STABILIZING OFF-POLICY REINFORCEMENT LEARNING FOR LLMS VIA BALANCED POLICY OPTIMIZATION WITH ADAPTIVE CLIPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) has recently become the core paradigm for aligning and strengthening large language models (LLMs). Yet, applying RL in off-policy settings—where stale data from past policies are used for training—improves sample efficiency, but remains challenging: policy entropy declines sharply, optimization often becomes unstable and may even collapse. Through theoretical and empirical analysis, we identify two key insights: (i) **an imbalance in optimization**, where negative-advantage samples dominate the policy gradient, suppressing useful behaviors and risking gradient explosions; and (ii) **the derived Entropy-Clip Rule**, which reveals that the fixed clipping mechanism in PPO-like objectives systematically blocks entropy-increasing updates, thereby driving the policy toward over-exploitation at the expense of exploration. Building on these insights, we propose **BA**lanced **P**olicy **O**ptimization with Adaptive Clipping (**BAPO**), a simple yet effective method that dynamically adjusts clipping bounds to adaptively re-balance positive and negative contributions, preserve entropy, and stabilize RL optimization. Across diverse off-policy scenarios—including sample replay and partial rollout—BAPO achieves fast, stable, and data-efficient training. On AIME 2024 and AIME 2025 benchmarks, our 7B BAPO model surpasses open-source counterparts such as SkyWork-OR1-7B, while our 32B BAPO model not only achieves state-of-the-art results among models of the same scale but also outperforms leading proprietary systems like o3-mini and Gemini-2.5-Flash-Thinking.

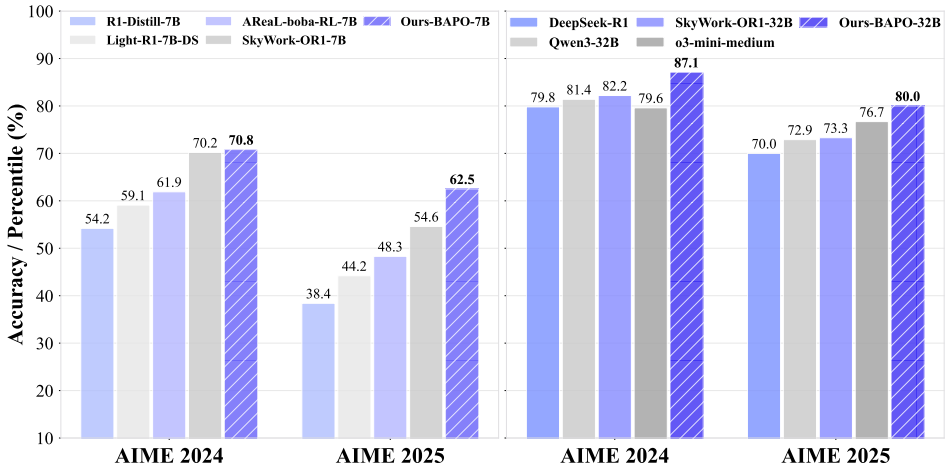

Figure 1: Performance of BAlanced Policy Optimization with Adaptive Clipping (BAPO).

1 INTRODUCTION

Reinforcement learning (RL) has become a pivotal paradigm for optimizing large language models (LLMs) (Zhang et al., 2025), delivering significant improvements in complex tasks such as reasoning

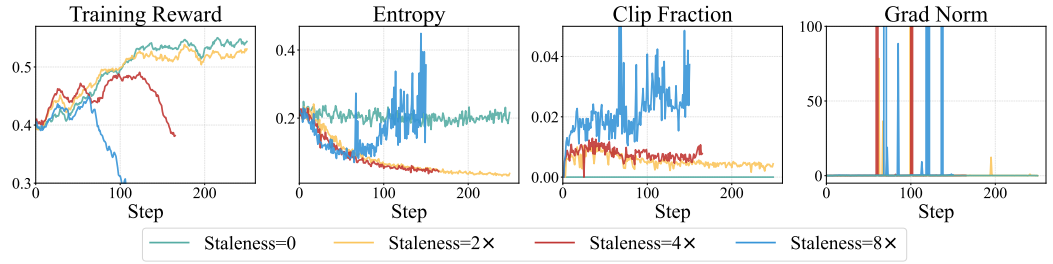

Figure 2: Preliminary results with different data staleness. As the staleness increases, the model suffers from unstable optimization, decreasing entropy, and even a sudden collapse in training.

