# EAPO: ENTROPY-DRIVEN ADAPTIVE POSITIVE-NEGATIVE SAMPLE WEIGHTING FOR POLICY OPTIMIZATION IN OPEN-ENDED QA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Reasoning Models are typically trained via reinforcement learning from verifiable rewards (RLVR). However, existing approaches adopt fixed weights for positive and negative samples, and their behavior in open-ended question answering (QA) remains less explored.In this paper, we investigate the roles of positive and negative samples in reinforcement learning for open-ended QA. We propose a reward-mean-based strategy for distinguishing positive from negative samples, and observe that negative samples are more closely related to response diversity, whereas positive samples contribute more to response quality and convergence stability.Building on these observations, we propose **EAPO**, an **E**ntropy-driven **A**daptive **P**olicy **O**ptimization method that adaptively computes the weighting coefficients of positive samples based on the ratio of the current policy entropy to the initial entropy. During the entropy-decreasing phase, the weight assigned to positive samples is reduced to preserve exploration, whereas during the entropy-increasing phase it is amplified to reinforce stability. Experiments on two publicly available open-ended medical QA datasets show that EAPO achieves a better balance between response diversity and training stability than fixed-weight baselines.

## 1 INTRODUCTION

Large Reasoning Models (LRMs) have recently attracted considerable attention owing to their remarkable performance on mathematical(Mirzadeh et al., 2025; Wang et al., 2026), coding(Chen et al., 2025b; Yang et al., 2025), and scientific reasoning tasks(Yan et al., 2025; Phan et al., 2025). Such models are typically trained via reinforcement learning from verifiable rewards (RLVR) (Guo et al., 2025; Team et al., 2025; Huang et al., 2026; Wu et al., 2026; Xu et al., 2026a; Ingle et al., 2026; Xu et al., 2026b), where multiple long chain-of-thought trajectories are sampled and assigned a binary reward according to final-answer correctness, and the policy is then updated based on reward signal.

Recent studies(Tang et al., 2025; Zhu et al., 2026) have investigated the impact of positive and negative samples in RLVR. Zhu et al. (2026) finds that using only positive samples improves Pass@1 but harms diversity, leading to a decrease in Pass@k for higher k, whereas using only negative samples maintains exploration and yields stronger performance at high k. Tang et al. (2025) further explores adjusting the influence strength of positive and negative samples at different granularities (sample-level and token-level), thereby modulating the training process of RLVR.

However, existing policy optimization methods (Zhu et al., 2026) often employ fixed weights for positive and negative samples. Strong positive reinforcement can concentrate the policy distribution, whereas strong negative reinforcement can increase stochasticity; the resulting trade-off motivates adapting their relative weights during training. Although Tang et al. (2025) introduces asymmetric advantage shaping at the token level, its experimental analysis focuses on RLVR. Whether the same conclusions hold under continuous rewards in open-ended QA requires separate empirical evaluation. Open-ended QA can provide reference answers without a single binary correctness criterion, making sample polarity dependent on the chosen reward function.

In this paper, we investigate positive- and negative-sample learning on two Chinese open-ended medical QA datasets, RJUA and CMD. For each prompt, rollouts with rewards at or above the group



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.

mean have non-negative group-normalized advantages and are termed positive samples; those below the mean have negative advantages and are termed negative samples. This partition expresses relative reward, not absolute correctness. In the reported runs, the two groups exert different pressures on policy entropy and yield metric-dependent quality trade-offs. Our fixed-weight sweeps further show sensitivity to coefficient selection, while appropriately selected fixed weights remain competitive.

Building upon these observations, we propose **EAPO**, an **E**ntropy-driven **A**daptive positive-negative sample weighting method for **P**olicy **O**ptimization. EAPO uses the ratio of current to initial entropy to adjust the positive-sample weight online while keeping the negative-sample weight fixed. Its contribution is a lightweight entropy-feedback rule within sample-level asymmetric weighting. Across the evaluated medical QA settings, EAPO achieves competitive final performance and smaller score ranges under changes to initialization in three of four metric–dataset pairs. The strongest fixed-weight configuration on RJUA and NSR on CMD nevertheless obtain higher final average scores. We therefore position EAPO primarily as a practical approach to moderating training dynamics and reducing sensitivity to weight selection within the tested settings.

Our contributions are summarized as follows:

- We extend sample-polarity analysis to continuous-reward medical QA using the sign of group-normalized advantage, and examine the resulting entropy and evaluation-score trade-offs.
- We characterize how fixed asymmetric weights affect entropy trajectories and evaluation scores, documenting dataset- and metric-dependent trade-offs on medical QA.
- We propose EAPO, a lightweight entropy-feedback weighting rule, and report competitive medical QA performance with reduced initialization sensitivity in three of four evaluated metric–dataset pairs.

## 2 Related Work

**Reinforcement Learning for Large Reasoning Models.** Reinforcement learning has become a central paradigm for improving the reasoning ability of large language models. Early alignment methods commonly optimize language model policies with PPO-style objectives and KL constraints (Schulman et al., 2017). More recent reasoning-oriented methods adopt reinforcement learning with verifiable rewards (RLVR)(Chen et al., 2025a; Xu et al., 2026b; Xie et al., 2025; Zheng et al., 2025; Wang et al., 2025), where models generate multiple candidate reasoning trajectories and receive outcome-based rewards. DeepSeekMath (Shao et al., 2024) introduces GRPO, which estimates advantages from groups of sampled rollouts and removes the need for a separate value model. DeepSeek-R1 (Guo et al., 2025) further demonstrates that large-scale RL can elicit complex reasoning behaviors from base models. Follow-up systems such as DAPO (Yu et al., 2026) improve the stability and scalability of long-CoT RL through algorithmic and system-level refinements.

**Positive and Negative Sample Signals in RLVR.** Recent studies(Zhu et al., 2026; Tang et al., 2025; Arnal et al., 2026) have begun to revisit the role of sample polarity in RLVR. Rather than treating incorrect or low-reward rollouts as uninformative failures, negative samples can provide useful signals for exploration. Zhu et al. (2026) show that negative reinforcement can be surprisingly effective for LLM reasoning, preserving diversity and improving high-$k$ performance. Tang et al. (2025) further analyze positive and negative samples in RLVR and show that positive samples tend to sharpen existing correct reasoning patterns, whereas negative samples encourage exploration of alternative reasoning paths. These studies primarily analyze RLVR; their implications for the continuous-reward medical QA setting considered here require empirical examination.

**Entropy-guided Adaptive Policy Optimization.** Entropy has long been used as a proxy for policy exploration in reinforcement learning. Maximum-entropy RL encourages policies to maintain stochasticity and improves exploration robustness (Ziebart et al., 2008; Haarnoja et al., 2018; Tian et al., 2026), while PPO-style objectives often include entropy bonuses or KL penalties to prevent unstable policy updates (Schulman et al., 2017). In LLM reinforcement learning, entropy collapse often corresponds to shortened outputs, reduced diversity, and premature convergence, whereas excessive entropy may lead to unstable or low-confidence generations. EAPO builds on entropy-guided optimization and asymmetric sample weighting by using the current-to-initial entropy ratio

to adjust positive-sample weights online. Its distinguishing design choice is to adapt sample-level reinforcement strength using observed entropy, and we evaluate this rule on medical QA.

