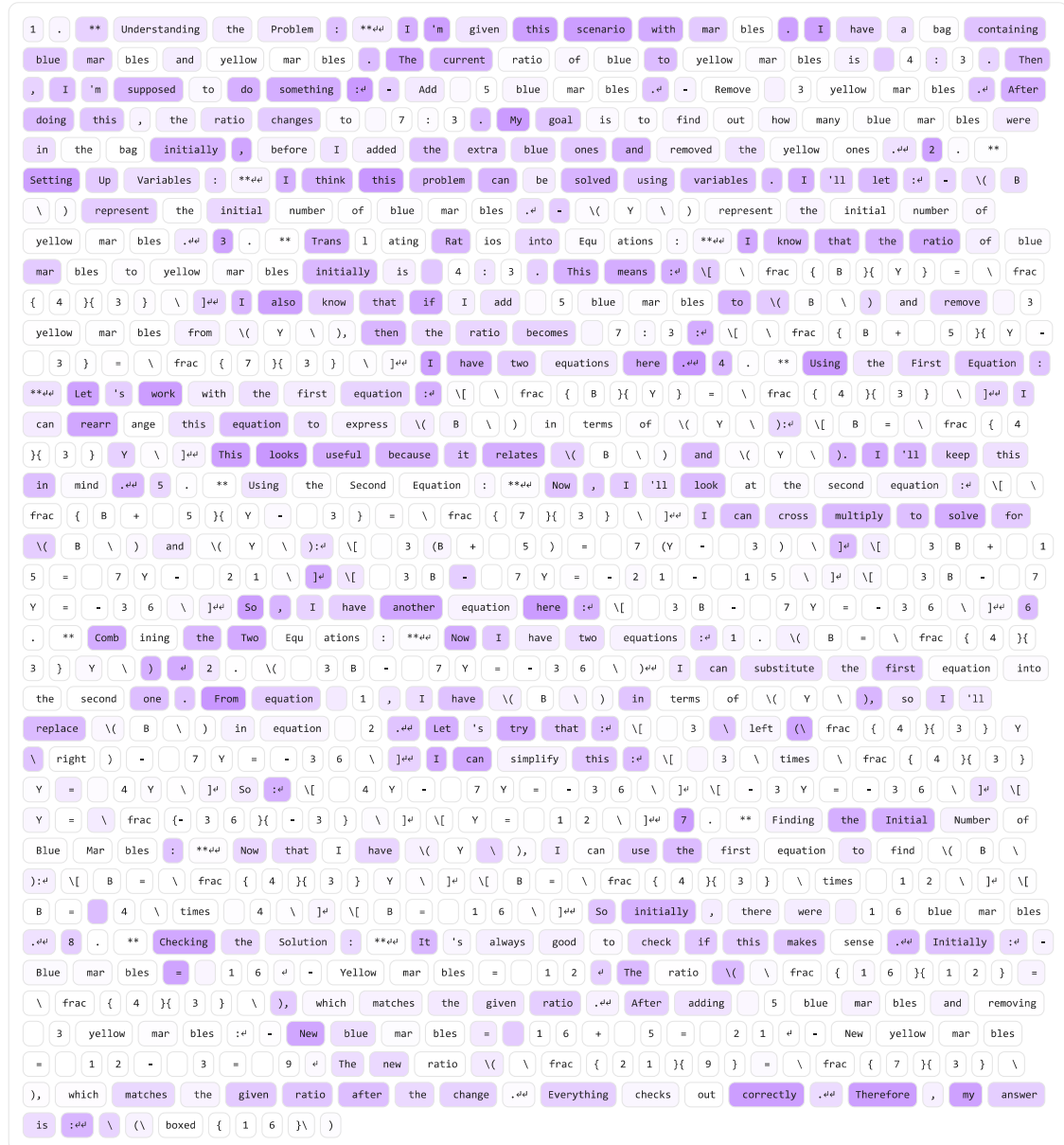


Figure 14: Advantage distribution of MINER. The problem is sourced from the MATH500 dataset. The darker the token, the more advantage credit is assigned. The maximum token advantage equals to the sequence advantage value.

uses that would violate upstream access conditions (e.g., non-research use when restricted) and we do not claim additional rights over third-party resources.

G Data Cleaning

We use only publicly available open-source datasets. These datasets consist of task/problem content and are not intended to contain personally identifying information. We additionally performed sanity checks via automated pattern matching (e.g., emails/phone numbers/URLs) and spot-checking, and did not observe PII. We do not redistribute any third-party data and only release code/model under the upstream licenses/terms; any examples shown are sanitized.