(Jaech et al., 2024; Guo et al., 2025), coding (Anthropic, 2025), and agentic decision-making (Bai et al., 2025). Among RL methods, off-policy RL—where the rollout policy (behavior policy) differs from the training policy (target policy)—emerges as particularly promising (Roux et al., 2025; Arnal et al., 2025). It offers high sample efficiency and tolerance to data staleness, making it well-suited for extremely long-horizon and challenging scenarios, while also aligning more naturally with features in modern AI infrastructures such as partial rollout (Team et al., 2025; Fu et al., 2025).

However, applying off-policy RL to LLMs introduces substantial challenges (Yu et al., 2025; Arnal et al., 2025). As shown in Figure 2, increasing data staleness leads to unstable optimization, exploding gradient and even collapse. Meanwhile, policy entropy declines sharply, reflecting reduced exploratory capacity and a bias toward over-exploitation. By contrast, on-policy training—where rollout and target policies coincide—remains stable across metrics, consistent with prior studies (Tang et al., 2024; Roux et al., 2025; Arnal et al., 2025).

To understand the instability of off-policy training, we conduct a comprehensive theoretical and empirical analysis to reveal two key insights. We first demonstrate an **imbalance in optimization**: policy updates are often dominated by negative-advantage samples, producing excessive penalization signals that suppress even neutral or correct actions and may cause gradient explosions (Gülçehre et al., 2023). We then derive and empirically validate the **Entropy-Clip Rule** in the widely-used PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024), showing that the clipping mechanism in PPO-like objectives blocks many low-probability positive tokens while over-penalizing low-probability negatives. This systematically excludes entropy-increasing updates, sharpens the output distribution, and drives policies toward over-exploitation at the cost of exploration.

Based on these insights, we propose **BA**lanced **P**olicy **O**ptimization with Adaptive Clipping (**BAPO**), a new method for stable and effective off-policy RL. BAPO dynamically adjusts the clipping bounds to re-balance positive and negative contributions for each update step, incorporate low-probability positives while filtering excessive negatives, and preserve policy entropy—achieving a better balance between exploration and exploitation. An overview of our approach is illustrated on the right side of Figure 3.

Experiments across diverse off-policy scenarios—including sample replay, partial rollout, and varying degrees of staleness—on base models such as DeepSeek-R1-Distill-Qwen-7B (Guo et al., 2025) and OctoThinker-Llama3.2-3B-Long-Zero (Wang et al., 2025b) show that BAPO consistently yields significant improvements. Our 7B model achieves scores of 70.8 on AIME24 and 62.5 on AIME25, surpassing open-source counterparts such as SkyWork-OR1-7B (He et al., 2025). Moreover, our 32B model reaches 87.1 on AIME24 and 80.0 on AIME25, outperforming both comparably scaled open-source models like Qwen3-32B (Yang et al., 2025a) and leading proprietary systems including o3-mini-medium (OpenAI, 2025) and Gemini-2.5-Flash-Thinking (Comanici et al., 2025).

Our contributions are summarized as follows:

- We identify and analyze two key insights behind instability in off-policy RL for LLMs: the imbalanced optimization and the Entropy-Clip Rule. (§3)
- We propose BAPO, a new RL algorithm that dynamically adjusts clipping bounds to balance positive and negative signals, preserving entropy for exploration, and stabilizing training. (§4)
- We validate BAPO across multiple backbones, model scales, and off-policy settings, showing that it achieves stable optimization and competitive results with proprietary systems. (§5)