# 3 Effects of Positive and Negative Samples in Open-Ended QA

In this section, we first present the task formulation. We then introduce the proposed reward-mean-based strategy for distinguishing positive and negative samples in the open-ended QA reinforcement learning setting, together with the design of the reward function. Finally, we analyze the respective effects of positive and negative samples on reinforcement learning for open-ended QA.

## 3.1 Task Formulation

Given a prompt $x \in \mathcal{D}$, a language model policy $\pi_\theta$ generates a response $y$. For each prompt $x$, we sample $N$ rollouts $\{y_i\}_{i=1}^N \sim \pi_\theta(\cdot \mid x)$ with corresponding rewards $\{r_i\}_{i=1}^N$. Unlike mathematical reasoning tasks where rewards are binary, open-ended QA employs a continuous reward function $r(x, y)$ that evaluates answer quality along multiple dimensions. The training objective is to update $\theta$ via policy gradient to increase the probability of higher-quality responses.

We follow previous work(Tang et al., 2025; Zhu et al., 2026) and decompose the open-ended QA reinforcement learning objective into learning from positive-advantage and negative-advantage rollouts. Here, sample polarity is defined relative to the group reward mean rather than by correct or incorrect answers. The overall objective can thereby be expressed as the sum of two sub-objectives:

$$\mathcal{L}(\theta) = w^+ \mathcal{L}_{\mathrm{PSR}}(\theta) + w^- \mathcal{L}_{\mathrm{NSR}}(\theta), \tag{1}$$

where $w^+$ and $w^-$ denote the weights of positive and negative samples, respectively. The two sub-objectives correspond to the two learning paradigms described above:

$$\mathcal{L}_{\mathrm{PSR}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}} \left[ \sum_{y:\, r(x,y)\geq \bar{r}_x} \pi_\theta(y \mid x) \right] \tag{2}$$

$$\mathcal{L}_{\mathrm{NSR}}(\theta) = -\mathbb{E}_{x\sim\mathcal{D}} \left[ \sum_{y:\, r(x,y)< \bar{r}_x} -\pi_\theta(y \mid x) \right] \tag{3}$$

where $\mathcal{L}_{\mathrm{PSR}}(\theta)$ denotes *positive sample reinforcement* (PSR), and $\mathcal{L}_{\mathrm{NSR}}(\theta)$ denotes *negative sample reinforcement* (NSR). $\bar{r}_x$ denotes the mean reward over the rollouts within a group, whose formal definition is provided in Subsection 3.2.

## 3.2 Reward-Mean-Based Positive and Negative Sample Identification

Unlike RLVR, open-ended QA does not presuppose fixed ground-truth correctness criteria. However, in GRPO (Guo et al., 2025), the group-normalized advantage for rollout $i$ under prompt $x$ is:

$$A_i = \frac{r_i - \bar{r}_x}{\sigma_x + \epsilon}, \quad \bar{r}_x = \frac{1}{N}\sum_{j=1}^{N} r_j \tag{4}$$

$\sigma_x$ denotes the intra-group reward variance. The group mean $\bar{r}_x$ precisely serves as the critical boundary at which the direction of the reinforcement learning update reverses. Rollouts satisfying $r_i \geq \bar{r}_x$ are assigned a non-negative advantage $A_i \geq 0$ and are regarded as *positive samples*, indicating that the model is encouraged to generate such trajectories with higher probability. Conversely, rollouts with $r_i < \bar{r}_x$ are assigned a negative advantage $A_i < 0$ and are treated as *negative samples*, in which case the model is driven to reduce the likelihood of producing them. Formally, we define:

$$\mathcal{D}_x^+ = \{i \mid r_i \geq \bar{r}_x\}, \quad \mathcal{D}_x^- = \{i \mid r_i < \bar{r}_x\} \tag{5}$$

where $\mathcal{D}_x^+$ is the positive samples set and $\mathcal{D}_x^-$ is the negative samples set. The mean-based boundary is query-adaptive and follows the sign of the group-normalized advantage. It provides a relative partition under the selected reward function, without certifying the factual or clinical correctness of individual responses.

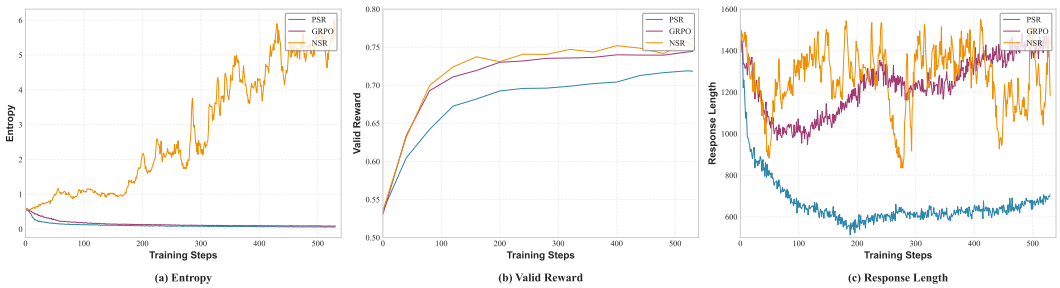


Figure 1: Training Dynamics on the RJUA dataset.

### 3.3 Reward Design

For the open-ended QA task, we design a composite reward combining four complementary signals:

$$r(x, y) = \alpha_1 r_{\text{fmt}}(y) + \alpha_2 r_{\text{rouge}}(y, \hat{d}) + \alpha_3 r_{\text{rerank}}(y, \hat{d}) + \alpha_4 r_{\text{llm}}(x, y, \hat{d}) \tag{6}$$

where $\hat{d}$ is the ground-truth response with weights $\alpha_1, \alpha_2, \alpha_3$ and $\alpha_4$.

**Format Reward** $r_{\text{fmt}}(y) \in \{0, 1\}$ checks whether the response contains all required structural tags.

**Rouge-L Reward** $r_{\text{rouge}}(y, \hat{d}) \in [0, 1]$ computes the character-level Rouge-L F1 score between the response and the ground-truth, measuring lexical overlap.

**Reranker Reward** $r_{\text{rerank}}(y, \hat{d}) \in [0, 1]$ uses a neural reranker to score the semantic relevance between the ground-truth $\hat{d}$ as query and the generated response $y$ as document, capturing semantic similarity beyond lexical overlap.

**LLM-as-a-Judge Reward** $r_{\text{llm}}(x, y, \hat{d}) \in [0, 1]$ evaluates the quality of the model's reasoning process within the `<think>` tag via a large language model judge. The detailed prompt and evaluation dimensions are provided in Appendix A.

