SURPRISING SUCCESS, REPEATED FAILURE: ENTROPY-GUIDED CREDIT ASSIGNMENT FOR EXPLORATION IN LLM REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) enhances reasoning in large language models (LLMs) through outcome-level feedback, yet recent approaches to finer-grained credit assignment often require auxiliary models, additional sampling, or privileged information. Although policy entropy provides a readily available signal, prioritizing uncertain positions under both reinforcement and penalization concentrates penalties where failed responses still retain alternatives for recovery, which can suppress opportunities for exploration. To address this, we introduce *Entropic Advantage Policy Optimization* (**EAPO**), an entropy-guided credit assignment method that treats success and failure asymmetrically. Specifically, motivated by the observation that success under uncertainty is less repeatable while confident failures tend to recur, EAPO couples normalized policy entropy with the sign of the response advantage to reinforce ***surprising success*** and correct ***repeated failure***. It assigns stronger reinforcement to high-entropy decisions in successful responses and stronger penalties to low-entropy decisions in failed responses, while attenuating penalties at uncertain positions to preserve opportunities for recovery. By redistributing the response advantage across tokens, EAPO derives token-level credit directly from existing rollout signals without additional supervision. We validate EAPO on a range of reasoning tasks across both base and reasoning backbones, demonstrating that it achieves the best overall performance. We further show that EAPO promotes more effective exploration, broadening problem coverage and generating more diverse candidate answers.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) (Shao et al., 2024; Lambert et al., 2025; Guo et al., 2025a) improves the reasoning capabilities of large language models (LLMs) through outcome-level supervision. The resulting reward evaluates a trajectory as a whole, although the decisions within it may contribute differently to the final outcome. This motivates finer-grained credit assignment that allocates learning feedback to individual reasoning decisions. However, existing approaches to obtaining such feedback require auxiliary models, additional sampling, or access to privileged information (Kazemnejad et al., 2025; Cui et al., 2026; Kim et al., 2026a).

These requirements motivate the use of *policy entropy*, a measure of next-token uncertainty directly available from the rollout policy, as a signal for token-level feedback allocation. High entropy reflects competing continuations and potential branching points for exploration, whereas low entropy indicates concentrated preferences (Wang et al., 2025; Cheng et al., 2026a). Beyond entropy regularization for exploration (Mnih et al., 2016), existing methods select high-entropy tokens for policy updates (Wang et al., 2025) or reshape token-level advantages (Cheng et al., 2026a; He et al., 2026). However, these methods treat uncertainty in the same way under success and failure, either adding nonnegative entropy bonuses regardless of outcome or favoring high-entropy positions under both positive and negative feedback. Such a shared preference places the strongest penalties on uncertain positions in failed responses, where competing continuations may still lead to recovery through alternative reasoning paths, potentially constraining further exploration. This raises a central question: *should uncertainty guide both reinforcement and penalization in the same way?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

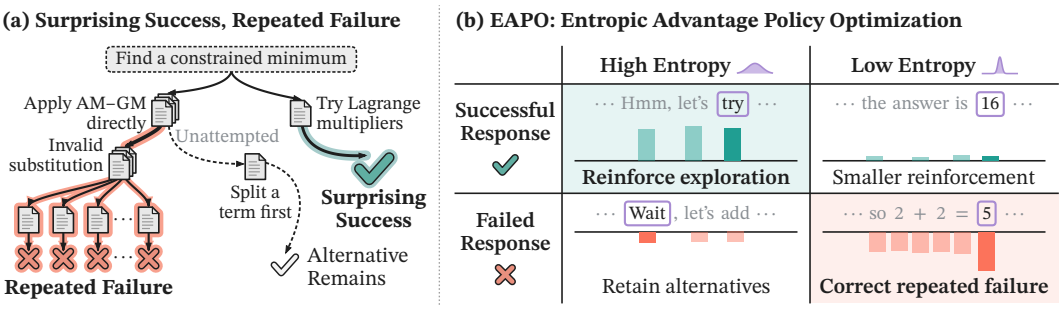

Figure 1: **Conceptual illustration.** *(a)* Success under uncertainty is rarely sampled and thus less repeatable, whereas confident failures keep recurring across many rollouts. *(b)* EAPO reweights token-level advantages by entropy to reinforce ***surprising success*** at high-entropy positions and correct ***repeated failure*** at low-entropy positions, while preserving uncertain alternatives.

We examine this question through the asymmetry between reinforcing success and penalizing failure. In a successful trajectory, a high-entropy decision selects among competing continuations, so the observed success may reflect an exploratory path that has not yet become a stable behavioral preference. Stronger reinforcement at these points can help consolidate successful exploration into more repeatable behavior. In contrast, low-entropy decisions within an unsuccessful trajectory can reflect concentrated preferences that favor similar failing continuations, motivating stronger correction of confident patterns associated with failure to discourage their recurrence. Meanwhile, in the same unsuccessful trajectory, high-entropy positions retain competing alternatives, so concentrating penalties on these positions risks constraining opportunities for exploration and recovery. These two observations motivate reinforcing ***surprising success*** and correcting ***repeated failure*** while preserving uncertain alternatives, without treating entropy as a measure of individual-token correctness.

To incorporate this asymmetry into token-level credit assignment, we introduce **E**ntropic **A**dvantage **P**olicy **O**ptimization (**EAPO**), which couples policy entropy with the correctness feedback (i.e., the advantage sign) to reverse the entropy preference between reinforcement and penalization (Figure 1). Specifically, EAPO reweights the response advantage across tokens, assigning stronger reinforcement to high-entropy decisions when the advantage is positive and stronger penalties to low-entropy decisions when it is negative. By reinforcing surprising success and correcting repeated failure in this way, EAPO makes successful exploration more repeatable while attenuating penalties at uncertain positions within unsuccessful responses to preserve alternative reasoning paths.

We validate EAPO on a range of tasks, including mathematical, logical, and algorithmic reasoning, across both base and reasoning backbones. EAPO achieves the best overall performance, surpassing existing methods for promoting exploration in both average accuracy and problem coverage. Our analyses show that EAPO broadens problem coverage under varying test-time sampling budgets and maintains greater answer diversity on challenging problems, while also confirming the benefits of reinforcing high-entropy decisions and penalizing low-entropy decisions. Together, these findings highlight the benefits of treating uncertainty differently under success and failure to promote more effective exploration in LLM reasoning while preserving alternative reasoning paths.

2 Preliminaries

Group Relative Policy Optimization. Reinforcement learning with verifiable rewards (RLVR) trains a policy π_θ using rewards computed by task-specific verifiers (Lambert et al., 2025; Guo et al., 2025a). As one instantiation, group relative policy optimization (GRPO) (Shao et al., 2024) samples a group of $G > 1$ responses $\{y_1, \ldots, y_G\} \sim \pi_\theta(\cdot \mid x)$ for each prompt x and normalizes the corresponding response-level rewards $r^i = r(x, y_i)$ to obtain the response advantage

$$\hat{A}^i = \frac{r^i - \mathrm{mean}(\{r^j\}_{j=1}^G)}{\mathrm{std}(\{r^j\}_{j=1}^G) + \varepsilon}. \tag{1}$$

The resulting advantage is shared uniformly across all tokens within the response, $\hat{A}_t^i = \hat{A}^i$, and therefore provides no position-specific weighting of the learning signal, treating all token decisions within a response as equally responsible for the final reward.

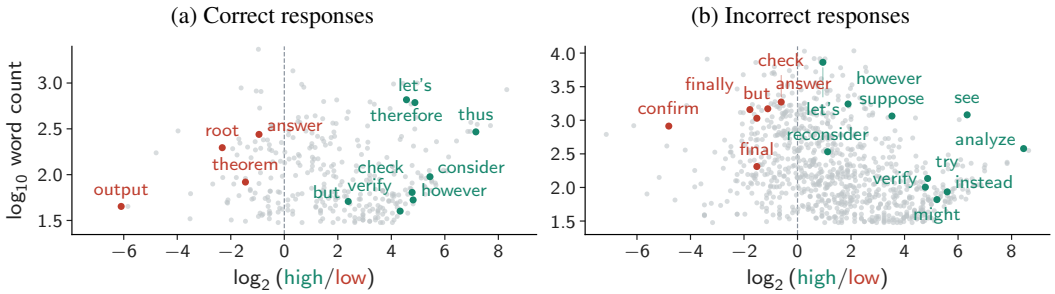

Figure 2: **Word usage by token entropy.** Green and red mark words with higher relative frequencies in the top and bottom 10% entropy tails of each Qwen3-4B-Base response, respectively. Exploratory expressions are enriched at high-entropy positions in both correct and incorrect responses, whereas conclusion markers are enriched at low-entropy positions in incorrect responses.

Policy Entropy. For a generation prefix $(x, y_{i,<t})$ and vocabulary $\mathcal{V}$, we define the *policy entropy* as the Shannon entropy (Shannon, 1948) of the policy's next-token distribution:

$$H_{i,t} = -\sum_{v \in \mathcal{V}} \pi_\theta(v \mid x, y_{i,<t}) \log \pi_\theta(v \mid x, y_{i,<t}). \tag{2}$$

Unlike sampled-token surprisal, $-\log \pi_\theta(y_{i,t} \mid x, y_{i,<t})$, which measures the unexpectedness of a single realized token, $H_{i,t}$ characterizes uncertainty over the full next-token distribution. Higher entropy indicates greater uncertainty among possible continuations, whereas lower entropy reflects more concentrated preferences, with high-entropy positions often associated with branching and exploratory behavior in reasoning (Wang et al., 2025; Cheng et al., 2026a).

3 Surprising Success, Repeated Failure

Then, how does policy entropy relate to reasoning success? While prior work has highlighted high-entropy positions as branching points for exploring alternatives (Wang et al., 2025; Cheng et al., 2026a), this link to exploration does not by itself indicate whether success under uncertainty can be reproduced or whether failure can be avoided even when the policy is confident. In this section, we examine the relationship between policy entropy and successful exploration in reasoning.

High-entropy success is less repeatable, while failure retains alternatives. To examine how entropy relates to reproducible success and recovery from failure, we analyze Qwen3-4B/8B-Base on 20 moderately difficult mathematical reasoning problems per model. For each problem, we randomly select two correct and two incorrect responses. Within each selected response, we compare contiguous 32-token windows with high and low mean token entropy, matching their relative-position ranges to control for where resampling begins in the reasoning trajectory. From the fixed prefix preceding each window, we sample 64 continuations and measure their overlap within the first 32 generated tokens and their final-answer accuracy. Further details are provided in Section C.1.