H Case Study

In this section, we present the advantage distribution of MINER trained on Qwen3-8B-Base. We randomly choose one problem sourced from MATH500 to illustrate the advantage score, as it contains numerous PH prompts. The case shown in Fig. 14 demonstrates that most deterministic tokens are not rewarded, which prevents overfitting to already-mastered actions. Noticeably, MINER enhances the frequency of rewarding reasoning-required actions, including “Checking” for verification, “So” for implication, “Let” for solution progression, and “Combining” for conditions gathering. This case reflects meaningful learning signals for utilizing PH prompts via our method, significantly enhancing the data efficiency for RLVR.

Model	1	2	4	8	16	32	64	128
Base	8.00 ± 9.19	12.67 ± 8.92	17.67 ± 9.25	22.33 ± 10.57	28.33 ± 10.22	36.67 ± 12.92	37.67 ± 12.15	45.00 ± 8.22
GRPO	22.00 ± 10.02	24.67 ± 9.72	28.33 ± 12.88	31.00 ± 9.97	38.00 ± 13.88	43.33 ± 15.22	51.67 ± 10.90	57.00 ± 7.55
DAPO	22.33 ± 12.35	24.33 ± 6.19	29.67 ± 8.79	36.33 ± 10.57	40.00 ± 9.22	45.33 ± 9.65	50.33 ± 3.53	52.33 ± 1.53
REINFORCE++	15.00 ± 7.53	19.33 ± 9.49	24.33 ± 9.30	27.33 ± 10.39	31.67 ± 13.18	39.33 ± 13.06	46.00 ± 12.04	51.00 ± 9.16
GSPO	15.67 ± 9.25	20.00 ± 6.67	25.00 ± 10.63	27.67 ± 13.69	32.67 ± 13.63	40.00 ± 15.87	51.33 ± 18.48	61.33 ± 13.49
MINER	25.67 ± 14.08	30.33 ± 11.39	36.67 ± 13.86	43.33 ± 17.08	45.67 ± 8.19	54.67 ± 10.02	61.00 ± 9.18	67.00 ± 5.97

Table 8: Pass@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AIME2024 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	7.33 ± 8.02	10.67 ± 11.65	15.33 ± 11.52	20.33 ± 9.97	26.67 ± 7.58	30.33 ± 10.35	37.00 ± 14.59	44.00 ± 14.71
GRPO	19.33 ± 11.49	25.67 ± 11.00	32.67 ± 10.97	35.67 ± 12.86	43.33 ± 14.16	49.00 ± 9.92	56.33 ± 9.02	59.33 ± 5.49
DAPO	19.33 ± 9.19	24.00 ± 7.35	28.00 ± 11.86	31.33 ± 10.60	36.33 ± 10.35	43.00 ± 6.35	44.33 ± 5.69	48.00 ± 4.00
R++	13.33 ± 11.58	16.00 ± 8.39	19.00 ± 7.02	24.00 ± 8.16	31.33 ± 9.72	31.00 ± 9.53	37.00 ± 6.13	39.33 ± 4.16
GSPO	12.67 ± 8.16	17.33 ± 13.58	20.00 ± 10.69	26.33 ± 8.43	28.33 ± 11.46	35.00 ± 8.72	39.67 ± 8.13	44.67 ± 2.97
MINER	23.00 ± 9.97	26.00 ± 12.60	30.67 ± 13.63	37.00 ± 13.39	42.00 ± 13.79	47.00 ± 8.49	51.67 ± 7.67	55.33 ± 6.83

Table 9: Pass@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AIME2025 benchmark.

Model	1	2	4	8	16	32	64	128
Base	32.75 ± 28.90	46.00 ± 28.26	58.25 ± 27.01	71.75 ± 19.97	81.75 ± 14.51	87.50 ± 7.86	91.25 ± 3.97	95.00 ± 4.62
GRPO	62.75 ± 20.46	72.25 ± 16.61	79.75 ± 14.44	87.25 ± 11.54	90.75 ± 5.27	94.75 ± 3.25	96.25 ± 1.97	96.50 ± 1.22
DAPO	63.50 ± 15.09	71.25 ± 14.99	78.75 ± 10.99	84.00 ± 9.08	88.00 ± 6.62	90.50 ± 3.29	93.50 ± 2.22	94.00 ± 1.90
R++	57.50 ± 20.98	69.75 ± 19.34	79.00 ± 13.12	82.75 ± 9.77	88.50 ± 6.40	91.75 ± 3.97	92.00 ± 2.40	93.50 ± 2.22
GSPO	57.25 ± 19.68	72.00 ± 18.97	77.25 ± 14.01	82.75 ± 12.22	89.50 ± 10.52	94.25 ± 3.47	95.50 ± 3.00	97.00 ± 1.00
MINER	68.25 ± 21.69	79.00 ± 16.98	85.75 ± 11.72	91.75 ± 6.37	94.00 ± 1.90	95.75 ± 2.37	97.00 ± 1.00	97.00 ± 1.00