### 3.4 Experimental Setup

We conduct experiments on Qwen3-8B (Team, 2025). Following the approach of W-REINFORCE (Zhu et al., 2026), we perform reinforcement learning separately using only positive samples and only negative samples for LLM, and include GRPO (Guo et al., 2025), which utilizes both types of samples. The experimental setups are given in Subsection 6. Details of the training and evaluation datasets are provided in Subsection E.1. Evaluation methods are shown in Subsection E.2.

### 3.5 Different Training Dynamics of Positive and Negative Sample Reinforcement

**PSR and NSR show contrasting entropy trends in the reported RJUA runs.** As shown in Figure 1(a), PSR is associated with rapidly decreasing entropy, whereas NSR maintains higher entropy. Reinforcing above-mean-reward rollouts can concentrate probability mass, while suppressing below-mean-reward rollouts can redistribute it toward alternatives. These observed trends are consistent with findings in W-REINFORCE (Zhu et al., 2026) and A3PO (Tang et al., 2025) under RLVR.

**NSR shows faster reward gains than PSR in the reported comparison.** As shown in Figure 1(b), PSR yields limited reward improvement, while NSR improves more quickly under the selected reward function.

**Response lengths differ between PSR and NSR in this setting.** As shown in Figure 1(c), PSR tends toward shorter responses, while NSR produces longer responses in the reported runs. Response length alone does not establish reasoning depth or answer quality, and this comparison does not imply that response length changes monotonically with sample weight.

Table 1: Performance comparison of different RL paradigms on RJUA test set. Avg: Average.

| Methods | Rouge-L | Reranker | LAAJ | Avg |
|---|---|---|---|---|
| GRPO | 0.312 | 0.976 | **0.744** | **0.677** |
| PSR | 0.293 | 0.940 | 0.738 | 0.657 |
| NSR | **0.327** | **0.971** | 0.702 | 0.667 |

Table 2: Performance comparison of different RL paradigms on CMD test set. Avg: Average.

| Methods | Rouge-L | Reranker | LAAJ | Avg |
|---|---|---|---|---|
| GRPO | 0.186 | 0.783 | 0.649 | **0.539** |
| PSR | 0.147 | 0.688 | **0.783** | **0.539** |
| NSR | **0.194** | **0.792** | 0.544 | 0.510 |

### 3.6 Multi-View Trade-offs of Positive and Negative Sample Reinforcement

We evaluate PSR, NSR, and GRPO on Qwen3-8B using Rouge-L, Reranker, and LLM-as-a-Judge (LAAJ), as shown in Tables 1 and 2. We use CMD to illustrate the metric-dependent trade-offs below; their magnitudes and method rankings depend on the dataset.

**PSR obtains higher LAAJ scores than NSR on CMD, with lower reference-based scores.** Compared with NSR, PSR has a 0.239 higher LAAJ score on CMD (0.783 versus 0.544), but lower Rouge-L and Reranker scores. This disagreement indicates a trade-off among evaluation criteria. Without independent judge validation or direct diversity measurements, it does not establish greater clinical quality or distributional collapse.

**NSR obtains higher Rouge-L and Reranker scores on CMD.** NSR has the highest Rouge-L and Reranker scores among the three methods on CMD, but a lower LAAJ score than PSR and GRPO. On RJUA, GRPO has a higher Reranker score than NSR. This pattern illustrates dataset-dependent trade-offs among the reported metrics.

The relative performance of PSR, NSR, and GRPO therefore depends on the evaluation criterion and dataset.

## 4 The Impact of Asymmetric Weights in Open-ended QA Reinforcement Training

The preceding analyses show different entropy trajectories and metric-dependent outcomes for PSR, NSR, and GRPO. We next examine how fixed asymmetric weights affect these trade-offs in the evaluated medical QA settings.

### 4.1 Training Dynamics of Asymmetric Weights for Positive and Negative Samples

We systematically investigate the impact of weight imbalance between PSR and NSR on the training process. The experimental design follows a controlled ablation: with the positive sample weight $w^+ = 1.0$, the negative sample weight $w^- \in \{0.05, 0.1, 0.2, 1, 5\}$, as shown in the first row of Figure 2; with the negative sample weight $w^- = 1.0$, the positive sample weight $w^+ \in \{0.05, 0.1, 0.2, 1, 5\}$, as shown in the second row of Figure 2. The two sweeps probe relative polarity strength from complementary directions. For a given ratio, absolute weight scale may also interact with clipping, KL regularization, and optimization dynamics.

**Negative-dominant weighting is associated with higher entropy in the reported sweep.** With $w^+ = 1$ fixed, small negative weights ($\leq 0.2$) tend to produce PSR-like entropy reduction in Figure 2. Larger negative weights retain more entropy, and the $w^- = 5$ run exhibits stronger reward gains alongside late-stage fluctuations. The response-length trajectories overlap and fluctuate, so they do not establish a monotonic relationship between negative-sample weight and length.

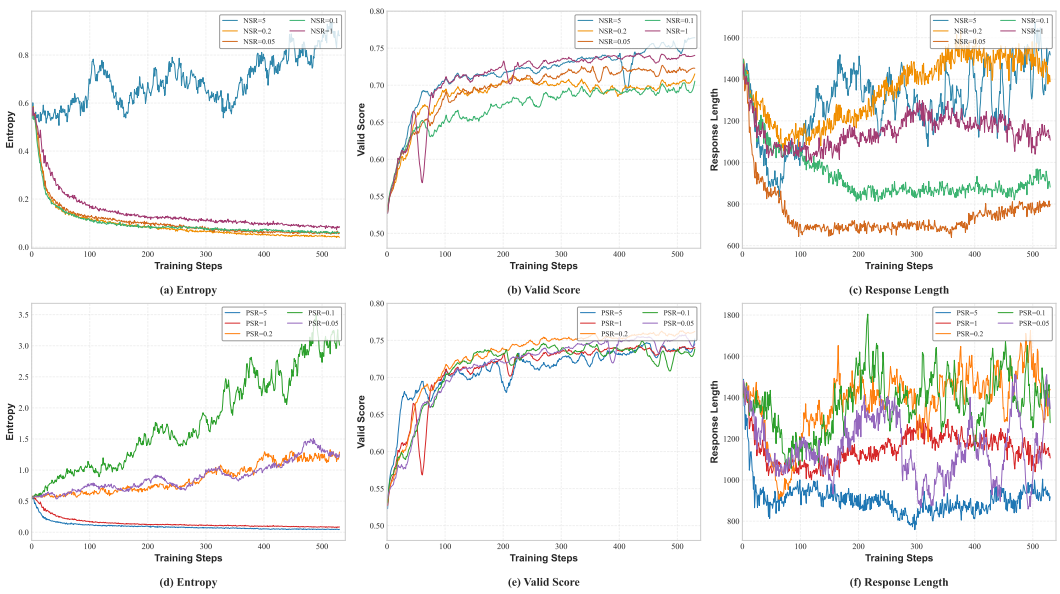


Figure 2: Training dynamics under asymmetric weighting on the RJUA dataset. The corresponding results on the CMD dataset are provided in Appendix B.