As shown in Table 1, originally correct responses are substantially less likely to succeed again when resampled from high- rather than low-entropy windows, which makes success through uncertainty a ***surprising success***. In contrast, originally incorrect responses tend to remain incorrect when resampled from low-entropy windows, indicating ***repeated failure*** despite the policy's confidence. Meanwhile, resampling from high-entropy windows of the same incorrect responses yields higher accuracy, suggesting that these responses retain alternative paths to success despite recurrent failure at low-entropy windows. High-entropy windows also yield lower overlap across both outcomes, supporting their role in exploratory branching. We provide a qualitative example illustrating these trends in Section F. These findings highlight the need to make surprising success more repeatable and correct repeated failure at low-entropy positions, while preserving alternative paths to recovery at high-entropy positions in unsuccessful responses.

Table 1: Continuation resampling by original response outcome and window entropy. $\Delta\bar{r}$ denotes the reward difference (High − Low).

	Overlap		Reward		
Source	Low	High	Low	High	$\Delta\bar{r}$ (pp)
Qwen3-4B-Base					
Correct	29.1	4.1	0.578	0.390	−18.8
Incorrect	27.8	7.3	0.070	0.163	+9.3
Qwen3-8B-Base					
Correct	29.6	4.3	0.625	0.418	−20.7
Incorrect	30.2	6.9	0.032	0.159	+12.7

Token patterns reflect successful exploration and confident failure. To characterize the reasoning behaviors behind these trends, we analyze 2,376 Qwen3-4B-Base responses to mathematical reasoning problems, comparing normalized word frequencies between the top and bottom 10% entropy tails of each correct and incorrect response. As shown in Figure 2, high-entropy positions in correct responses are enriched with exploratory expressions (e.g., "let's" and "consider"), whereas conclusion markers (e.g., "finally" and "confirm") are more prevalent at low-entropy positions in incorrect responses. These patterns mirror the resampling results, suggesting that ***surprising success*** arises from *successful exploration* that has not yet become a stable preference, while ***repeated failure*** reflects *confident failure*, in which the policy firmly commits to an incorrect conclusion. Meanwhile, exploratory expressions (e.g., "try" and "instead") also appear at high-entropy positions in incorrect responses, indicating that failed responses still retain alternatives for recovery, consistent with the gains from high-entropy resampling in Table 1. Together, these observations motivate asymmetric credit assignment: stronger reinforcement of high-entropy decisions can make successful exploration repeatable, whereas concentrating penalties on low-entropy decisions can correct confident failure while sparing the exploratory alternatives that unsuccessful responses retain.

4 Entropic Advantage Policy Optimization

We present **E**ntropic **A**dvantage **P**olicy **O**ptimization (**EAPO**), which operationalizes this asymmetric credit assignment by redistributing the response advantage across completion tokens using only the policy's own entropy, assigning stronger reinforcement to uncertain decisions in positive-advantage rollouts and stronger penalties to confident decisions in negative-advantage rollouts.

4.1 Asymmetric Advantage Reweighting via Sign–Entropy Coupling

An entropy preference shared across advantage signs would prioritize the same positions under reinforcement and penalization, so favoring high entropy would also concentrate penalties on uncertain decisions in negative-advantage responses. To address this, we couple entropy with the advantage sign, reversing the preference between reinforcement and penalization.

Policy Entropy Normalization. Absolute token entropies can vary substantially across responses and fluctuate sharply within a response, making absolute entropy-based reweighting sensitive to noise and potentially destabilizing policy optimization. To address this, we normalize token entropies using shared percentiles within each rollout batch $\mathcal{B}$. Specifically, we evaluate the entropy $H_{i,t}$ in Equation (2) under the old policy π_{old}. For response i, let $\mathcal{I}_i$ denote valid completion positions, excluding prompt and padding tokens, and let $T_i = |\mathcal{I}_i| > 0$. We define

$$h_{i,t} = \text{stopgrad}\left[\text{clip}\left(\frac{H_{i,t} - Q_{\text{lo}}}{Q_{\text{hi}} - Q_{\text{lo}} + \varepsilon_H}, 0, 1\right)\right], \qquad t \in \mathcal{I}_i, \tag{3}$$

where Q_{lo} and Q_{hi} are the 10th and 90th percentiles of valid completion-token entropies in $\mathcal{B}$. We use clipping to limit the influence of extreme entropies and stopgrad to use the entropies as detached credit assignment signals, preventing gradients from propagating through the reweighting factors.

Asymmetric Advantage Redistribution. We design token-level credit to depend jointly on the advantage sign and normalized entropy. Given the response advantage $\hat{A}^i$ in Equation (1), we express this coupling through the signed uncertainty $\text{sign}(\hat{A}^i)h_{i,t}$. For $t \in \mathcal{I}_i$ and $\kappa \geq 0$, we define

$$w_{i,t} = \frac{\exp\Big(\kappa\,\text{sign}(\hat{A}^i)h_{i,t}\Big)}{\frac{1}{T_i}\sum_{u\in\mathcal{I}_i}\exp\Big(\kappa\,\text{sign}(\hat{A}^i)h_{i,u}\Big)}, \qquad \hat{A}^{i,\text{E}}_t = \hat{A}^i w_{i,t}. \tag{4}$$

The exponential weighting in Equation (4) favors larger signed uncertainty, with κ controlling concentration. For $\kappa > 0$, this assigns greater weight to high-entropy decisions when $\hat{A}^i > 0$ and to low-entropy decisions when $\hat{A}^i < 0$, with the ratio between any two weights within a response bounded by e^{κ}. Dividing by the within-response mean of the exponential scores ensures that the advantage is redistributed across tokens rather than rescaled for the response as a whole. We then multiply $\hat{A}^i$ by $w_{i,t}$ to obtain the token-level advantage $\hat{A}^{i,\text{E}}_t$. Since the weights are positive, they preserve the sign of $\hat{A}^i$ while determining how strongly each decision is reinforced or penalized.

Policy Optimization. EAPO simply replaces the response-level advantage $\hat{A}^i$ with the token-level advantage $\hat{A}_t^{i,\mathrm{E}}$ in the original policy objective, leaving the rest of the optimization procedure unchanged. Since token-level credit is derived directly from the policy's own entropy and the existing response advantage, EAPO requires no auxiliary models, token-level supervision, or substantial additional computation, making it readily applicable to existing policy optimization pipelines.

4.2 Theoretical Interpretation

We now turn to the theoretical interpretation of EAPO, connecting its asymmetric credit allocation to KL anchoring in token space and examining how penalty attenuation affects conditional entropy and distributional distortion among unchosen alternatives compared with uniform weighting.

Reinterpreting KL Anchoring in Token Space. KL-regularized RLVR anchors the policy to a reference policy π_{ref} through $D_{\mathrm{KL}}(\pi_\theta \| \pi_{\mathrm{ref}})$ while optimizing verifiable rewards (Shao et al., 2024). We transfer this principle from policy space to credit allocation over token positions, with uniform credit as the reference and signed uncertainty $\mathrm{sign}(\hat{A}^i)h_{i,t}$ as the utility.

Proposition 1 (KL-regularized uncertainty allocation). *Let $u_{i,t} = 1/T_i$ be uniform credit over $\mathcal{I}_i$, and let $\Delta(\mathcal{I}_i)$ denote the probability simplex on these positions. For $\kappa > 0$, the problem*

$$\max_{q \in \Delta(\mathcal{I}_i)} \left\{ \mathrm{sign}(\hat{A}^i)\mathbb{E}_{t\sim q}[h_{i,t}] - \frac{1}{\kappa} D_{\mathrm{KL}}(q\|u_i) \right\} \tag{5}$$

has the unique solution $q^*_{i,t} \propto u_{i,t} \exp\Big(\kappa\, \mathrm{sign}(\hat{A}^i)h_{i,t}\Big)$.

We defer the proof to Section E.1. Our formulation in Equation (4) is therefore an operationalization of asymmetric credit assignment based on signed uncertainty, obtained by transferring KL anchoring to token space with uniform allocation as the reference. Increasing κ weakens this anchor and strengthens the preference for high entropy under positive advantages and low entropy under negative advantages, while $\kappa \to 0$ recovers uniform credit.

Understanding the Role of Penalty Attenuation. Motivated by the observation that updates with negative advantages can shift probability toward already likely alternatives (Ren & Sutherland, 2025), we analyze how this effect depends on update strength to understand EAPO's penalty attenuation at uncertain positions. Specifically, at a fixed prefix, we consider a single policy-gradient step on the logits with negative advantage $\hat{A} < 0$ and update strength $\beta = \eta|\hat{A}|w \geq 0$, where w is the allocation weight and $\eta > 0$ is the effective step size. Let $\nu^{(\beta)}$ denote the conditional distribution over unchosen tokens after the update, with $\nu = \nu^{(0)}$ denoting their initial conditional distribution.

Proposition 2 (Concentration under negative updates). *For a single negative logit update with finite logits and the initial distribution held fixed, the sampled-token probability strictly decreases with β, while $H(\nu^{(\beta)})$ is nonincreasing and $D_{\mathrm{KL}}(\nu^{(\beta)}\|\nu)$ is nondecreasing.*

Stronger penalties suppress the sampled token and favor already likely alternatives, sharpening the distribution over the remaining alternatives. Within this single-step logit model, reducing the penalty at a given position limits this sharpening and KL distortion. The proof and additional analyses are provided in Sections E.2 and E.3. Together with the findings in Section 3, this analysis supports attenuating penalties at uncertain positions while concentrating correction on repeated failure.

5 Experiments

5.1 Experimental Setup

Benchmarks & Metrics. We evaluate EAPO on six competition-level mathematical reasoning benchmarks: AIME24/25/26 (MAA, 2026), HMMT26 (HMMT, 2026), AMC23 (MAA, 2023), and the level-5 subset of MATH500 (Lightman et al., 2024). We report avg@32, the mean accuracy over 32 sampled responses per problem, and pass@32, the probability of generating at least one correct response within 32 samples, which we estimate using the unbiased estimator of Chen et al. (2021).

Table 2: Results of diverse RLVR methods with base and reasoning models across six mathematical reasoning benchmarks. **Bold** denotes the best performance within each backbone.