Table 10: Pass@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AMC23 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	0.33 ± 1.00	1.33 ± 3.33	3.33 ± 5.33	6.67 ± 8.67	8.00 ± 7.22	11.00 ± 8.58	14.33 ± 11.11	20.67 ± 9.35
GRPO	8.00 ± 9.19	10.00 ± 5.63	16.00 ± 10.60	19.67 ± 8.27	22.33 ± 5.67	26.67 ± 6.30	30.00 ± 5.27	31.33 ± 3.06
DAPO	6.67 ± 8.02	12.67 ± 9.35	13.67 ± 6.35	17.00 ± 6.49	23.67 ± 6.93	22.00 ± 4.97	23.67 ± 5.39	26.67 ± 3.27
REINFORCE++	4.33 ± 6.86	6.00 ± 7.33	11.67 ± 11.92	16.33 ± 11.11	21.00 ± 10.35	27.33 ± 8.16	32.33 ± 5.90	33.33 ± 3.33
GSPO	5.00 ± 7.39	8.67 ± 9.06	9.33 ± 10.53	16.33 ± 10.59	19.00 ± 10.43	26.33 ± 7.00	28.33 ± 4.33	29.67 ± 1.00
MINER	10.00 ± 8.97	16.33 ± 6.19	17.33 ± 9.27	21.67 ± 13.30	26.67 ± 14.02	33.67 ± 14.29	38.00 ± 12.62	44.00 ± 7.13

Table 11: Pass@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the HMMT25 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	7.67 ± 8.93	10.00 ± 9.19	10.33 ± 5.30	13.00 ± 5.97	15.67 ± 5.53	17.00 ± 5.30	17.00 ± 2.33	17.67 ± 1.53
GRPO	20.00 ± 9.13	20.33 ± 10.49	21.33 ± 6.83	25.33 ± 5.86	26.67 ± 2.00	26.33 ± 3.53	26.67 ± 0.00	26.67 ± 0.00
DAPO	22.67 ± 12.46	19.67 ± 13.16	23.67 ± 9.11	25.00 ± 6.86	25.00 ± 3.53	27.67 ± 3.00	27.33 ± 4.67	28.33 ± 2.86
REINFORCE++	16.67 ± 10.16	16.33 ± 11.39	19.33 ± 7.19	19.67 ± 3.30	21.67 ± 3.67	22.33 ± 1.53	23.00 ± 1.00	22.67 ± 1.33
GSPO	17.00 ± 14.13	14.67 ± 8.69	18.00 ± 7.86	20.33 ± 5.53	20.33 ± 3.67	21.00 ± 1.53	21.00 ± 1.53	22.00 ± 2.53
MINER	27.67 ± 15.39	27.33 ± 15.44	26.33 ± 6.86	29.33 ± 8.02	33.67 ± 7.02	35.00 ± 3.97	36.00 ± 3.33	35.67 ± 2.33

Table 12: Maj@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AIME2024 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	6.33 ± 9.16	6.67 ± 9.22	8.33 ± 8.13	12.33 ± 8.49	12.33 ± 7.64	15.33 ± 6.83	18.67 ± 4.83	19.67 ± 4.06
GRPO	20.00 ± 14.02	18.67 ± 9.86	22.33 ± 10.79	25.33 ± 7.65	26.67 ± 7.57	28.00 ± 6.11	29.67 ± 5.83	28.67 ± 4.16
DAPO	22.33 ± 9.69	16.00 ± 8.97	20.00 ± 8.60	22.67 ± 6.30	23.33 ± 6.30	24.00 ± 5.33	24.67 ± 3.86	24.67 ± 3.63
REINFORCE++	12.00 ± 6.39	13.00 ± 8.49	16.00 ± 7.58	16.00 ± 6.00	17.00 ± 3.16	18.00 ± 2.67	16.67 ± 0.00	16.67 ± 0.00
GSPO	10.33 ± 8.79	13.00 ± 6.49	14.33 ± 7.16	19.00 ± 8.86	19.00 ± 6.49	19.00 ± 4.63	19.00 ± 3.67	19.00 ± 2.63
MINER	23.33 ± 11.86	22.33 ± 8.33	22.67 ± 5.19	26.00 ± 9.72	26.00 ± 7.72	29.33 ± 5.49	28.00 ± 5.27	28.67 ± 2.67