**Strong positive weighting is associated with entropy reduction in the reported sweep.** With $w^- = 1$ fixed, the $w^+ = 5$ run shows very low entropy and short responses. The evidence is strongest for entropy reduction under highly positive-dominant weighting; intermediate response-length trajectories are less clearly ordered.

**The sweeps reveal sensitivity to relative sample weighting.** Across the tested weights from 0.05 to 5, the displayed reward trajectories vary more in the negative-weight sweep than in the positive-weight sweep. This comparison characterizes sensitivity to coefficient selection in these training runs, where both relative polarity and optimization scale may affect the trajectories.

### 4.2 Relationship Between the P/N Sample Weight Ratio and Performance

Table 3: Performance under different positive/negative reward weight configurations.

| Methods | $w^+$ | $w^-$ | $w^+/w^-$ | Rouge-L | Reranker |
|---|---|---|---|---|---|
| PSR | 1 | 0 | $\infty$ | 0.293 | 0.940 |
| NSR | 0 | 1 | 0 | 0.327 | 0.971 |
| GRPO | 1 | 1 | 1 | 0.301 | 0.970 |
| P/N ↓ | 1 | 0.05 | 20 | 0.299 | 0.933 |
| | 1 | 0.1 | 10 | 0.232 | 0.922 |
| | 1 | 0.2 | 5 | 0.274 | 0.931 |
| | 1 | 5 | 0.2 | **0.335** | **0.976** |
| P/N ↑ | 0.05 | 1 | 0.05 | 0.263 | 0.903 |
| | 0.1 | 1 | 0.1 | 0.322 | 0.958 |
| | 0.2 | 1 | 0.2 | **0.337** | **0.983** |
| | 5 | 1 | 5 | 0.302 | 0.970 |

We evaluate the performance under different positive-to-negative sample weight ratios based on Qwen3-8B on the RJUA test set, and the evaluation results are shown in Table 3. The experimental results on CMD are provided in Appendix C.

**RJUA scores vary across the tested weight configurations.** With the positive weight fixed at 1, setting the negative weight to 5 gives Rouge-L 0.335 and Reranker 0.976, compared with 0.327 and




Figure 3: Training dynamics of EAPO against various baseline methods on the RJUA dataset.

0.971 for NSR. With the negative weight fixed at 1, the tested positive weights yield non-monotonic scores, with the $w^+ = 5$ configuration below $w^+ = 0.2$ on both metrics. These comparisons highlight the dependence of final scores on coefficient selection.

**A 1:5 ratio performs well in the tested RJUA sweep.** Among the configurations in Table 3, $(w^+, w^-) = (0.2, 1)$ has the highest Rouge-L and Reranker scores. The $(1, 5)$ configuration has the same ratio and also performs well. The 1:5 ratio is a favorable choice within this RJUA grid; the CMD results indicate that the preferred setting depends on the dataset and metric.

## 5 Entropy-Driven Adaptive Positive-Negative Sample Weighting

The preceding results motivate adapting relative sample weights to observed training dynamics. Fixed coefficients can perform well, but do not respond directly to changes in entropy. We therefore introduce EAPO as a lightweight entropy-feedback weighting rule within asymmetric sample weighting.

### 5.1 Method

EAPO adjusts the positive-sample weight using the ratio of current policy entropy $H_t$ to initial entropy $H_0$. This is a simple feedback design intended to moderate concentration and dispersion of the policy. Appendix D gives an algebraic interpretation of the ratio; it does not prove convergence or optimality relative to alternative controllers. The dynamic positive sample weight is defined as:

$$w_t^+ = \mathrm{clip}\left( w_0 \cdot \frac{H_t}{H_0},\ w_{\min},\ w_{\max} \right) \tag{7}$$

where $w_0$ is the base positive weight and $[w_{\min}, w_{\max}]$ bounds the weight within a prescribed range. The negative sample weight $w^-$ is held fixed as a reference. The ratio $H_t/H_0$ normalizes entropy by its initial level; this normalization alone does not establish equivalent behavior across models.

Given the dynamic weight, the weighted advantage for rollout $y_i$ is:

$$\hat{A}_i = \begin{cases} w_t^+ \cdot A_i & i \in \mathcal{D}_x^+ \\ w^- \cdot A_i & i \in \mathcal{D}_x^- \end{cases} \tag{8}$$

When $H_t < H_0$, the unclipped positive weight is below $w_0$, weakening positive reinforcement relative to the reference setting. When $H_t > H_0$, it is above $w_0$, strengthening positive reinforcement, subject to the clipping bounds. These comparisons are relative to the initial entropy, not necessarily to the preceding training step. They describe the controller's intended feedback, without guaranteeing stable convergence or reliable answers.

The training objective of EAPO is as follows:

Table 4: Performance comparison of different reinforcement learning methods. RL@8 denotes the best Rouge-L score among 8 inference samples, while RR@8 denotes the best Reranker score among 8 inference samples.

| Methods | Rouge-L | Reranker | RL@8 | RR@8 | Avg |
|---|---|---|---|---|---|
| GRPO | 0.312 | **0.976** | 0.344 | 0.987 | 0.655 |
| PSR | 0.293 | 0.940 | 0.327 | 0.979 | 0.635 |
| NSR | 0.327 | 0.971 | 0.361 | 0.990 | 0.662 |
| W-REINFORCE | 0.322 | 0.958 | **0.366** | **0.998** | 0.661 |
| EAPO | **0.330** | 0.973 | 0.360 | 0.995 | **0.664** |

$$\mathcal{J}(\theta) = \mathbb{E}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|y_i|}\sum_{t=1}^{|y_i|}\min\big(r_t\hat{A}_{i,t}, \\ \text{clip}(r_t, 1-\epsilon, 1+\epsilon)\,\hat{A}_{i,t}\big) - \beta\, D_{\text{KL}}[\pi_\theta \| \pi_{\text{ref}}]\Bigg] \tag{9}$$

$r_t = \frac{\pi_\theta(y_{i,t}|x, y_{i,<t})}{\pi_{\theta_{\text{old}}}(y_{i,t}|x, y_{i,<t})}$ is the importance ratio. $\beta$ is the KL divergence coefficient, and the KL divergence penalty term is added to constrain policy shift.

# 6 Experiments and Results

## 6.1 Experimental Settings

We select the following methods as baselines: **GRPO**: both $w^+$ and $w^-$ are set to 1.0. **W-REINFORCE**: fixed asymmetric weights, setting $w^+$ to 0.1. **PSR**: training with positive samples only with $w^+ = 1.0, w^- = 0$. **NSR**: training with negative samples only with $w^+ = 0, w^- = 1.0$. All experiments are conducted using Qwen3-8B as the backbone.