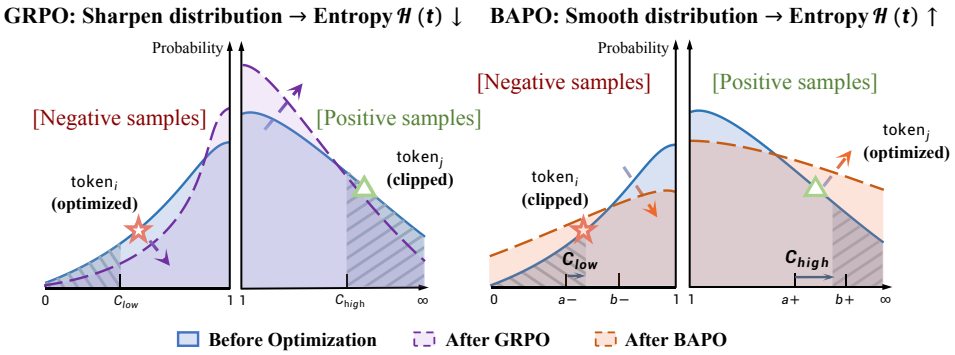

Figure 3: An illustration of our proposed BAPO. **(Left)** Baseline methods like GRPO use symmetric fixed clipping bounds, reinforcing high-probability positive tokens while penalizing excessive low-probability negatives, leading to sharp distributions and entropy collapse. **(Right)** BAPO dynamically adjusts the clipping bounds c_{low} and c_{high} based on the loss contributions from positive tokens. It excludes overly negative tokens ☆ to maintain a smoother distribution and incorporates previously clipped positive tokens △ to preserve entropy balance.

2 PRELIMINARIES

Policy gradient. In the field of LLM RL (Trung et al., 2024; Jaech et al., 2024), policy gradient-based (PG) algorithms (Williams, 1992) are widely used. Specifically, given an input prompt $\boldsymbol{x}$, an LLM π_θ sequentially generates a T-token response $\boldsymbol{y} = (y_1, ..., y_T)$:

$$\pi_\theta(\boldsymbol{y}|\boldsymbol{x}) = \prod_{t=1}^{T} \pi_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \ . \tag{1}$$

Given a training dataset $\mathcal{D} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ and reward function R, the RL objective is to maximize the expected reward:

$$J(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\ \boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{x})} \left[R(\boldsymbol{x}, \boldsymbol{y})\right] \ . \tag{2}$$

PG algorithms then leverage gradient ascent to optimize the policy with the following gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\ \boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{x})} \left[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t}) \cdot A_t\right] \ , \tag{3}$$

where A_t denotes the estimated advantage at time step t, i.e., how much better action y_t is than the expected action under the current policy.

Importance sampling and PPO objective. To improve sample efficiency and adapt to modern infrastructure, mainstream RL algorithms for LLMs typically adopt a PPO-like surrogate objective (Schulman et al., 2017):

$$J^{\text{PPO}}(\theta) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D},\ \boldsymbol{y}\sim\pi_{\theta_{\text{rollout}}}(\cdot|\boldsymbol{x})} \sum_{t=1}^{T} \left[\min(r_t \cdot A_t, \text{clip}(r_t, 1-\varepsilon, 1+\varepsilon) \cdot A_t)\right] \ , \tag{4}$$

where $r_t = \frac{\pi_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})}{\pi_{\theta_{\text{rollout}}}(y_t|\boldsymbol{x}, \boldsymbol{y}_{<t})}$ is the importance weight that corrects for the distribution mismatch, estimating the expected advantage of tokens generated by the behavior policy $\pi_{\theta_{\text{rollout}}}$ under the target policy π_θ. The clipping mechanism in PPO serves to implicitly enforce a trust region between the behavior and target policies, preventing overly large policy updates that could destabilize training. The hyperparameter $\varepsilon \in (0, 1)$ determines the width of this clipping interval.

We then analyze data with positive and negative advantages respectively. The policy gradient can then be expressed as:

$$\nabla J^{\text{PPO}} = \underbrace{\sum_{A_t>0} \pi_\theta(y_t) \cdot \mathbb{I}\{r_t < 1+\varepsilon\} \cdot A_t \cdot \nabla \log \pi_\theta(y_t)}_{\text{positive tokens}} + \underbrace{\sum_{A_t<0} \pi_\theta(y_t) \cdot \mathbb{I}\{r_t > 1-\varepsilon\} \cdot A_t \cdot \nabla \log \pi_\theta(y_t)}_{\text{negative tokens}}, \tag{5}$$

where $\mathbb{I}$ represents the indicator function.