Method	AIME24		AIME25		AIME26		HMMT26		AMC23		MATH500-H	
	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32
Qwen3-4B-Base	10.7	43.3	5.6	30.0	6.8	23.3	3.9	24.2	42.3	90.0	43.0	82.8
GRPO	12.5	33.3	10.4	36.7	7.5	33.3	5.9	24.2	54.5	90.0	56.1	85.1
EntropyAdv	13.1	46.7	12.5	36.7	10.0	33.3	6.3	27.3	53.8	90.0	56.7	86.6
HAPO	12.2	36.7	11.0	40.0	8.9	30.0	6.5	24.2	52.0	87.5	53.2	85.1
80/20	11.9	33.3	12.7	40.0	8.6	30.0	6.6	24.2	54.1	**95.0**	52.8	84.3
RLRT	15.3	50.0	12.7	33.3	14.2	40.0	8.9	**36.4**	54.9	**95.0**	57.4	86.6
EAPO (Ours)	**20.2**	**53.3**	**17.1**	**46.7**	**15.7**	**50.0**	**13.4**	33.3	**56.3**	92.5	**63.3**	**91.8**
Qwen3-8B-Base	12.8	46.7	13.4	43.3	9.0	33.3	5.8	27.3	54.0	92.5	47.7	85.1
GRPO	15.4	40.0	13.4	40.0	10.5	33.3	5.5	24.2	57.3	87.5	58.6	88.1
EntropyAdv	15.9	33.3	15.2	43.3	9.9	40.0	6.0	27.3	57.8	87.5	59.2	85.8
HAPO	15.6	40.0	12.9	26.7	8.4	36.7	5.1	27.3	61.5	87.5	57.4	87.3
80/20	17.2	46.7	15.1	33.3	13.9	40.0	8.5	27.3	60.6	92.5	63.1	87.3
RLRT	21.7	56.7	**20.1**	**46.7**	17.3	56.7	10.6	**30.3**	64.2	95.0	64.6	91.0
EAPO (Ours)	**24.1**	**63.3**	19.3	**46.7**	**17.9**	**60.0**	**11.2**	**30.3**	**65.7**	**97.5**	**65.9**	**95.5**
Qwen3-4B	67.9	83.3	61.7	86.7	60.6	**86.7**	40.1	66.7	95.2	**100.0**	90.0	95.5
GRPO	70.6	83.3	62.8	86.7	63.2	83.3	40.9	69.7	94.9	**100.0**	89.8	96.3
EntropyAdv	73.0	**86.7**	62.2	86.7	64.0	**86.7**	42.2	69.7	95.9	**100.0**	**90.6**	96.3
HAPO	72.0	**86.7**	64.1	83.3	62.0	83.3	41.6	66.7	94.9	**100.0**	89.9	96.3
80/20	71.4	**86.7**	63.2	83.3	61.0	**86.7**	40.8	66.7	94.3	**100.0**	90.5	96.3
RLRT	72.5	**86.7**	63.2	86.7	62.7	83.3	42.8	69.7	95.9	**100.0**	**90.6**	96.3
EAPO (Ours)	**73.5**	**86.7**	**65.0**	**90.0**	**64.8**	**86.7**	**43.2**	**72.7**	**97.7**	**100.0**	90.2	**97.0**
Olmo-3-7B-Think-DPO	74.1	90.0	60.1	80.0	66.3	**90.0**	42.7	72.7	95.5	**100.0**	90.1	**96.3**
GRPO	75.9	**93.3**	62.2	83.3	68.4	**90.0**	**46.0**	69.7	95.8	**100.0**	91.4	**96.3**
EntropyAdv	74.1	**93.3**	64.3	83.3	69.4	**90.0**	45.7	72.7	94.6	**100.0**	90.4	**96.3**
HAPO	74.2	**93.3**	63.3	86.7	66.4	**90.0**	43.8	69.7	95.7	**100.0**	90.6	**96.3**
80/20	75.6	**93.3**	65.1	83.3	69.4	**90.0**	44.9	75.8	96.0	**100.0**	91.3	**96.3**
RLRT	**77.3**	**93.3**	63.4	83.3	68.1	86.7	44.6	69.7	95.2	**100.0**	91.3	**96.3**
EAPO (Ours)	76.7	**93.3**	**65.3**	**90.0**	**70.3**	**90.0**	45.5	**78.8**	**96.2**	**100.0**	**91.8**	**96.3**

Baselines. We compare EAPO against the initial backbone and five RLVR methods, including GRPO (Shao et al., 2024). EntropyAdv (Cheng et al., 2026a) adds an entropy term to the advantage to encourage exploration, while HAPO (He et al., 2026) applies a bounded, sign-preserving adjustment based on normalized policy entropy to emphasize high-entropy positions. 80/20 (Forking Tokens) (Wang et al., 2025) restricts policy-gradient updates to the highest-entropy 20% of tokens. Finally, RLRT (Kim et al., 2026a) reverses the self-distillation signal on correct rollouts, reinforcing tokens more probable under the student than under a teacher conditioned on privileged information.

Implementation Details. We primarily evaluate EAPO across two model types, with Qwen3-4B-Base and Qwen3-8B-Base (Qwen Team, 2025) as base backbones and Qwen3-4B (Qwen Team, 2025) and Olmo-3-7B-Think-DPO (Olmo Team, 2025) as reasoning backbones. All optimization-based methods are trained on DAPO-Math-17k-Processed with a DAPO-style training configuration (Yu et al., 2025). Further experimental and implementation details are provided in Section C.

5.2 Main Results

Table 2 compares EAPO with baseline methods across six mathematical reasoning benchmarks. EAPO achieves the best overall performance, with the highest mean avg@32 and pass@32 for all four backbones. Specifically, with Qwen3-4B-Base and Qwen3-8B-Base, EAPO attains mean accuracies of 31.0% and 34.0%, surpassing the strongest entropy-based baselines (EntropyAdv and 80/20, respectively) by substantial margins of 5.6 and 4.3 percentage points. Notably, EAPO also outperforms the strongest baseline, RLRT, in both metrics on both base backbones even without relying on privileged information. Meanwhile, the gains extend to reasoning backbones, where EAPO achieves mean accuracies of 72.4% with Qwen3-4B and 74.3% with Olmo-3-7B-Think-DPO, surpassing the strongest baseline in each case despite their substantially stronger initial performance. These results demonstrate consistent gains across benchmarks and model types.

OOD Generalization. To examine whether these improvements extend beyond mathematical reasoning, we evaluate the two base backbones on eight tasks from Reasoning Gym (Stojanovski et al.,

Table 3: OOD generalization on Reasoning Gym tasks (avg@16). **Bold** denotes best performance.

Method	Graph	Cube	Sudoku	Family	K&K	Anag.	Zebra	Palin.	Avg.
Qwen3-4B-Base	6.75	24.63	40.00	17.00	29.13	37.50	29.50	9.00	24.19
GRPO	7.50	28.13	41.88	25.38	38.63	43.50	41.88	11.38	29.78
EntropyAdv	8.38	27.00	37.00	27.00	41.00	45.38	41.88	14.50	30.27
HAPO	10.13	27.13	40.38	30.13	43.63	**47.88**	47.50	14.13	32.61
80/20	11.38	26.75	39.25	**31.50**	49.63	46.38	47.38	15.13	33.43
RLRT	7.38	26.75	38.50	26.50	43.38	42.88	44.13	13.63	30.39
EAPO (Ours)	**12.88**	**28.25**	**42.13**	30.88	**52.88**	46.00	**48.38**	**17.75**	**34.89**
Qwen3-8B-Base	12.25	22.88	39.38	38.38	33.88	38.63	44.88	9.00	29.91
GRPO	18.50	25.63	40.25	42.50	56.88	54.00	49.50	19.13	38.30
EntropyAdv	16.75	22.25	47.88	44.25	49.38	55.63	54.25	18.13	38.56
HAPO	20.13	24.63	40.88	45.63	59.13	56.50	54.25	18.88	40.00
80/20	17.75	23.13	37.38	44.75	60.13	52.63	51.00	18.88	38.21
RLRT	19.00	22.38	46.25	40.75	48.50	43.88	51.75	18.75	36.41
EAPO (Ours)	**20.50**	**26.25**	**52.25**	**46.25**	**60.50**	**62.63**	**54.50**	**21.25**	**43.02**

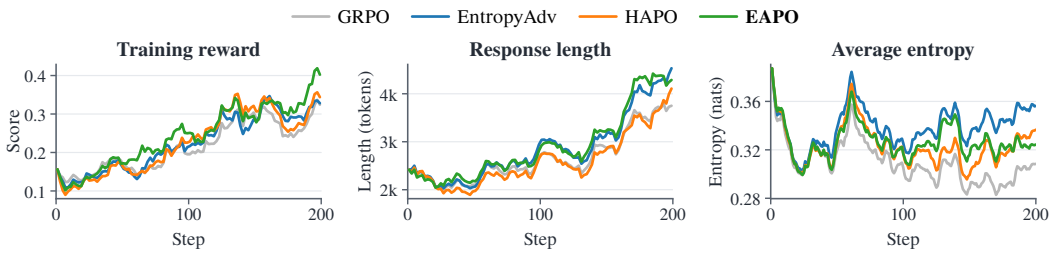

Figure 3: **Training dynamics.** *(Left)* Training reward on Qwen3-4B-Base; *(Center)* response length and *(Right)* average entropy on Qwen3-4B, using a trailing 15-step moving average.

2025), spanning logical, spatial, and algorithmic reasoning (with details of each task in Section C). As shown in Table 3, EAPO achieves the highest macro-averaged scores of 34.89% and 43.02% with Qwen3-4B-Base and Qwen3-8B-Base, improving upon the respective best baselines, 80/20 and HAPO, by 1.46 and 3.02 percentage points. These results suggest that the benefits of EAPO generalize to diverse reasoning tasks beyond the mathematical problems used for the optimization.

5.3 Training Dynamics

To understand how EAPO behaves during training, we examine the evolution of training reward, response length, and average entropy. Figure 3 shows that EAPO maintains higher training rewards than the baselines over most of training, with its advantage becoming particularly pronounced near the end, further supporting its superiority over the baselines even during training. Meanwhile, entropy-based RLVR approaches consistently exhibit an overall increase in response length, producing longer responses than GRPO. EAPO also follows this trend, sustaining longer reasoning trajectories as training progresses. However, longer responses alone do not establish more effective exploration. To examine whether response length explains the performance gains, we extend baseline responses following Muennighoff et al. (2025) and compare accuracy at comparable mean response lengths in Section D.3, where additional continuations do not close the gap to EAPO.