We set the global training batch size to 32. Both the maximum prompt length and the maximum response length are set to 2048 tokens. The model is trained for a total of 10 epochs with a learning rate of $1e^{-6}$ and a KL coefficient $\beta = 0.001$. clipping range $\epsilon = 0.2$. Default hyperparameters for EAPO: $w_0 = 0.2, w_{min} = 0, w_{max} = 2.0$. During the rollout stage, $N = 4$ trajectories are sampled per prompt. All experiments are conducted on 4×NVIDIA H800 GPUs. We adopt DeepSeek-V3.2(DeepSeek-AI, 2025) as the evaluator to score the model's reasoning process under the LLM-as-a-Judge protocol.

Details of the training data and evaluation protocol are provided in Appendix E.

## 6.2 Main Results

Figure 3 compares the reported training runs on RJUA. PSR and GRPO exhibit strong entropy reduction, while NSR and the displayed fixed-weight PSR0.2-NSR1 run show increasing entropy. EAPO follows a more moderate entropy trajectory and avoids the extreme response-length behavior visible in these comparisons. These qualitative comparisons characterize training behavior in the shown runs.

The RJUA results in Table 4 report single-generation and best-of-$k$ scores, where $k$ is the number of inference samples. Appendix G reports an additional comparison on Qwen2.5-7B-Instruct(Qwen et al., 2025). We interpret these results as follows:

**Final-score comparisons are metric-dependent.** PSR has the lowest scores among the methods in the main RJUA comparison, while NSR is competitive on single-generation Rouge-L and Reranker. The selected W-REINFORCE baseline has the highest RL@8 and RR@8, but lower single-generation scores than EAPO and NSR. Best-of-$k$ scores describe sampled performance under these metrics; they do not by themselves establish output diversity, per-response reliability, or training stability.

**EAPO is competitive, with small and mixed final-score differences.** On RJUA, EAPO's Avg is 0.664, compared with 0.662 for NSR and 0.661 for the selected W-REINFORCE baseline. However, the best fixed-weight configuration in Table 6 reaches 0.672. On CMD (Table 7), EAPO scores 0.516, below NSR at 0.521 and the selected W-REINFORCE configuration at 0.518, and above GRPO at 0.503 and PSR at 0.457. These results do not establish consistent final-score superiority, and statistical significance is not established by the reported point estimates. Appendix F reports the fixed-weight comparisons and the initialization-sensitivity analysis.

**Initialization sensitivity is reduced in most of the tested comparisons.** Across the three initial weights in Table 8, EAPO has smaller score ranges than W-REINFORCE for RJUA Rouge-L, RJUA Reranker, and CMD Reranker. CMD Rouge-L is an exception: EAPO's range is 0.011, compared with 0.009 for W-REINFORCE. The supported practical claim is therefore competitive final performance with more moderate observed training dynamics and reduced initialization sensitivity in three of four metric–dataset pairs. This evidence concerns the tested initial weights and medical QA settings, rather than a tuning-free method or robustness to all controller choices.

## 7 Conclusion

We study positive- and negative-advantage rollouts under continuous rewards on two Chinese medical QA datasets and introduce EAPO, a lightweight entropy-feedback rule for adapting their relative weighting. The reported experiments show contrasting entropy dynamics and metric-dependent trade-offs between sample polarities. EAPO achieves competitive final scores, exhibits more moderate training dynamics in the shown runs, and reduces sensitivity to initialization in three of four tested metric–dataset pairs. The best fixed-weight configuration on RJUA and NSR on CMD obtain higher final average scores. These findings support EAPO as a practical weighting strategy within the evaluated settings. Broader task evaluation, repeated-run uncertainty estimates, independent judge validation, and controller ablations are needed to assess its wider applicability.

## Limitations

Our empirical evidence is limited to two Chinese medical QA datasets and Qwen-family backbones of up to 8B parameters. The additional backbone does not establish cross-domain generality or a model-capacity scaling trend. Final-score differences are small and mixed, and the reported results do not establish statistical significance across repeated training runs. Initialization sensitivity is evaluated at only three values; clipping bounds, alternative feedback functions, and reward-component choices have not been systematically ablated. The mean-based partition is relative to the selected reward and can be affected by outliers or changes in component weighting; it is not an absolute correctness criterion. LLM-as-a-Judge scores have not been independently validated against human judgments or alternative judges, and may reflect shared reward and evaluation biases. Clinical validity or safety cannot be inferred from these scores alone. All public models and baselines used in this work are employed under their original licenses or service terms, while proprietary large language models are accessed through their official services. Future work includes broader task and model evaluation, repeated-run uncertainty estimates, reward and controller ablations, and independent assessment of judge reliability.

## References

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Remi Munos. Asymmetric reinforce for off-policy reinforcement learning: Balancing positive and negative rewards. *Advances in Neural Information Processing Systems*, 38:9640–9664, 2026.

Peter Chen, Xiaopeng Li, Ziniu Li, Wotao Yin, Xi Chen, and Tianyi Lin. Exploration vs exploitation: Rethinking rlvr through clipping, entropy, and spurious reward. *arXiv preprint arXiv:2512.16912*, 2025a.

Yongchao Chen, Harsh Jhamtani, Srinagesh Sharma, Chuchu Fan, and Chi Wang. Steering large language models between code execution and textual reasoning. In *International Conference on Learning Representations*, volume 2025, pp. 99676–99707, 2025b.

DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of open large language models, 2025.

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

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Kexin Huang, Haoming Meng, Junkang Wu, Jinda Lu, Chiyu Ma, Ziqian Chen, Xue Wang, Bolin Ding, Jiancan Wu, Xiang Wang, et al. On the direction of rlvr updates for llm reasoning: Identification and exploitation. *arXiv preprint arXiv:2603.22117*, 2026.

Yash Ingle, Jaival Chauhan, Ankit Yadav, and Sudhakar Mishra. Adaptive negative reinforcement for llm reasoning: Dynamically balancing correction and diversity in rlvr. *arXiv preprint arXiv:2605.07137*, 2026.

Shiwei Lyu, Chenfei Chi, Hongbo Cai, Lei Shi, Xiaoyan Yang, Lei Liu, Xiang Chen, Deng Zhao, Zhiqiang Zhang, Xianguo Lyu, Ming Zhang, Fangzhou Li, Xiaowei Ma, Yue Shen, Jinjie Gu, Wei Xue, and Yiran Huang. Rjua-qa: A comprehensive qa dataset for urology, 2023.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 94743–94765, 2025.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

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

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

Xinyu Tang, Yuliang Zhan, Zhixun Li, Wayne Xin Zhao, Zhenduo Zhang, Zujie Wen, Zhiqiang Zhang, and Jun Zhou. Rethinking sample polarity in reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2512.21625*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Qwen Team. Qwen3 technical report, 2025. URL `https://arxiv.org/abs/2505.09388`.