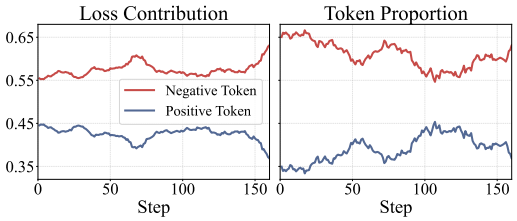

Figure 4: Contribution of positive and negative tokens to the policy-gradient loss and their proportion of tokens during training.

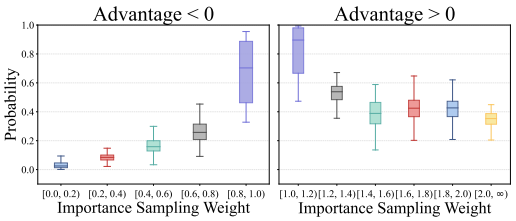

Figure 5: Relationship between token probability and importance sampling weight.

3 Motivation: Imbalanced Optimization and Entropy-Clip Rule

In this section, we first conduct preliminary experiments to show the influence of data staleness on the RL optimization process. Next, we perform in-depth empirical and theoretical analysis to reveal the underlying mechanisms and provide new insights.

Training instability with data staleness. We perform experiments under different levels of data staleness using the popular GRPO algorithm. Results in Figure 2 show that, compared to on-policy training, off-policy RL typically suffers from instability, and entropy decreases rapidly, reflecting reduced exploratory capacity (He et al., 2025). As staleness increases, the entropy decline becomes more severe and a larger number of tokens are clipped; meanwhile, training becomes more unstable. In the following paragraphs, we attempt to explain this phenomenon from different perspectives and summarize the motivation behind our method.

Excessive negative samples lead to imbalanced optimization. Within the PPO-like objective for policy updates, we analyze tokens with positive and negative advantages separately, as shown in Equation 5. Empirical results in Figure 4 reveal a pronounced imbalance: positive samples constitute a minority both in number and in their contribution to the policy-gradient loss. We attribute this skew to two main factors: (i) the model tends to generate longer trajectories on difficult queries, thereby producing more tokens in negative samples (Figure 6); and (ii) in early stages of training, the model has not yet acquired sufficient capability, resulting in a higher proportion of negative samples. This observation may help explain the effectiveness of certain curriculum-based approaches (Xi et al., 2024; Yuan et al., 2025).

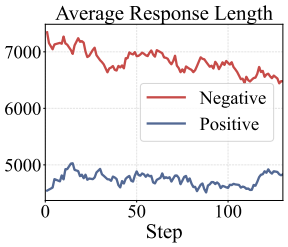

Figure 6: Average model response length during training.

In the RL training of LLMs, reinforcing positive samples is often more efficient for driving performance gains than attempting to "suppress" the vast number of negative samples (Gülçehre et al., 2023; Zhu et al., 2025). To this end, prior work has proposed amplifying positive signals through the clip-higher technique (Yu et al., 2025). However, merely enlarging the clipping upper bound does not mitigate the influence of negative data, thus failing to prevent them from dominating the optimization process. Moreover, as shown in Equation 5, the accumulation of low-probability negative tokens (i.e., $\pi_\theta(y_t) \to 0$, driving the log term toward $-\infty$) may trigger gradient explosion, further destabilizing training (Yang et al., 2025c).

The Entropy-Clip Rule exposes insufficient entropy promotion in optimization, leading to entropy collapse. Theoretically, we derives Equation 6 (see Appendix C for detailed derivations) for PPO surrogate objective to reveal the factors that influence the policy entropy (Roux et al., 2025):

$$\Delta\mathcal{H}(\pi_\theta) \approx -\eta \cdot \mathrm{Cov}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{x})}\left[\log \pi_\theta(y_t|\boldsymbol{x}, \boldsymbol{y}_{<\boldsymbol{t}}), A_t \cdot \mathcal{X}(y_t) + C\right] , \tag{6}$$

where C is a constant, and

$$\mathcal{X}(y_t) = \begin{cases} 1, & \textbf{\textit{if}}\ A_t > 0\ \&\ r_t < 1+\epsilon \\ & \textbf{\textit{or}}\ A_t < 0\ \&\ r_t > 1-\epsilon \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