The average entropy exhibits an interesting trend throughout training. While retaining higher entropy than GRPO during later training, EAPO ends with lower average entropy than EntropyAdv and HAPO. A key distinction is that EAPO redistributes the original response-level advantage across tokens while preserving its within-response mean, whereas these baselines apply entropy-dependent additive adjustments. In particular, EntropyAdv adds a nonnegative entropy term to the advantage to encourage high-entropy actions and exhibits the highest final entropy in our experiments. EAPO, however, achieves stronger reasoning performance without a substantial increase in entropy, suggesting that higher average entropy is not a prerequisite for more effective exploration.

5.4 Analysis of Exploration

To assess whether EAPO's asymmetric design broadens exploration, we compare pass@k, response diversity, and epistemic marker frequency across different exploration methods.

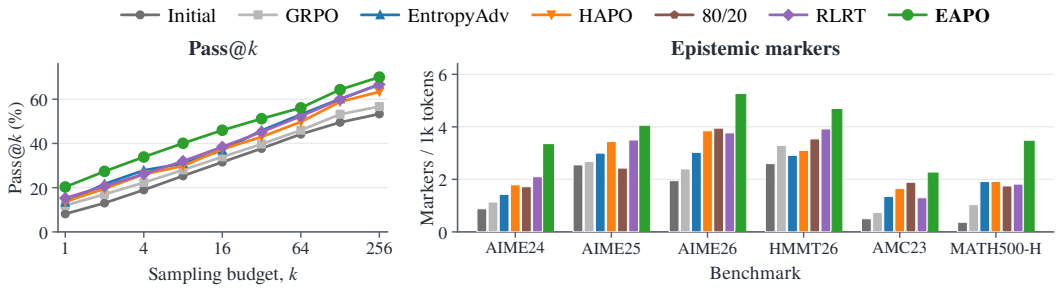

Figure 4: **Analysis of exploration.** *(Left)* Pass@k curves for Qwen3-4B-Base on AIME26. *(Right)* Frequency of epistemic markers (per 1,000 generated tokens) across six mathematics benchmarks.

Exploration across Sampling Budgets. We evaluate pass@k on AIME26 with Qwen3-4B-Base, varying the sampling budget from 1 to 256. As shown in Figure 4 (Left), entropy-based methods, EntropyAdv and HAPO, expand coverage over GRPO, but the gains remain modest. RLRT, which uses privileged information to guide exploration, also yields only a limited improvement over GRPO. In contrast, EAPO consistently achieves the highest pass@k across all sampling budgets, maintaining a substantial margin over GRPO and the other baselines. In particular, at $k = 256$, EAPO solves an additional problem that none of the baselines solve, suggesting that its exploration strategy broadens problem coverage even after extensive sampling.

Response Diversity & Exploratory Cues. We further analyze the generated responses to assess whether EAPO also promotes exploratory behavior during reasoning. Since precisely measuring diversity across full reasoning trajectories is impractical, we use final-answer diversity as a tractable proxy on challenging problems. Specifically, we consider all problems from the main Qwen3-4B-Base experiments, using $N = 32$ sampled responses per problem and selecting those with avg@32 $\leq 25\%$ under the initial model. We measure normalized answer entropy (H_{norm}) and the collision rate, the fraction of pairs of distinct sampled responses that produce the same final answer:

$$C = \frac{\sum_i n_i(n_i - 1)}{N(N-1)}, \tag{6}$$

where n_i counts responses with final answer i, with $\sum_i n_i = N$. Table 4 shows that the other RLVR methods exhibit final-answer diversity comparable to GRPO, whereas EAPO achieves higher normalized answer entropy and a lower collision rate. This suggests that EAPO maintains broader exploration on challenging problems, producing a more diverse range of candidate solutions even when they are incorrect. Such diversity offers more opportunities to discover correct solutions, consistent with the broader coverage in the pass@k analysis above.

Table 4: Normalized answer entropy and collision rate on low-accuracy problems.

Method	H_{norm} (↑)	**Collision** (↓)
GRPO	0.657	0.166
EntropyAdv	0.659	0.159
HAPO	0.647	0.173
80/20	0.672	0.161
RLRT	0.659	0.168
EAPO (Ours)	**0.751**	**0.101**

Beyond answer diversity, we further examine exploratory behavior within the generated responses via *epistemic markers*, tokens that signal exploratory behavior during reasoning (e.g., "wait"), following Kim et al. (2026b). Figure 4 (Right) shows that, while other exploration-oriented methods tend to generate epistemic tokens more frequently than GRPO, EAPO exhibits the highest frequency per 1,000 generated tokens across all six mathematical benchmarks by a substantial margin. This pattern suggests that EAPO's combination of exploration promotion and penalty attenuation encourages the generation of such cues, supporting continued exploration of alternative reasoning paths.

5.5 Ablation Studies

Effect of Sign–Entropy Coupling. To isolate how entropy preferences under reinforcement and penalization affect performance, we independently vary the allocation direction for positive- and negative-advantage responses. Specifically, we replace $\text{sign}(\hat{A}^i)$ in Equation (4) with b_+ when $\hat{A}^i > 0$ and b_- when $\hat{A}^i < 0$. Each coefficient is chosen from $\{-1, 0, +1\}$, corresponding to low-entropy preference, uniform credit, or high-entropy preference, respectively. Thus, $(b_+, b_-) = (+1, -1)$ recovers EAPO, whereas $(0, 0)$ yields uniform token credit. We evaluate all nine combinations on Qwen3-4B-Base, reporting macro-averaged scores across the six benchmarks. Table 5 shows that performance improves as the entropy preference shifts from low to high for positive advantages and

Table 5: **Ablation on sign–entropy coupling.** Macro average of avg@32 / pass@32 (%) across six benchmarks. Our default EAPO configuration is shaded. Best scores are **bolded**.

	Entropy	**Penalization** ($\hat{A}^i < 0$): b_-		
	Preference	Low (-1)	Uniform (0)	High ($+1$)
Reinforcement ($\hat{A}^i > 0$): b_+	High ($+1$)	**31.00 / 61.27**	26.73 / 53.83	25.36 / 52.09
	Uniform (0)	26.08 / 51.26	24.48 / 50.43	22.89 / 50.03
	Low (-1)	23.99 / 49.84	23.07 / 48.62	22.56 / 47.83

Table 6: **Sensitivity to κ.** We report avg@32 / pass@32 (%) across different values of κ.

κ	**AIME25**	**AIME26**
0	10.42 / 36.67	7.50 / 33.33
$\log 2$	10.73 / 33.33	10.63 / 26.67
$\log 4$	17.08 / **46.67**	15.73 / **50.00**
$\log 8$	**18.23** / **46.67**	15.83 / 40.00
$\log 16$	15.73 / 43.33	**16.15** / 46.67

from high to low for negative advantages. Specifically, fixing high-entropy reinforcement ($b_+ = +1$) and shifting penalization from high to low entropy improves accuracy by 5.64 percentage points. Consequently, $(b_+, b_-) = (+1, -1)$, the allocation used by EAPO, achieves the highest avg@32 and pass@32, improving avg@32 over uniform credit by 6.52 percentage points, which supports EAPO's asymmetric allocation directions for reinforcement and penalization.

Sensitivity to κ. EAPO introduces a single hyperparameter, κ, to control the concentration of entropy-guided token credit, with $\kappa = \log K$ allowing token weights within a response to differ by up to a factor of K. During training, we observe that larger κ generally leads to faster improvements but also greater instability, indicating a trade-off between credit concentration and optimization stability. We select $\kappa = \log 4$ as the default based on the observed training stability. Table 6 shows that increasing κ generally improves overall performance. As κ approaches zero, the allocation converges to uniform token credit (i.e., GRPO), and the accompanying performance drop supports using entropy to distinguish token contributions to reinforcement and penalization.

6 Related Work

Fine-Grained Credit Assignment. Recent work supplements sparse outcome-level rewards in RLVR with finer-grained signals from learned process rewards (Wang et al., 2024; Cui et al., 2026), intermediate value estimates (Kazemnejad et al., 2025; Guo et al., 2025b), or teacher- and oracle-conditioned likelihoods (Yang et al., 2026a; Li et al., 2026a). These often require auxiliary models, rollouts, or privileged information. Other approaches use the policy's internal signals, deriving token importance from outcome sensitivity (Li et al., 2026b; Pala et al., 2026) or combining entropy with correctness (Chen et al., 2026). Token-level advantages are also reweighted using selected-token probabilities or surprisal to modulate policy updates (Tang et al., 2026; Ding & Zhang, 2026).

RLVR for Exploratory Reasoning. To broaden exploration in RLVR, existing approaches allocate more samples to difficult problems (Yang et al., 2026b), revisit promising states (Dou et al., 2025), or branch over alternative continuations (Wu et al., 2026; Cao et al., 2026). Explicit diversity objectives encourage complementary outcomes and reasoning traces (Song et al., 2025; Hu et al., 2026; Bahlous-Boldi et al., 2026). Other methods regulate policy concentration (Cui et al., 2025; Hao et al., 2026; Zhang et al., 2026) or preserve low-probability alternatives (Huang et al., 2026). Token-level credit assignment supports exploration through reversed teacher signals (Kim et al., 2026a) or updates focused on high-entropy tokens (Wang et al., 2025; He et al., 2026; Cheng et al., 2026a). In contrast, EAPO couples policy entropy with the advantage sign for asymmetric credit assignment, without requiring auxiliary models, additional rollouts, or privileged information.

7 Conclusion

In this work, we proposed EAPO, an entropy-guided credit assignment method for improving exploration in LLM reasoning. By coupling entropy with the advantage sign, EAPO reinforces uncertain decisions in successful responses and penalizes confident decisions in unsuccessful responses, while attenuating penalties on uncertain alternatives. Evaluations on reasoning tasks demonstrate consistent overall improvements over RLVR baselines designed to promote exploration, across both base and reasoning models. Our analyses further show that EAPO effectively broadens problem coverage and increases response diversity. These findings highlight the value of asymmetric credit assignment for reinforcing ***surprising success*** and correcting ***repeated failure*** while sustaining exploration.

Ethics Statement

This work aims to improve exploration in LLM reasoning using existing models and reasoning datasets. However, models trained with EAPO may retain biases from their underlying models and data, and improved reasoning performance does not guarantee reliable or safe outputs. We therefore encourage appropriate safeguards when applying the method beyond the reasoning tasks studied here or in real-world applications.