Xueyun Tian, Minghua Ma, Bingbing Xu, Nuoyan Lyu, Wei Li, Heng Dong, Zheng Chu, Yuanzhuo Wang, and Huawei Shen. Learning from mistakes: Negative reasoning samples enhance out-of-domain generalization. *arXiv preprint arXiv:2601.04992*, 2026.

Jiakang Wang, Runze Liu, Lei Lin, Wenping Hu, Xiu Li, Fuzheng Zhang, Guorui Zhou, and Kun Gai. Aspo: Asymmetric importance sampling policy optimization. *arXiv preprint arXiv:2510.06062*, 2025.

Peng-Yuan Wang, Tian-Shuo Liu, Chenyang Wang, Ziniu Li, Yidi Wang, Shu Yan, Chengxing Jia, Xu-Hui Liu, Xinwei Chen, Jiacheng Xu, et al. A survey on large language models for mathematical reasoning. *ACM Computing Surveys*, 58(8):1–35, 2026.

Junkang Wu, Kexin Huang, Jiancan Wu, An Zhang, Xiang Wang, and Xiangnan He. Quantile advantage estimation: Stabilizing rlvr for llm reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

Can Xie, Ruotong Pan, Xiangyu Wu, Yunfei Zhang, Jiayi Fu, Tingting Gao, and Guorui Zhou. Unlocking exploration in rlvr: Uncertainty-aware advantage shaping for deeper reasoning. *arXiv preprint arXiv:2510.10649*, 2025.

Huimin Xu, Shuai Zhao, Xiaobao Wu, and Anh Tuan Luu. Understanding and preventing entropy collapse in rlvr with on-policy entropy flow optimization. *arXiv preprint arXiv:2605.11491*, 2026a.

Ming Xu. Medical dataset (cmd). `https://github.com/Toyhom/Chinese-medical-dialogue-data`, 2023.

Yifan Xu, Junren Chen, and Yifan Chen. How you begin is how you reason: Driving exploration in rlvr via prefix-tuned priors. *arXiv preprint arXiv:2605.08817*, 2026b.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*, 2025.

Dayu Yang, Tianyang Liu, Daoan Zhang, Antoine Simoulin, Xiaoyi Liu, Yuwei Cao, Zhaopu Teng, Xin Qian, Grey Yang, Jiebo Luo, et al. Code to think, think to code: A survey on code-enhanced reasoning and reasoning-driven code intelligence in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2586–2616, 2025.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in llm reasoning. *Advances in Neural Information Processing Systems*, 38:126546–126573, 2026.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

# A Prompt and Evaluation Dimensions

## A.1 Prompt

We conduct experiments on two open-ended medical question-answering datasets, RJUA and CMD, with the detailed description on Appendix E.1. The corresponding prompts presented as follows.

**Prompt on RJUA**

You are an expert physician with comprehensive clinical expertise in Chinese urology and andrology. Please read the patient's consultation carefully and respond according to the following steps: 1. Perform clinical diagnostic reasoning within the **<think>** and **</think>** tags. The reasoning should encompass information acquisition and integration, disease hypothesis generation, hypothesis verification and exclusion, and clinical reasoning strategy. 2. Provide the diagnostic conclusion together with recommendations for further diagnostic workup or treatment within the **<advice>** and **</advice>** tags.

Patient consultation: {question}

**Prompt on CMD**

You are an expert physician with comprehensive clinical expertise across all medical specialties in China. Please read the patient's consultation carefully and respond according to the following steps: 1. Perform clinical diagnostic reasoning within the **<think>** and **</think>** tags. The reasoning should encompass information acquisition and integration, disease hypothesis generation, hypothesis verification and exclusion, and clinical reasoning strategy. 2. Provide the diagnostic conclusion together with recommendations for further diagnostic workup or treatment within the **<advice>** and **</advice>** tags.

Patient consultation: {question}

## A.2 Evaluation Dimensions on LLM-as-a-Judge

We adopt an LLM-as-a-Judge paradigm to score the model's reasoning content in <*think*> tag during reinforcement training and the corresponding judge prompt is presented as follows:

**Prompt on LLM-as-a-Judge for <think>**

You are a senior clinical medicine expert and a specialist in medical-education assessment. Your task is to evaluate the quality of the **internal reasoning process** of a medical consultation system.

Note: You are required to assess only the model's reasoning process (the content within the <think> tags); the final response delivered to the patient is not subject to evaluation.

**[Patient Consultation]**
{query}

**[Reference Answer (Ground Truth)]**
{gold_answer}

**[Model's Reasoning Process (to be evaluated)]** {think_content}

Please rate the model's reasoning process along the following four dimensions, each on a 100-point scale.

**[Dimension 1: Logical Coherence of Clinical Reasoning — Weight 35%]**
**[[Dimension 2: Accuracy of Medical Knowledge — Weight 30%]**

**[[Dimension 3: Adequacy of Differential Diagnosis — Weight 20%]**

**[[Dimension 4: Depth and Organization of Reasoning — Weight 15%]**

Overall Score = Logical Coherence × 0.35 + Medical Knowledge Accuracy × 0.30 + Differential Diagnosis Adequacy × 0.20 + Depth and Organization × 0.15

The prompt used to evaluate model-generated answers under the LLM-as-a-Judge paradigm is presented as follows:

**Prompt on LLM-as-a-Judge for Response**

You are a senior clinical medicine expert and a specialist in medical-education assessment. Your task is to perform a multi-dimensional evaluation of the response produced by a medical consultation system.

**[Patient Consultation Query]**
{query}

**[Reference Answer (Ground Truth)]**
{gold_diagnosis}

**[Model Response (To Be Evaluated)]** {response_content}

Please rate the model response along the following five dimensions, with each dimension scored on a 100 scale.

**[Dimension 1: Medical Factual Accuracy — Weight 30%]**
**[[Dimension 2: Diagnostic Completeness — Weight 25%]**

**[[Dimension 3: Clinical Safety — Weight 25%]**

**[[Dimension 4: Reasoning Logic and Interpretability — Weight 10%]**

**[[Dimension 5: Expression Clarity and Empathy — Weight 10%]**

Overall Score = Medical Factual Accuracy × 0.30 + Diagnostic Completeness × 0.25 + Clinical Safety × 0.25 + Reasoning Logic × 0.10 + Expression Clarity × 0.10

# B The Impact of Asymmetric Weights in Open-ended QA Reinforcement Training

Figure 4 reports the training dynamics on the CMD dataset under different asymmetric weighting configurations, where the upper row varies $w^-$ with $w^+ = 1$ fixed (NSR series), and the lower row varies $w^+$ with $w^- = 1$ fixed (PSR series).

The displayed CMD trajectories show sensitivity to the chosen weights, with overlapping and fluctuating response-length curves. They illustrate a setting-specific trade-off among entropy, response length, and reward; response length does not show a clear monotonic ordering across the tested weights.