Table 13: Maj@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AIME2025 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	35.00 ± 31.95	31.50 ± 26.39	40.00 ± 24.74	46.00 ± 16.36	48.75 ± 14.16	51.00 ± 14.53	50.00 ± 12.72	50.75 ± 9.32
GRPO	62.00 ± 21.92	62.00 ± 17.65	66.00 ± 17.60	66.50 ± 14.35	69.50 ± 11.36	69.25 ± 7.49	70.75 ± 6.20	70.75 ± 4.87
DAPO	65.00 ± 16.14	62.75 ± 15.39	69.25 ± 13.22	69.00 ± 11.91	70.25 ± 9.89	73.75 ± 5.10	73.25 ± 4.65	73.75 ± 4.15
REINFORCE++	57.75 ± 20.72	58.75 ± 21.67	63.75 ± 16.72	67.00 ± 12.09	70.00 ± 9.24	72.50 ± 8.35	73.25 ± 7.54	74.25 ± 6.04
GSPO	56.75 ± 22.36	60.00 ± 21.14	64.75 ± 18.44	65.50 ± 15.02	69.25 ± 12.06	73.00 ± 9.45	73.00 ± 6.02	74.50 ± 3.50
MINER	68.50 ± 22.07	70.00 ± 17.77	76.25 ± 17.77	84.25 ± 10.27	84.75 ± 7.04	84.50 ± 4.90	85.00 ± 3.90	85.50 ± 1.90

Table 14: Maj@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the AMC23 benchmark with Qwen3-4B-Base as the base model.

Model	1	2	4	8	16	32	64	128
Base	1.00 ± 3.00	1.67 ± 2.63	0.67 ± 1.33	2.67 ± 1.33	2.33 ± 1.53	3.00 ± 1.00	3.00 ± 1.00	3.33 ± 0.00
GRPO	6.67 ± 5.33	8.67 ± 9.27	10.33 ± 5.67	10.67 ± 6.02	10.00 ± 2.00	12.67 ± 2.53	13.00 ± 1.00	12.67 ± 1.33
DAPO	8.00 ± 7.33	6.33 ± 8.93	8.33 ± 6.72	9.67 ± 3.30	10.33 ± 3.30	11.67 ± 2.63	12.00 ± 2.53	12.67 ± 1.33
REINFORCE++	5.67 ± 9.00	3.33 ± 4.53	5.67 ± 4.63	6.00 ± 2.97	6.33 ± 4.63	5.67 ± 2.63	5.33 ± 2.97	5.67 ± 3.16
GSPO	4.00 ± 7.49	6.00 ± 8.60	5.00 ± 3.67	6.33 ± 2.86	8.33 ± 4.86	9.00 ± 3.00	8.67 ± 1.63	8.67 ± 1.63
MINER	11.00 ± 8.49	11.00 ± 4.79	11.67 ± 5.69	12.67 ± 5.79	15.33 ± 3.86	17.67 ± 2.33	16.67 ± 0.00	16.67 ± 0.00

Table 15: Maj@K comparison across diverse k list [1,2,4,8,16,32,64,128] with repeated 10 runs on the HMMT25 benchmark with Qwen3-4B-Base as the base model.

Model	AIME2024		AIME2025		AMC23		HMMT25		MATH		OlympiadB		Avg	
	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K	P@1	P@K
Base	9.51	50.00	6.65	43.33	32.91	97.50	0.99	26.67	58.89	91.00	31.81	65.40	23.46	62.32
BAPO	17.86	50.00	16.69	56.67	61.04	95.00	8.57	33.33	82.81	94.40	51.32	70.40	39.72	66.63
KL-Cov	17.73	53.33	20.26	40.00	59.96	90.00	7.16	23.33	83.38	93.20	49.63	68.85	39.69	61.45
Clip-Cov	15.55	60.00	14.40	46.67	59.98	97.50	4.48	36.67	80.98	94.40	49.02	71.94	37.40	67.86
Entropy-Adv	17.71	63.33	15.99	50.00	61.86	97.50	6.51	30.00	81.73	94.80	50.17	73.67	38.99	68.22
MINER	25.86	73.33	22.97	60.00	69.65	97.50	10.81	46.67	86.93	95.80	57.07	76.25	45.55	74.92

Table 16: Comparison with other exploration-enhanced algorithms. Given numerous hyperparameter combinations, these methods show bad generalization with their officially suggested hyperparameters.