Reproducibility Statement

To facilitate reproducibility, we describe the formulation and optimization procedure of EAPO in Section 4 and specify the training data, model backbones, baselines, benchmarks, and evaluation metrics in Section 5.1. We use a fixed random seed during training and additionally conduct experiments with multiple seeds to assess sensitivity to randomness, with further experimental details and results provided in Sections C and D.

AI Use Statement

We use large language models to improve the clarity and readability of the manuscript, assist with code implementation and debugging, and identify potentially overlooked related work. The core research ideas and the initial draft of the manuscript are entirely our own, and all AI-assisted artifacts are reviewed by the authors.

References

Ryan Bahlous-Boldi, Isha Puri, Idan Shenfeld, Akarsh Kumar, Mehul Damani, Sebastian Risi, Omar Khattab, Zhang-Wei Hong, and Pulkit Agrawal. Vector policy optimization: Training for diversity improves test-time search. In *COLM 2026 Workshop on Efficient Reasoning*, 2026.

Lang Cao, Hui Ruan, Yongqian Li, Peng Chao, Wu Ning, Haonan Song, Renhong Chen, and Yitong Li. Treeadv: Tree-structured advantage redistribution for group-based rl. *arXiv preprint arXiv:2601.03703*, 2026.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Xinzhu Chen, Xuesheng Li, Zhongxiang Sun, and Weijie Yu. Beyond high-entropy exploration: Correctness-aware low-entropy segment-based advantage shaping for reasoning LLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 32970–32984, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 30377–30385. AAAI Press, 2026a.

Guangran Cheng, Chengqi Lyu, Songyang Gao, Wenwei Zhang, and Kai Chen. Group entropy-controlled policy optimization. *arXiv preprint arXiv:2607.16850*, 2026b.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, Qixin Xu, Weize Chen, Jiarui Yuan, Huayu Chen, Kaiyan Zhang, Xingtai Lv, Shuo Wang, Yuan Yao, Xu Han, Hao Peng, Yu Cheng, Zhiyuan Liu, Maosong Sun, Bowen Zhou, and Ning Ding. Process reinforcement through implicit rewards. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Yi Ding and Ruqi Zhang. Does on-policy distillation really distill? from noisy teacher to self-improvement. *arXiv preprint arXiv:2608.31046*, 2026.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47, 1975.

Shihan Dou, Muling Wu, Jingwen Xu, Rui Zheng, Tao Gui, and Qi Zhang. Improving RL exploration for LLM reasoning through retrospective replay. In *Natural Language Processing and Chinese Computing - 14th National CCF Conference, NLPCC 2025, Urumqi, China, August 7-9, 2025, Proceedings, Part I*, volume 16102 of *Lecture Notes in Computer Science*, pp. 594–606. Springer, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655(8122):487–496, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025a.

Yiran Guo, Lijie Xu, Jie Liu, Dan Ye, and Shuang Qiu. Segment policy optimization: Effective segment-level credit assignment in RL for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025b.

Zhezheng Hao, Hong Wang, Haoyang Liu, Jian Luo, Jiarui Yu, Hande Dong, Qiang Lin, Can Wang, and Jiawei Chen. Rethinking entropy interventions in RLVR: An entropy change perspective. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31105–31133, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Yuhang He, Haodong Wu, Siyi Liu, Hongyu Ge, Hange Zhou, Keyi Wu, Zhuo Zheng, Qihong Lin, Zixin Zhong, and Yongqi Zhang. Where hindsight credit can reside: A signed-capacity view of token updates in rlvr. *arXiv preprint arXiv:2604.11056*, 2026.

HMMT. Harvard–mit mathematics tournament (HMMT) problems and solutions, 2026. URL https://www.hmmt.org/www/archive/problems.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*, 2022.

Zican Hu, Shilin Zhang, Yafu Li, Jianhao Yan, Xuyang Hu, Leyang Cui, Xiaoye Qu, Chunlin Chen, Yu Cheng, and Zhi Wang. Diversity-incentivized exploration for versatile reasoning. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*, 2026.

Guanhua Huang, Tingqiang Xu, Mingze Wang, Qi Yi, Xue Gong, Siheng Li, Ruibin Xiong, Kejiao Li, Yuhao Jiang, and Bo Zhou. Low-probability tokens sustain exploration in reinforcement learning with verifiable reward. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24158–24188, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, Alessandro Sordoni, Siva Reddy, Aaron C. Courville, and Nicolas Le Roux. Vineppo: Refining credit assignment in RL training of llms. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, 2025.

Jeonghye Kim, Jiwon Jeon, Dongsheng Li, and Yuqing Yang. Rebellious student: Reversing teacher signals for reasoning exploration with self-distilled RLVR. In *COLM 2026 Workshop on Efficient Reasoning*, 2026a.

Jeonghye Kim, Xufang Luo, Minbeom Kim, Sangmook Lee, Dongsheng Li, and Yuqing Yang. Understanding reasoning in llms through strategic information allocation under uncertainty. *arXiv preprint arXiv:2603.15500*, 2026b.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pp. 611–626. ACM, 2023.

Hynek Kydlíček. Math-verify: Math verification library, 2025. URL https://github.com/huggingface/math-verify.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling, COLM 2025, Montreal, Canada, October 7-10, 2025*, 2025.

Yu Li, Rui Miao, Tian Lan, and Zhengling Qi. Oppo: Bayesian value recursion for token-level credit assignment in llm reasoning. *arXiv preprint arXiv:2605.21851*, 2026a.

Ziheng Li, Liu Kang, Feng Xiao, Luxi Xing, Qingyi Si, Zhuoran Li, Weikang Gong, Deqing Yang, Yanghua Xiao, and Hongcheng Guo. Outcome-grounded advantage reshaping for fine-grained credit assignment in mathematical reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24681–24693, San Diego, California, United States, July 2026b. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

MAA. American mathematics competitions (AMC) 12 problems and solutions, 2023. URL https://artofproblemsolving.com/wiki/index.php/AMC_12_Problems_and_Solutions.

MAA. American invitational mathematics examination (AIME) problems and solutions, 2026. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1928–1937, 2016.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 20275–20321, Suzhou, China, November 2025. Association for Computational Linguistics.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Olmo Team. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Tej Deep Pala, Vernon Toh, and Soujanya Poria. Grail: Gradient-reweighted advantages for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2606.04889*, 2026.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

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.

Yuda Song, Julia Kempe, and Rémi Munos. Outcome-based exploration for LLM reasoning. In *NeurIPS 2025 Workshop: Second Workshop on Aligning Reinforcement Learning Experimentalists and Theorists*, 2025.

Zafir Stojanovski, Oliver Stanley, Joe Sharratt, Richard Jones, Abdulhakeem Adefioye, Jean Kaddour, and Andreas Köpf. Reasoning gym: Reasoning environments for reinforcement learning with verifiable rewards. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Xinyu Tang, Yuliang Zhan, Zhixun Li, Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Rethinking sample polarity in reinforcement learning with verifiable rewards. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2928–2954, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. Trl: Transformers reinforcement learning, 2020. URL https://github.com/huggingface/trl.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce LLMs step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xiong-Hui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for LLM reasoning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Fang Wu, Weihao Xuan, Heli Qi, Aaron Tu, Ximing Lu, Li Erran Li, and Yejin Choi. Deepsearch: Overcome the bottleneck of reinforcement learning with verifiable rewards via tree-based search. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*, 2026.

Can Xie, Ruotong Pan, Xiangyu Wu, Zhang Yunfei, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in RLVR: Uncertainty-aware advantage shaping for deeper reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19057–19076, San Diego, California, United States, July 2026. Association for Computational Linguistics.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026a.

Zhicheng Yang, Zhijiang Guo, Yinya Huang, Yongxin Wang, Dongchun Xie, Hanhui Li, Yiwei Wang, Xiaodan Liang, and Jing Tang. Depth-breadth synergy in RLVR: unlocking LLM reasoning gains with adaptive exploration. In *Forty-third International Conference on Machine Learning, ICML 2026, Seoul, Korea (South), July 6-11, 2026*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Jiazheng Zhang, Ziche Fu, Junrui Shen, Yunbin Zhao, Yunke Zhang, Zhiheng Xi, Long Ma, Chenxin An, Zhihao Zhang, Shichun Liu, et al. Entropy polarity in reinforcement fine-tuning: Direction, asymmetry, and control. *arXiv preprint arXiv:2605.11775*, 2026.

A Limitations and Future Directions

EAPO uses the policy's own entropy as an intrinsic signal for token-level credit assignment, allowing it to guide exploration without auxiliary models, additional information, or extra compute. However, this signal reflects the model's learned preferences and confidence, making the resulting allocation sensitive to its existing biases. The reliability of entropy-guided credit assignment therefore depends on how well the model's uncertainty aligns with the actual contribution of individual reasoning decisions to the final outcome.

While our work effectively expands exploration in LLM reasoning, our analysis of the underlying mechanism focuses on alternative continuations from a given reasoning state, leaving the explicit combination of findings across different states or trajectories unexplored. Recent advances in scientific discovery, however, illustrate the value of maintaining diverse candidate solutions and iteratively reusing or combining their useful components to generate new hypotheses and algorithms (Novikov et al., 2025; Gottweis et al., 2026). Motivated by these approaches, an important direction for future work is to investigate how exploration can be strengthened in such iterative discovery settings.

B Extended Related Work

Uncertainty-Guided RLVR. Many existing RLVR approaches concentrate learning on high-entropy regions or seek to increase policy entropy to promote exploration. Specifically, 80/20 (Wang et al., 2025) restricts policy gradients to the highest-entropy 20% of tokens, while HAPO (He et al., 2026) increases advantage magnitudes at high-entropy positions under both advantage signs. EntropyAdv (Cheng et al., 2026a) and UCAS (Xie et al., 2026) promote exploration by adding a non-negative entropy bonus or subtracting a certainty-based penalty from token advantages, respectively, while limiting token-level shaping to one-sided additive adjustments. LESS (Chen et al., 2026), on the other hand, targets confident reasoning patterns in low-entropy segments, reinforcing those associated with success and suppressing those associated with failure. Recent approaches also regulate updates to low-probability tokens, protecting useful alternatives from excessive suppression (Huang et al., 2026) or suppressing low-probability branches to favor more confident reasoning alternatives (Ding & Zhang, 2026). These approaches weight or filter tokens by uncertainty without distinguishing between reinforcement and penalization. In contrast, EAPO couples policy entropy with the response advantage sign to reverse this preference, strengthening uncertain successes while concentrating correction on confident positions in failed responses and attenuating penalties at uncertain positions to preserve exploratory alternatives.