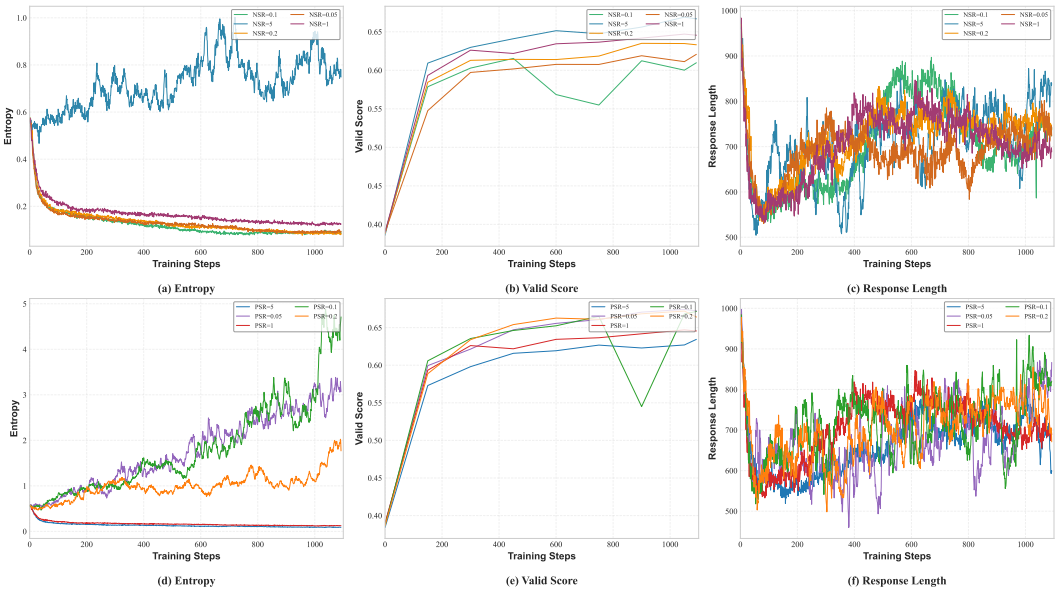


Figure 4: Training dynamics under asymmetric weighting on the CMD dataset.

Table 5: Performance under different positive/negative reward weight configurations. Benchmark: CMD.

| Methods | $w^+$ | $w^-$ | $w^+/w^-$ | Rouge-L | Reranker |
|---|---|---|---|---|---|
| PSR | 1 | 0 | $\infty$ | 0.147 | 0.688 |
| NSR | 0 | 1 | 0 | 0.194 | 0.792 |
| GRPO | 1 | 1 | 1 | 0.184 | 0.787 |
| P/N ↓ | 1 | 0.05 | 20 | 0.160 | 0.749 |
| | 1 | 0.1 | 10 | 0.158 | 0.724 |
| | 1 | 0.2 | 5 | 0.170 | 0.763 |
| | 1 | 5 | 0.2 | **0.187** | **0.793** |
| P/N ↑ | 0.05 | 1 | 0.05 | **0.194** | 0.778 |
| | 0.1 | 1 | 0.1 | 0.185 | **0.808** |
| | 0.2 | 1 | 0.2 | 0.186 | 0.754 |
| | 5 | 1 | 5 | 0.173 | 0.752 |

## C Relation Between the P/N Sample Weight Ratio and Performance

Table 5 reports the performance of various $(w^+, w^-)$ configurations on the CMD dataset.

On CMD, NSR exceeds PSR on both Rouge-L and Reranker in Table 5, but the fixed-weight results do not vary monotonically with the positive-sample weight. Together with Section 4.2, these results show dataset- and metric-dependent sensitivity to weighting.

## D Formula Derivation

To give an algebraic interpretation of the entropy-ratio rule, consider the following unclipped recursive update for nonzero reference entropies:

$$w_t = w_{t-1} \cdot \frac{H_t}{H_{t-1}} \tag{10}$$

where $H_t$ denotes the rollout entropy at training step $t$. Unrolling the recursion in Eq. (10) yields

$$\begin{aligned} w_t &= w_{t-1} \cdot \frac{H_t}{H_{t-1}} = w_{t-2} \cdot \frac{H_{t-1}}{H_{t-2}} \cdot \frac{H_t}{H_{t-1}} \\ &= \cdots = w_0 \cdot \frac{H_1}{H_0} \cdot \frac{H_2}{H_1} \cdots \frac{H_t}{H_{t-1}} \end{aligned} \tag{11}$$

Since the intermediate entropy terms cancel pairwise (telescoping product), the recursion collapses into the closed-form expression

$$\boxed{w_t \;=\; w_0 \cdot \frac{H_t}{H_0}} \tag{12}$$

where the symbols are defined as follows:

- $w_0$: the initial PSR weight; the full method also specifies clipping bounds;
- $H_0$: the rollout entropy recorded at the first training step (fixed thereafter);
- $H_t$: the rollout entropy at the current step $t$;
- Weight clipping: $w_t \in [0, 2]$ to prevent unbounded amplification.

Equation (12) is the algebraic form of the unclipped recursion. The controller specified in the main text applies the bounds in Eq. (7) to the current-to-initial entropy ratio; the telescoping identity does not establish equivalence to recursively clipping at each step. This derivation clarifies the update, but does not prove optimality of linear entropy feedback, convergence, or robustness relative to other controllers.

# E Datasets and Evaluation Methods

## E.1 Datasets

We evaluate our approach on two open-domain medical question answering datasets: CMD (Xu, 2023): A Chinese medical dialogue dataset. We take the subset of the top five departments, covering Internal Medicine, Surgery, Obstetrics and Gynecology, Pediatrics, and Dermatology. The training/test split is 9:1. RJUA (Lyu et al., 2023): A Chinese urological specialty diagnostic QA dataset, containing complete medical records and reference diagnoses. Both datasets are open-ended generation tasks with no single ground-truth answer; answer quality must be evaluated from multiple dimensions.

## E.2 Evaluation Methods

We report the following metrics: **ROUGE-L**: measures lexical overlap with the reference answer; it is not a direct diversity measure. **Reranker**: measures semantic relevance. **LLM-as-a-Judge (LAAJ)**: provides prompt-dependent quality scores; agreement with human assessment and clinical validity have not been established.

# F Different Fixed Weighting Configurations

Table 6 and Table 7 respectively present the performance comparison of different methods under various configurations of positive and negative sample weights.

The reported scores vary across fixed-weight configurations. On RJUA, the best fixed-weight Avg is 0.672, compared with 0.664 for EAPO; on CMD, NSR reaches 0.521 and EAPO reaches 0.516. These results motivate separating peak final performance from sensitivity to coefficient selection. EAPO adapts weights online to reduce dependence on a fixed coefficient, while initialization and clipping bounds remain design choices.

We further conduct a comparative analysis of the final performance of EAPO and W-REINFORCE under varying initial positive-sample weights, with the experimental results reported in Table 8.