We observe that changes in policy entropy are driven by the influence of unclipped tokens, which is determined by the covariance between their log probabilities and advantages. We term this as **the Entropy-Clip Rule**. The left side of Figure 3 illustrates how the optimization of different tokens influences the probability distribution, thereby affecting entropy. The Entropy-Clip Rule theoretically explains the following statement: Specifically, updating the policy with positive high-probability tokens (high advantage, high probability) and negative low-probability tokens (low advantage, low probability) sharpens the distribution and consequently reduces entropy. Conversely, updating the policy with negative high-probability tokens and positive low-probability tokens smooths the distribution, resulting in an increase in entropy (detailed proofs are available in Appendix C.4.2).

Empirically, our statistical analysis on token probabilities and their importance sampling (IS) weights further clarifies this dynamic. As shown in Figure 5, we find that tokens with either very high or very low IS weights tend to have low probabilities. However, in standard algorithms with symmetric clipping bounds (e.g., [0.8,1.2]), a majority of positive, low-probability tokens are prevented from contributing to the optimization. This systematic exclusion of entropy-increasing updates causes a continuous decline in entropy, ultimately crippling the model's exploratory capacity and resulting in a performance bottleneck.

Summary of motivation. Based on the above analysis, we can summarize two main motivations: (1) to balance the contributions of positive and negative tokens while preventing gradient explosion, and (2) to preserve policy entropy for sustaining exploration and preventing collapse.

4 Methodology

4.1 Validation Experiment: Asymmetric Clipping

The main idea of our method is to stabilize the training and maintain exploration ability of the policy by asymmetrically adjusting the trust region for positive and negative tokens, i.e., adjusting c_{low} and c_{high}.

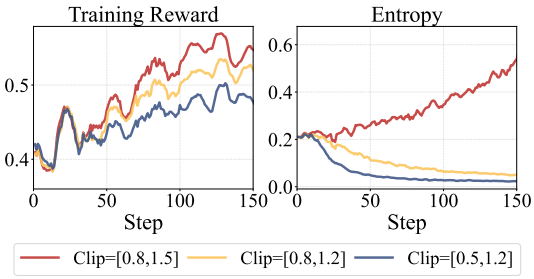

Figure 7: Training dynamics of asymmetric clipping experiments.

We then conduct preliminary validation experiments to examine whether asymmetrically adjusting the clipping range could influence the training dynamics. The results, shown in Figure 7, together with Figure 5, reveal that increasing the upper bound c_{high} (which introduces more low-probability positive tokens to policy updates) improves performance while counteracting the downward trend of entropy, albeit at a rapid pace. In contrast, relaxing the lower bound c_{low} (which introduces more low-probability negative tokens to policy updates) not only degrades performance but also accelerates entropy collapse. These findings confirm the effectiveness of entropy control through asymmetric clipping. Nevertheless, this approach remains relatively rigid and manually specified, providing limited flexibility and adaptation.

4.2 BAPO: Balanced Policy Optimization with Adaptive Clipping

To this end, we propose **BA**lanced **P**olicy **O**ptimization with Adaptive Clipping (**BAPO**), a new method to achieve stable, fast RL optimization for LLMs. The core insight of BAPO lies in its adaptive clipping mechanism, which dynamically adjusts the clipping bounds c_{high} and c_{low}, to regulate the positive contribution to the policy loss and maintain a balance in entropy throughout RL training. Formally, for each update with a batch, our goal is to find a pair of c_{high} and c_{low} that satisfy:

$$\frac{\left|\sum_{A_t>0} \pi_{\theta_{\text{rollout}}}(y_t)\cdot[\min(r_t\cdot A_t, \text{clip}(r_t,0,c_{\text{high}})\cdot A_t)]\right|}{\left|\sum_{A_t} \pi_{\theta_{\text{rollout}}}(y_t)\cdot[\min(r_t\cdot A_t, \text{clip}(r_t,c_{\text{low}},c_{\text{high}})\cdot A_t)]\right|} \geq \rho_0\ , \tag{8}$$