Outcome-Conditioned Credit Assignment. Several recent approaches also seek to promote exploration by assigning credit differently to successful and failed responses. GEPO (Cheng et al., 2026b) attenuates positive advantages in low-entropy response groups to reduce over-exploitation and negative advantages in high-entropy groups to preserve exploration, while leaving token-level credit uniform within each response. RLRT (Kim et al., 2026a) reweights token-level advantages using reversed teacher–student likelihood ratios to reinforce successful exploration, but requires privileged context and additional teacher forward passes while leaving failed responses uniformly weighted. A3PO (Tang et al., 2026) leverages sampled-token probability to amplify advantage magnitudes for selected tokens, leaving the advantages of the remaining tokens unchanged. However, this can increase the overall advantage magnitude without attenuating penalties on potentially exploratory decisions, and sampled-token probability does not fully characterize predictive uncertainty, as replacing entropy with surprisal destabilizes asymmetric credit allocation (Section D.1). Our approach, EAPO, uses full-vocabulary policy entropy to redistribute advantage across all token positions while preserving its within-response mean. By conditioning token weights on entropy and the response advantage sign, EAPO both reinforces uncertain successes and preserves exploratory alternatives in failed responses without privileged context or additional teacher forward passes.

C Additional Experimental Details

C.1 Additional Details on the Resampling Experiment

We provide further details on the experimental setup and selection statistics for the high- and low-entropy windows used in the resampling experiment in Section 3. Specifically, for each of Qwen3-4B-Base and Qwen3-8B-Base, we select 20 moderately difficult problems from AIME24/25/26, HMMT26, and AMC23, where moderate difficulty is defined as 6–14 correct responses out of 32 attempts by the corresponding model. For each problem, we randomly select two correct and two incorrect responses from these attempts. Within each selected response, we exclude the first and last 20% of tokens and consider pairs of non-overlapping 32-token windows whose starting positions differ by at most 20% of the response length. We select the pair with the largest difference in mean token entropy. We then fix the original prefix preceding each selected window and sample 64 continuations, measuring overlap as the mean pairwise common-prefix length over their first 32 generated tokens and reward as their final-answer accuracy.

We additionally analyze the relative starting positions of the selected high- and low-entropy windows for each model, expressed as a percentage of the original response length. Table 7 reports their means and standard deviations. Across both models, the high- and low-entropy windows have mean starting positions near the middle of the responses, with neither type consistently preceding the other, suggesting that window selection does not systematically favor earlier high-entropy windows or later low-entropy windows.

Table 7: Starting positions of selected windows (% of response length; mean $\pm$ std.).

Model	High entropy	Low entropy
Qwen3-4B-Base	47.29 ± 6.20	53.73 ± 7.11
Qwen3-8B-Base	51.44 ± 8.02	49.86 ± 6.42

C.2 Benchmark Details

We evaluate on 297 problems across six mathematical reasoning benchmarks: AIME24, AIME25, and AIME26 (MAA, 2026), with 30 problems each; HMMT26 (HMMT, 2026), with 33 problems; AMC23 (MAA, 2023), with 40 problems; and MATH500-H (Lightman et al., 2024), the level-5 subset of MATH500 containing 134 problems. We automatically grade the generated answers against the reference answers using Math-Verify (Kydlíček, 2025). We sample 32 responses per mathematical reasoning problem and report avg@32 as the mean response accuracy and pass@32 as the probability of generating at least one correct response, using the unbiased estimator of Chen et al. (2021) for pass@k. We average scores over problems within each benchmark and give equal weight to the six benchmarks when reporting an overall mean.

To assess out-of-distribution generalization, we evaluate on eight tasks from Reasoning Gym (Stojanovski et al., 2025), covering logical, spatial, and algorithmic reasoning. The task headers in Table 3 correspond to graph coloring (Graph), color cube rotation (Cube), mini sudoku (Sudoku), family relationships (Family), knights & knaves (K&K), group anagrams (Anag.), zebra puzzles (Zebra), and palindrome generation (Palin.). We use the "easy" task configurations by default, but modify the settings for three tasks: graph coloring uses 12–14 vertices and an edge probability of 0.15, mini sudoku uses 4–6 empty cells, and zebra puzzles use three people and three characteristics. For each task, we evaluate 50 problems, sampling 16 responses per problem and using the task-specific verifier to assess full correctness, without counting partial credit as success. We report avg@16 within each task and compute the macro average by assigning equal weight to all eight tasks before rounding.

C.3 Training Hyperparameters

Table 8 summarizes the shared training hyperparameters for EAPO and the baselines, including the optimizer, batching, generation, and adapter settings. We train on DAPO-Math-17k-Processed using a DAPO-style configuration (Yu et al., 2025), including asymmetric clipping, token-level loss aggregation, and a soft overlong penalty. Unless otherwise stated, all optimization-based methods use LoRA (Hu et al., 2022) and are optimized with AdamW (Loshchilov & Hutter, 2019). We implement training with TRL (von Werra et al., 2020) and use vLLM (Kwon et al., 2023) for rollout generation. All experiments are conducted using a single NVIDIA H200 GPU. Parameters not listed in the table follow the defaults of the corresponding implementations.

Table 8: Training hyperparameters shared by EAPO and the baselines. Parameters not listed here follow the defaults of the original implementations.

Category	**Parameter**	**Value**
Data	Maximum response length	10,240 (base); 38,912 (reasoning)
	Soft overlong penalty onset	8,192 (base); 32,768 (reasoning)
Batching	Prompts per rollout iteration	16
	Responses per prompt (G)	8
	Generation batch size	128
	Mini-batch size	64
	Rollout iterations	100
Optimization	Optimizer	AdamW
	Learning rate	1×10^{-5}
	Learning-rate schedule	Constant with warmup
	Warmup steps	10
	Weight decay	0.01
	Maximum gradient norm	1.0
	Random seed	42
LoRA	Rank (r)	32
	Alpha (α)	64
	Dropout	0
	Target modules	Attention and MLP projections
Generation	Inference engine	vLLM
	Temperature / top-p	1.0 / 1.0
	Top-k filtering	Disabled
Policy loss	Clipping thresholds $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$	$(0.2, 0.28)$
	Loss aggregation	Token mean
	KL penalty coefficient	0
	Entropy bonus coefficient	0
	Importance sampling	Token-level TIS (cap $= 2.0$)
	Advantage std normalization	Group
EAPO	Concentration (κ)	$\log 4$
	ε_H	10^{-8}

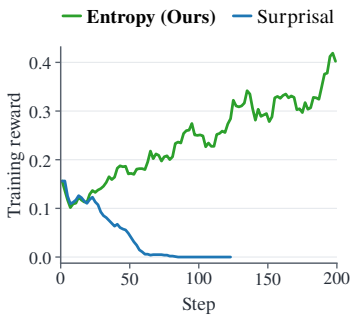

Figure 5: Entropy vs. Surprisal as signals for credit allocation.

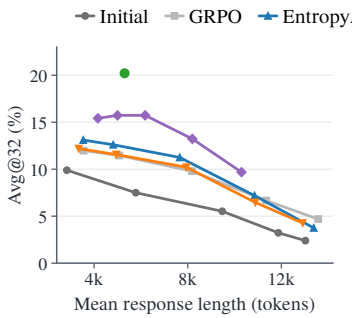

Figure 6: Response length vs. accuracy with budget forcing.

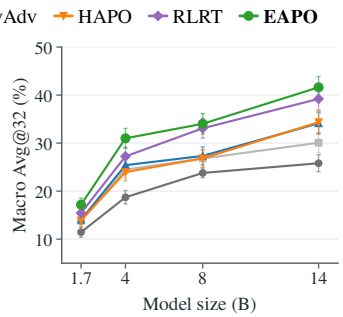

Figure 7: Scaling with model size in the Qwen3-Base family.

D Additional Experimental Results

D.1 Entropy vs. Surprisal

We examine whether sampled-token surprisal can substitute for policy entropy in EAPO. Specifically, the surprisal-based variant uses the normalized negative log-probability of the sampled token,

$$\mathcal{S}_{i,t} = -\log \pi_{\text{old}}(y_{i,t} \mid x, y_{i,<t}), \tag{7}$$

in place of entropy as the signal for the sign-dependent credit allocation in Equation (4).

Figure 5 shows the training dynamics of entropy-based EAPO and its surprisal-based variant on Qwen3-4B-Base. EAPO continues to improve its training reward, whereas the surprisal-based variant shows declining reward after a few training steps and eventually collapses to zero reward. Although the expectation of surprisal under the policy equals entropy at a fixed prefix, sampling a low-probability token can yield high surprisal even under a concentrated policy. Consequently, surprisal-based reweighting may conflate rare token selections with uncertain decisions, potentially destabilizing credit allocation. In contrast, entropy summarizes the full next-token distribution independently of which token is sampled. These results suggest that the sampled-token surprisal of A3PO (Tang et al., 2026) is less suitable than entropy for dense asymmetric advantage redistribution.

D.2 Extended Ablation on Sign–Entropy Coupling

We provide an additional ablation on sign–entropy coupling with Qwen3-8B-Base, extending the analysis in Section 5.5. As shown in Table 9, performance improves as the entropy preference shifts from high to low for penalization and from low to high for reinforcement, mirroring the trends observed with Qwen3-4B-Base. Specifically, with high-entropy reinforcement fixed ($b_+ = +1$), shifting penalization from high to low entropy improves avg@32 by 4.25 percentage points. Similarly, with low-entropy penalization fixed ($b_- = -1$), shifting reinforcement from low to high entropy improves avg@32 by 6.22 percentage points, indicating that both directions contribute to the gains of EAPO. Consequently, the default EAPO configuration $(b_+, b_-) = (+1, -1)$, which combines both preferences, achieves the highest avg@32 and pass@32 among all nine combinations. These results further support EAPO's asymmetric allocation directions for reinforcement and penalization.

Table 9: **Sign–entropy coupling on Qwen3-8B-Base.** Macro average of avg@32 / pass@32 (%) across six benchmarks. Our default EAPO configuration is shaded. Best scores are **bolded**.

	Entropy	**Penalization** ($\hat{A}^i < 0$): b_-		
	Preference	Low (-1)	Uniform (0)	High ($+1$)
Reinforcement ($\hat{A}^i > 0$): b_+	High ($+1$)	**34.02 / 65.55**	32.14 / 63.62	29.77 / 52.40
	Uniform (0)	30.55 / 59.52	26.78 / 52.18	27.48 / 52.56
	Low (-1)	27.80 / 54.68	25.26 / 50.27	24.63 / 49.82

Table 10: LoRA vs. full fine-tuning on Qwen3-4B-Base across mathematical reasoning benchmarks.