Table 6: Effect of fixed asymmetric weights $(w^+, w^-)$ within W-REINFORCE compared with GRPO, PSR, NSR, and our proposed EAPO across Rouge-L, Reranker, Best-RL@8, and Best-RR@8 metrics. Benchmark: RJUA.

| Methods | Rouge-L | Reranker | RL@8 | RR@8 | Avg |
|---|---|---|---|---|---|
| GRPO | 0.312 | 0.977 | 0.344 | 0.987 | 0.655 |
| PSR | 0.293 | 0.940 | 0.327 | 0.979 | 0.635 |
| NSR | 0.327 | 0.971 | 0.361 | 0.990 | 0.662 |
| *W-REINFORCE* | | | | | |
| $w^+ = 0.05,\ w^- = 1$ | 0.263 | 0.903 | 0.346 | 0.996 | 0.627 |
| $w^+ = 0.1,\ w^- = 1$ | 0.322 | 0.958 | 0.366 | **0.998** | 0.661 |
| $w^+ = 0.2,\ w^- = 1$ | **0.337** | **0.983** | **0.372** | 0.997 | **0.672** |
| $w^+ = 1,\ w^- = 1$ | 0.301 | 0.970 | 0.326 | 0.994 | 0.648 |
| $w^+ = 5,\ w^- = 1$ | 0.302 | 0.970 | 0.335 | 0.989 | 0.649 |
| $w^+ = 1,\ w^- = 0.05$ | 0.299 | 0.933 | 0.324 | 0.983 | 0.635 |
| $w^+ = 1,\ w^- = 0.1$ | 0.232 | 0.922 | 0.264 | 0.973 | 0.598 |
| $w^+ = 1,\ w^- = 0.2$ | 0.274 | 0.931 | 0.288 | 0.978 | 0.618 |
| $w^+ = 1,\ w^- = 5$ | 0.335 | 0.976 | 0.371 | 0.995 | 0.669 |
| **EAPO** | 0.330 | 0.973 | 0.360 | 0.995 | 0.664 |

Table 7: Effect of fixed asymmetric weights $(w^+, w^-)$ within W-REINFORCE compared with GRPO, PSR, NSR, and our proposed EAPO across Rouge-L, Reranker, Best-RL@8, and Best-RR@8 metrics. Benchmark: CMD.

| Methods | Rouge-L | Reranker | RL@8 | RR@8 | Avg |
|---|---|---|---|---|---|
| GRPO | 0.186 | 0.783 | 0.206 | 0.838 | 0.503 |
| PSR | 0.147 | 0.688 | 0.173 | 0.821 | 0.457 |
| NSR | **0.194** | 0.792 | 0.220 | **0.879** | **0.521** |
| *W-REINFORCE* | | | | | |
| $w^+ = 0.05,\ w^- = 1$ | 0.194 | 0.778 | **0.220** | 0.874 | 0.517 |
| $w^+ = 0.1,\ w^- = 1$ | 0.185 | **0.808** | 0.207 | 0.871 | 0.518 |
| $w^+ = 0.2,\ w^- = 1$ | 0.186 | 0.754 | 0.216 | 0.876 | 0.508 |
| $w^+ = 1,\ w^- = 1$ | 0.184 | 0.787 | 0.207 | 0.852 | 0.508 |
| $w^+ = 5,\ w^- = 1$ | 0.173 | 0.752 | 0.199 | 0.838 | 0.490 |
| $w^+ = 1,\ w^- = 0.05$ | 0.160 | 0.749 | 0.184 | 0.827 | 0.480 |
| $w^+ = 1,\ w^- = 0.1$ | 0.158 | 0.724 | 0.183 | 0.806 | 0.467 |
| $w^+ = 1,\ w^- = 0.2$ | 0.170 | 0.763 | 0.195 | 0.852 | 0.495 |
| $w^+ = 1,\ w^- = 5$ | 0.187 | 0.793 | 0.210 | 0.862 | 0.513 |
| **EAPO** | 0.187 | 0.798 | 0.213 | 0.864 | 0.516 |

Across the three tested positive weights, W-REINFORCE's RJUA Rouge-L ranges from 0.263 to 0.337 and Reranker from 0.903 to 0.983. These score ranges describe sensitivity to coefficient selection in the reported experiments, rather than variance across repeated training runs.

For EAPO, the corresponding ranges are 0.011 for RJUA Rouge-L and 0.030 for RJUA Reranker, compared with 0.074 and 0.080 for W-REINFORCE. On CMD, EAPO also has a smaller Reranker range (0.008 versus 0.054), but a slightly larger Rouge-L range (0.011 versus 0.009). Thus, the evidence supports reduced sensitivity in three of four metric–dataset pairs over the tested initial weights. The scope of this result is initialization sensitivity over the tested values; clipping bounds and alternative feedback functions remain unevaluated.

Table 8: Sensitivity of W-REINFORCE and EAPO to the positive-sample weight. Range denotes the gap between the maximum and minimum values across the three settings.

| Methods | RUJA | | CMD | |
|---|---|---|---|---|
| | Rouge-L | Reranker | Rouge-L | Reranker |
| W-REINFORCE | | | | |
| $w^+ = 0.05$ | 0.263 | 0.903 | 0.194 | 0.778 |
| $w^+ = 0.1$ | 0.322 | 0.958 | 0.185 | 0.808 |
| $w^+ = 0.2$ | 0.337 | 0.983 | 0.186 | 0.754 |
| Range (max−min) | 0.074 | 0.080 | 0.009 | 0.054 |
| EAPO | | | | |
| $w_0 = 0.05$ | 0.321 | 0.943 | 0.195 | 0.806 |
| $w_0 = 0.1$ | 0.319 | 0.961 | 0.184 | 0.800 |
| $w_0 = 0.2$ | 0.330 | 0.973 | 0.187 | 0.798 |
| Range (max−min) | 0.011 | 0.030 | 0.011 | 0.008 |

Table 9: Performance comparison on Qwen2.5-7B-Instruct. Benchmark: RJUA.

| Methods | Rouge-L | Reranker | RL@8 | RR@8 | Avg |
|---|---|---|---|---|---|
| GRPO | 0.323 | 0.968 | 0.346 | 0.990 | 0.657 |
| PSR | 0.287 | 0.944 | 0.294 | 0.970 | 0.624 |
| NSR | 0.246 | 0.858 | 0.359 | **0.994** | 0.614 |
| W-REINFORCE | 0.336 | 0.973 | 0.371 | 0.992 | 0.668 |
| **EAPO** | **0.341** | **0.977** | **0.388** | 0.994 | **0.675** |

## G Results on an Additional Backbone

Table 9 reports an additional experiment on Qwen2.5-7B-Instruct using RJUA. EAPO has the highest reported Rouge-L, Reranker, RL@8, and Avg (0.675) among the compared methods, and ties NSR on RR@8 at the displayed precision. This extends the empirical evidence to one additional Qwen-family backbone within medical QA. Cross-domain transfer and model-capacity scaling require broader evaluation.