Algorithm 1: BAPO

Input: Initialized LLM policy π_θ, training dataset $\mathcal{D}$, reward function R, staleness E, movable range of clipping bounds $[a^-, b^-]$ and $[a^+, b^+]$, step size of upper bound δ_1, step size of lower bound δ_2, positive token contribution threshold ρ_0

for *step* $s = 1...S$ **do**
  Procedure Sample and filter out responses
    Update the old LLM policy $\pi_{\theta_{\text{rollout}}} \leftarrow \pi_\theta$;
    Sample the s-th batch $\mathcal{D}_s$ from $\mathcal{D}$;
    Sample G responses $\{\boldsymbol{y}_i\}_{i=1}^{G} \sim \pi_{\theta_{\text{rollout}}}(\cdot|\boldsymbol{x})$, where $\boldsymbol{x} \in \mathcal{D}_s$;
    Compute reward and advantage for each $\boldsymbol{y}_i$ based on reward function R ;
  for *staleness* $= 0...E$ **do**
    Procedure Dynamically adjusting the clipping bounds c_{high} *and* c_{low}
      Initialize clipping bounds $c_{\text{low}} = a^-$ and $c_{\text{high}} = a^+$;
      while *the positive token contribution* $\rho < \rho_0$ **and** $c_{\text{low}} + \delta_2 \leq b^-$ **do**
        if $c_{\text{high}} + \delta_1 \leq b^+$ **then**
          $c_{\text{high}} \leftarrow c_{\text{high}} + \delta_1$
        else
          $c_{\text{low}} \leftarrow c_{\text{low}} + \delta_2$
        end
      end
    Procedure Update the LLM policy π_θ
      Update the LLM policy π_θ by maximizing the following objective:
        $J^{\text{BAPO}}(\theta) = \mathbb{E}_{\boldsymbol{y}\sim\pi_{\theta_{\text{rollout}}}(\cdot|\boldsymbol{x})} \sum_{t=1}^{T} [\min(r_t \cdot A_t, \text{clip}(r_t, c_{\text{low}}, c_{\text{high}}) \cdot A_t)]$
  end
end

where ρ_0 is the target contribution of positive signals to the policy gradient loss. Specifically, BAPO gradually increases c_{high} and c_{low} with step sizes of δ_1 and δ_2, respectively, until the condition in Equation 8 is met. We present an overview of BAPO in Figure 3 and summarize it in Algorithm 1.

Overall, BAPO offers several significant benefits. First, by dynamically adjusting c_{high} and c_{low} for each step, we can increase the contribution of positive tokens to the policy-gradient loss while preventing negative tokens from excessively dominating the optimization objective. Second, based on our earlier analysis of the relationship between IS weights and token probabilities in Figure 5, BAPO incorporates more low-probability positive tokens and filters out more low-probability negative tokens, both of which contribute to maintaining entropy. Third, by setting the target contribution from positive tokens, BAPO prevents uncontrolled entropy growth, avoids situations where positive tokens overwhelm the loss, and mitigates tail degradation—where the model overfits to easy problems but fails to handle more challenging ones (Ding et al., 2025).

4.3 Analysis

Stable and fast training of BAPO. As shown in Figure 9, BAPO enables a more stable optimization process, characterized by rapidly increasing training rewards, greater contributions from positive tokens, steady gradient normalization, and stable policy entropy—resulting in an improved balance between exploration and exploitation.

We further visualize the adjustment process of the clipping bounds in BAPO. As shown in Figure 8, the averaged upper and lower clipping bounds both fluctuate during training, confirming that BAPO dynamically adjusts the clipping for both types of data and adaptively balances their contributions to the loss. In contrast to approaches such as DAPO (Yu et al., 2025) or the asymmetric clipping in Section 4.1, which rely on empirical tuning, BAPO eliminates the need for complex manual hyperparameter tuning, making it simple yet effective.

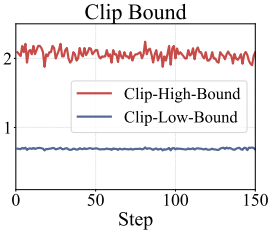

Figure 8: Clipping bounds.