Method	Mode	AIME24		AIME25		AIME26		HMMT26		AMC23		MATH500-H	
		Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32	Avg@32	Pass@32
GRPO	LoRA	12.5	33.3	10.4	36.7	7.5	33.3	5.9	24.2	**54.5**	90.0	56.1	85.1
	Full	**13.1**	**36.7**	**12.8**	36.7	**10.2**	33.3	**6.9**	**27.3**	53.7	**92.5**	**58.2**	**87.3**
EntropyAdv	LoRA	**13.1**	**46.7**	**12.5**	36.7	10.0	33.3	6.3	27.3	**53.8**	**90.0**	56.7	**86.6**
	Full	13.0	36.7	11.0	36.7	10.0	**36.7**	**7.7**	**30.3**	52.5	87.5	**57.2**	84.3
HAPO	LoRA	**12.2**	36.7	**11.0**	40.0	8.9	30.0	**6.5**	**24.2**	52.0	87.5	53.2	**85.1**
	Full	11.9	36.7	10.5	40.0	8.9	30.0	6.3	21.2	**55.3**	**92.5**	**54.9**	82.8
80/20	LoRA	**11.9**	33.3	**12.7**	40.0	**8.6**	30.0	**6.6**	24.2	54.1	**95.0**	52.8	84.3
	Full	11.6	33.3	9.2	40.0	6.9	30.0	6.1	**27.3**	**55.3**	90.0	**53.0**	**87.3**
RLRT	LoRA	15.3	50.0	12.7	33.3	**14.2**	**40.0**	**8.9**	**36.4**	54.9	**95.0**	57.4	86.6
	Full	**19.1**	**56.7**	**13.9**	**36.7**	14.0	36.7	8.5	30.3	**57.0**	92.5	**59.7**	**88.1**
EAPO (Ours)	LoRA	**20.2**	**53.3**	17.1	**46.7**	**15.7**	**50.0**	**13.4**	33.3	56.3	92.5	**63.3**	91.8
	Full	18.3	50.0	**18.2**	43.3	15.2	46.7	12.3	**36.4**	**57.7**	92.5	62.8	91.8

D.3 Performance under Matched Token Budgets

We examine whether longer responses alone explain the performance gains of EAPO, extending baseline generations with *budget forcing* following Muennighoff et al. (2025). Specifically, we evaluate Qwen3-4B-Base on all 30 AIME24 problems with 32 responses per problem, comparing the initial model, GRPO, EntropyAdv, HAPO, and RLRT against EAPO without forced continuations. For each baseline, we remove the termination token (or `</think>`) and append `Wait` to resume generation, allowing up to $K \in \{0, 1, 2, 3, 4\}$ continuations under a total output budget of 32,768 tokens. We use the same sampling settings as during training.

Figure 6 suggests that additional continuations do not necessarily translate into more effective exploration. Specifically, EAPO achieves an avg@32 of 20.2% with a response length of 5.3k tokens, whereas the strongest baseline, RLRT, peaks at 15.7% with a mean response length of approximately 6.2k tokens. Most approaches show little benefit from early continuations and lose accuracy as responses are extended further. These results suggest that the performance gain of EAPO stems from enhanced exploration, with increased response length arising as a consequence of this process.

D.4 Scaling with Model Size

To examine the scaling behavior of EAPO as model size increases, we evaluate the Qwen3-Base family from 1.7B to 14B parameters, reporting macro-averaged avg@32 across the six mathematical reasoning benchmarks with ± 1 standard error bars.[1] As shown in Figure 7, performance improves with model size, while the relative performance trends among methods remain broadly consistent across scales, with EAPO achieving the highest mean accuracy throughout. These results suggest that the benefits of asymmetric entropy-guided credit assignment persist as model capacity increases.

D.5 LoRA vs. Full Fine-Tuning

Exploration can depend on retaining useful low-probability alternatives, motivating us to examine whether the benefits of entropy-guided credit assignment depend on the fine-tuning strategy. Our main experiments use LoRA (Hu et al., 2022), which constrains weight updates to a low-rank form and may therefore produce different changes in token probabilities from full fine-tuning. To assess whether EAPO's gains also hold under full fine-tuning, we compare EAPO and the baselines under LoRA and full fine-tuning on Qwen3-4B-Base across the six mathematical reasoning benchmarks. To ensure comparable training conditions, we use a learning rate of 1×10^{-6} for full fine-tuning while keeping all other experimental settings identical to those used for LoRA.

Table 10 shows that LoRA and full fine-tuning yield similar overall performance, with absolute differences in macro-averaged avg@32 of less than one percentage point on average across methods.

[1] We compute standard error as $B^{-1}\sqrt{\sum_{b=1}^{B} s_b^2/N}$, where B is the number of benchmarks (6), N is the number of generations (32), and s_b is the sample standard deviation of benchmark b's accuracy over generations.

Moreover, EAPO maintains the best overall performance under both strategies, with its margin over the strongest baseline exceeding these shifts in performance. These results suggest that the benefits of asymmetric entropy-guided credit assignment are robust to the choice of fine-tuning strategy.

D.6 Performance across Random Seeds

To evaluate whether the performance gains of EAPO are robust to random seed selection, we evaluate GRPO, EntropyAdv, HAPO, and EAPO on Qwen3-4B-Base across three distinct random seeds, extending our main experiment (which used a fixed seed of 42). Figure 8 reports the mean training reward and standard deviation across the three runs, with a 15-step moving average applied to the curves. EAPO maintains higher mean reward than all three baselines over most of training, and the gap becomes particularly pronounced in the final stages, where its standard-deviation band lies above those of the baselines. This separation suggests that the reward improvement is substantial relative to the observed variation across seeds. Overall, these results demonstrate the consistency of EAPO's gains beyond single-seed evaluation.

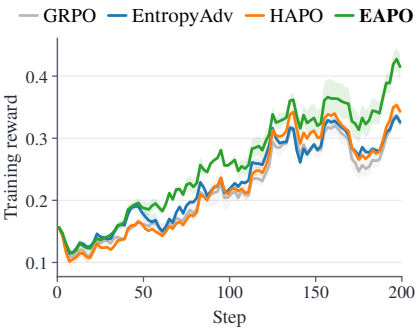

Figure 8: Training reward across three seeds (mean ± standard deviation).

E Proofs and Additional Theoretical Analysis

We provide the proofs of Propositions 1 and 2 (Sections E.1 and E.2) and present an information-theoretic motivation for entropy-guided allocation under a binary decision model (Section E.3).

E.1 KL-Regularized Uncertainty Allocation

The exponential allocation in Equation (4) is motivated by extending KL anchoring from policy optimization (Shao et al., 2024) to token-level credit allocation, favoring greater signed uncertainty while anchoring credit to a uniform reference. We formalize this motivation through Equation (5), where u_i is the reference distribution and $\mathrm{sign}(\hat{A}^i)h_{i,t}$ is the utility. The resulting optimization is a finite-dimensional instance of the Gibbs variational principle (Donsker & Varadhan, 1975).

Proposition 1 (Restated). *Let $u_{i,t} = 1/T_i$ be uniform credit over $\mathcal{I}_i$, and let $\Delta(\mathcal{I}_i)$ denote the probability simplex on these positions. For $\kappa > 0$, the problem*

$$\max_{q \in \Delta(\mathcal{I}_i)} \left\{ \mathrm{sign}(\hat{A}^i)\mathbb{E}_{t\sim q}[h_{i,t}] - \frac{1}{\kappa} D_{\mathrm{KL}}(q\|u_i) \right\} \tag{8}$$

has the unique solution $q^*_{i,t} \propto u_{i,t} \exp\Big(\kappa\, \mathrm{sign}(\hat{A}^i)h_{i,t}\Big)$.

Proof. Let $s_i = \mathrm{sign}(\hat{A}^i)$ and define

$$Z_i = \sum_{t\in\mathcal{I}_i} u_{i,t} \exp(\kappa s_i h_{i,t}), \qquad q^*_{i,t} = \frac{u_{i,t}\exp(\kappa s_i h_{i,t})}{Z_i}. \tag{9}$$

For any $q \in \Delta(\mathcal{I}_i)$, the objective in Equation (8) can be written as

$$s_i\mathbb{E}_{t\sim q}[h_{i,t}] - \frac{1}{\kappa} D_{\mathrm{KL}}(q\|u_i) = \frac{\log Z_i}{\kappa} - \frac{1}{\kappa} D_{\mathrm{KL}}(q\|q^*). \tag{10}$$

Since $D_{\mathrm{KL}}(q\|q^*) \geq 0$, with equality if and only if $q = q^*$, the allocation q^* is the unique global maximizer. Moreover, $u_{i,t} = 1/T_i$ and Equation (4) give $q^*_{i,t} = w_{i,t}/T_i$. □

E.2 Concentration of Unchosen Alternatives

We examine how stronger negative updates change the distribution over unchosen tokens, providing a theoretical rationale for EAPO's penalty attenuation. At a fixed prefix, let $\boldsymbol{z}$ be the initial logits, $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$ the corresponding token distribution, and a the sampled token. For fixed $\hat{A} < 0$ and allocation weight w, we consider one policy-gradient step on $\boldsymbol{z}$:

$$\boldsymbol{z}^{(\beta)} = \boldsymbol{z} - \beta(\boldsymbol{e}_a - \boldsymbol{p}), \qquad \beta = \eta|\hat{A}|w \geq 0, \qquad \boldsymbol{p}^{(\beta)} = \mathrm{softmax}(\boldsymbol{z}^{(\beta)}), \tag{11}$$

where $\boldsymbol{e}_a$ is the one-hot vector for a, $\eta > 0$ is the effective step size, and β controls the strength of the negative update. We exclude the sampled token and renormalize the remaining probabilities to obtain the conditional distribution over unchosen tokens:

$$\nu_v^{(\beta)} = \frac{p_v^{(\beta)}}{1 - p_a^{(\beta)}}, \qquad v \in \mathcal{V} \setminus \{a\}. \tag{12}$$

Here, $p_v^{(\beta)}$ is the probability of token v after the update and $\nu = \nu^{(0)}$ is the initial conditional distribution. The entropy of $\nu^{(\beta)}$ measures concentration among the alternatives, while $D_{\mathrm{KL}}(\nu^{(\beta)} \| \nu)$ measures their deviation from the initial proportions.

Proposition 2 (Restated). *For a single negative logit update with finite logits and the initial distribution held fixed, the sampled-token probability strictly decreases with β, while $H(\nu^{(\beta)})$ is nonincreasing and $D_{\mathrm{KL}}(\nu^{(\beta)} \| \nu)$ is nondecreasing.*

Proof. Let us assume finite logits over a finite vocabulary with $|\mathcal{V}| \geq 2$, so all initial and updated probabilities are positive and $0 < p_a^{(\beta)} < 1$. For any $v \neq a$, subtracting the updated logits gives

$$\log \frac{p_a^{(\beta)}}{p_v^{(\beta)}} = \log \frac{p_a}{p_v} - \beta(1 - p_a + p_v). \tag{13}$$

Every coefficient $1 - p_a + p_v$ is positive, so each ratio $p_v^{(\beta)}/p_a^{(\beta)}$ strictly increases with β. Therefore, $p_a^{(\beta)} = (1 + \sum_{v \neq a} p_v^{(\beta)}/p_a^{(\beta)})^{-1}$ establishes strict decrease of the sampled-token probability.

For unchosen tokens, the update adds βp_v to each logit. Conditioning on this set cancels the full softmax normalizer, yielding

$$\nu_v^{(\beta)} = \frac{\nu_v e^{\beta p_v}}{Z_a(\beta)}, \qquad Z_a(\beta) = \sum_{u \neq a} \nu_u e^{\beta p_u}. \tag{14}$$

Let $\psi_a(\beta) = \log Z_a(\beta)$. Differentiation gives $\psi_a'(\beta) = \mathbb{E}_{v \sim \nu^{(\beta)}}[p_v]$ and $\psi_a''(\beta) = \mathrm{Var}_{v \sim \nu^{(\beta)}}(p_v)$. The same calculation gives $\frac{d}{d\beta}\mathbb{E}_{\nu^{(\beta)}}[\log \nu_v] = \mathrm{Cov}_{\nu^{(\beta)}}(\log \nu_v, p_v)$.

Writing $H(\nu^{(\beta)}) = -\mathbb{E}_{\nu^{(\beta)}}[\log \nu_v] - \beta\psi_a'(\beta) + \psi_a(\beta)$, we obtain

$$\frac{d}{d\beta} H(\nu^{(\beta)}) = -\mathrm{Cov}_{\nu^{(\beta)}}(\log \nu_v, p_v) - \beta \mathrm{Var}_{\nu^{(\beta)}}(p_v) \leq 0. \tag{15}$$

To verify the inequality, let V, U be independent draws from $\nu^{(\beta)}$. Then

$$\mathrm{Cov}_{\nu^{(\beta)}}(\log \nu_v, p_v) = \frac{1}{2}\mathbb{E}[(\log \nu_V - \log \nu_U)(p_V - p_U)] \geq 0, \tag{16}$$

because $\log \nu_v = \log p_v - \log(1 - p_a)$ is increasing in p_v.

Similarly, Equation (14) implies

$$D_{\mathrm{KL}}(\nu^{(\beta)} \| \nu) = \beta\psi_a'(\beta) - \psi_a(\beta), \qquad \frac{d}{d\beta} D_{\mathrm{KL}}(\nu^{(\beta)} \| \nu) = \beta \mathrm{Var}_{\nu^{(\beta)}}(p_v) \geq 0. \tag{17}$$

The entropy and KL remain unchanged when all alternatives have the same initial probability. □

E.3 Implications for Entropy-Guided Allocation

We further examine how feedback information varies with entropy under success and failure, providing an information-theoretic interpretation of EAPO's asymmetric credit allocation. Consider a binary decision with correct branch $B \in \{0, 1\}$ and belief $P_p(B = 1) = p \in (0, 1)$. We assume that the policy samples a independently of B, with $\Pr(a = 1) = p$, and receives feedback $R = \mathbf{1}\{a = B\}$. We define *feedback information* as the expected posterior correction conditional on the outcome:

$$J_r(p) = \mathbb{E}[D_{\mathrm{KL}}(P_p(B \mid a, R) \| P_p(B)) \mid R = r], \qquad r \in \{0, 1\}. \tag{18}$$

Proposition 3 (Asymmetric information from success and failure)**.** *For the binary decision model, $J_1(p)$ is strictly increasing in the binary entropy $H(p)$, whereas $J_0(p)$ is strictly decreasing in $H(p)$.*

Proof. Let $D = p^2 + (1 - p)^2$. The pair (a, R) identifies B, so the posterior KL is $-\log P_p(B)$. Conditional on success, the probabilities of $B = 1$ and $B = 0$ are p^2/D and $(1-p)^2/D$, whereas on failure both are $1/2$ because each mismatched pair has probability $p(1 - p)$. Substituting into Equation (18) gives $J_1(p) = -\frac{p^2 \log p + (1-p)^2 \log(1-p)}{D}$ and $J_0(p) = -\frac{1}{2}\log(p(1 - p))$. By symmetry and continuity, it suffices to consider $p > 1/2$, where $H'(p) = -\log \frac{p}{1-p} < 0$ and

$$J_1'(p) = -\frac{2p - 1}{D} - \frac{2p(1 - p)}{D^2} \log \frac{p}{1 - p} < 0, \qquad J_0'(p) = \frac{2p - 1}{2p(1 - p)} > 0. \tag{19}$$

□

Proposition 3 shows that success leads to a larger expected correction of the prior belief at uncertain decisions, whereas failure does so at confident decisions. This matches EAPO's preference for high entropy under success and low entropy under failure.

Uniform allocation gives every decision the same share of credit, while EAPO shifts credit toward decisions with greater expected feedback information. For T decisions with the same outcome r, let $J_t = J_r(p_t)$ and let q_t^{E} denote the fraction of credit assigned to decision t by Equation (4). Then

$$\mathbb{E}_{q^{\mathrm{E}}}[J_t] - \mathbb{E}_u[J_t] = \frac{\mathrm{Cov}_u\big(J_t, e^{\kappa s h_t}\big)}{\mathbb{E}_u[e^{\kappa s h_t}]} \geq 0, \tag{20}$$

where $u_t = 1/T$, $s = 2r - 1$, and h_t is the normalized entropy from Equation (3). The covariance is nonnegative because entropy normalization preserves the ordering and larger credit shares correspond to larger J_t. Thus, in this binary model, the average feedback information weighted by EAPO is larger than the uniform average when $\kappa > 0$ and the normalized entropies are nonuniform. Favoring high entropy under both outcomes would instead shift credit toward less informative decisions under failure. This supports EAPO's asymmetric credit allocation by showing that it prioritizes more informative decisions under both outcomes, while Section E.2 supports attenuating penalties at uncertain decisions to preserve alternatives for further exploration.

F EXAMPLE OF SURPRISING SUCCESS, REPEATED FAILURE

Figure 9 illustrates ***repeated failure*** and ***surprising success*** in 32 rollouts from the initial Qwen3-4B-Base model on a MATH500 problem: all 28 AM–GM rollouts fail, whereas one of four Lagrange-multiplier rollouts succeeds. Specifically, the model repeatedly applies AM–GM to x^4, $4y^2$, and $4z^4$, resulting in the invalid substitution $x^4y^2z^4 = (xyz)^4$. Yet a successful alternative remains within AM–GM: splitting $4y^2$ into $2y^2 + 2y^2$ allows the same inequality to establish the correct minimum of 16. Meanwhile, the less frequently attempted Lagrange-multiplier approach also reaches this minimum, illustrating a successful alternative that remains rare under the current policy and should be reinforced to make such success more repeatable.

Problem. Let x, y, and z be positive real numbers such that $xyz = 2$. Find the minimum value of

$$x^4 + 4y^2 + 4z^4.$$

Sample 0

We will apply the AM-GM Inequality to the terms x^4, $4y^2$, and $4z^4$:

$$[\ldots] \quad x^4 + 4y^2 + 4z^4 \geq 3\sqrt[3]{16x^4y^2z^4}, \qquad x^4y^2z^4 = (xyz)^4 = 2^4 = 16.$$

$$[\ldots] \quad x^4 + 4y^2 + 4z^4 = 4\sqrt{2} + 4\sqrt{2} + 4\sqrt{2} = \boxed{12}.$$

Sample 4

We start by applying the AM-GM inequality to the terms x^4, $4y^2$, and $4z^4$:

$$[\ldots] \quad x^4 + 4y^2 + 4z^4 \geq 3\sqrt[3]{16x^4y^2z^4}, \qquad x^4y^2z^4 = (xyz)^4 = 2^4 = 16.$$

[. . .] However, we need to recheck the equality condition and the value of k.

$$[\ldots] \quad x^4 + 4y^2 + 4z^4 = 2^4 + 4 \cdot 1^2 + 4 \cdot 1^4 = 16 + 4 + 4 = \boxed{24}.$$

Sample 10

To find the minimum value, we need to check if equality can be achieved in the AM-GM inequality. Equality holds when all the terms are equal:

$$x^4 = 4y^2 = 4z^4.$$

$$[\ldots] \quad x^4 = 4y^2 = 4z^4 = 4\sqrt{2}, \qquad x^4 + 4y^2 + 4z^4 = 3 \cdot 4\sqrt{2} = \boxed{12\sqrt{2}}.$$

Sample 5

We will use the method of Lagrange multipliers to find the minimum value of the function $f(x, y, z) = x^4 + 4y^2 + 4z^4$ subject to the constraint $g(x, y, z) = xyz - 2 = 0$.

$$-\frac{4x^3}{yz} = -\frac{8y}{xz} \implies x^4 = 2y^2, \qquad -\frac{4x^3}{yz} = -\frac{16z^3}{xy} \implies x^4 = 4z^4.$$

$$[\ldots] \quad (x, y, z) = (\sqrt{2}, \sqrt{2}, 1), \qquad x^4 + 4y^2 + 4z^4 = 4 + 8 + 4 = \boxed{16}.$$

Figure 9: **Qualitative example of *surprising success* and *repeated failure*.** Among 32 rollouts of this problem, 28 use the AM–GM inequality and all fail, while four use Lagrange multipliers and only one succeeds. The model repeatedly applies AM–GM and makes the invalid substitution $x^4y^2z^4 = (xyz)^4$, yet splitting $4y^2$ into $2y^2 + 2y^2$ would enable the same approach to reach the correct minimum of 16. The less frequently sampled Lagrange-multiplier approach yields the sole observed correct response, motivating reinforcement of this rare success to make it more repeatable.