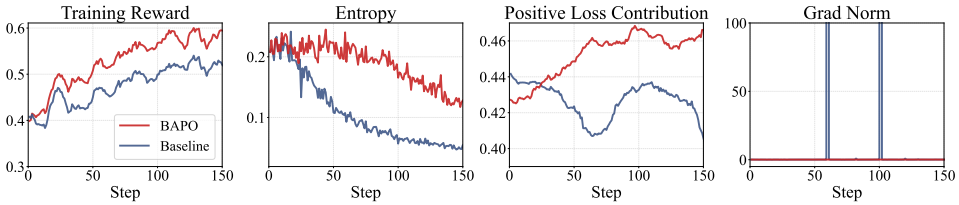

Figure 9: Training dynamics of BAPO.

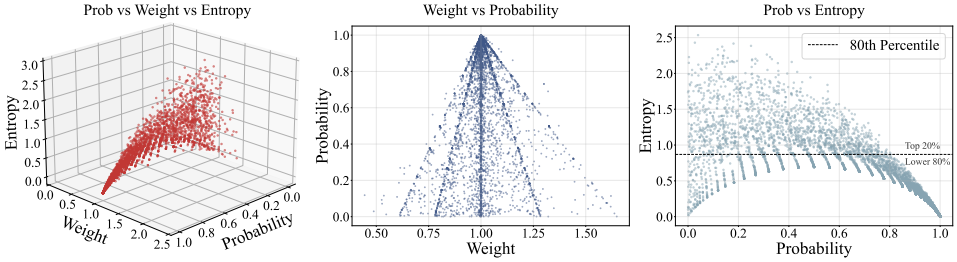

Figure 10: Relationship among token probabilities, importance sampling weights, and entropy.

Effectiveness of BAPO across different staleness. We conduct experiments using the R1-Distill model (Guo et al., 2025) on the SkyWork-OR1-RL dataset (He et al., 2025), with a maximum sequence length of $32k$. The results in Figure 11 show that under different data staleness, our method consistently outperforms both the baseline and the clip-higher approach, demonstrating its superiority.

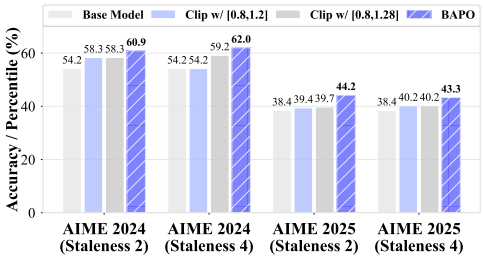

Figure 11: Results with different data staleness.

The working mechanism of BAPO and its connection to prior work. To better understand the working mechanism of BAPO, we present the relationship among token probabilities, IS weights, and entropy during training in Figure 10. We find that as IS weights deviate further from 1, the corresponding token probabilities decrease, and such low-probability tokens often exhibit higher entropy. Based on this observation, we explain how BAPO relates to prior work. For example, Clip-Higher in Yu et al. (2025) sets the clipping upper bound to 1.28, thereby including more low-probability positive tokens in training, which stabilizes entropy while balancing the contributions of positive and negative tokens. Similarly, Wang et al. (2025a) retain only the top 20% highest-entropy tokens for training, ensuring stable entropy throughout optimization and preserving the model's exploratory capability, and the target entropy technique in He et al. (2025) plays a similar role, which aligns with our motivation.

5 Experiments and Discussion

5.1 Experimental Setups

Datasets and Models. We use SkyWork-OR1-RL-Data (He et al., 2025) as our RL dataset, as it is widely adopted and of high quality. For evaluation, we employ both the AIME 2024 and the newly released AIME 2025 (AIME, 2025) benchmarks. Our experiments cover a range of backbone models, including `DeepSeek-R1-Distill-Qwen-7B`, `DeepSeek-R1-Distill-Qwen-32B` (Guo et al., 2025), and `OctoThinker-Llama3.2-3B-Long-Zero` (Wang et al., 2025b). In addition, we incorporate two our own supervised fine-tuning (SFT) models, `BP-Math-7B` and `BP-Math-32B`, which are derived from `Qwen2.5-Math` (Yang et al., 2024) through fine-tuning.

Implementation details. We leverage GRPO as the basis for BAPO. Both our preliminary and validation experiments are conducted using `DeepSeek-R1-Distill-Qwen-7B`, with the maximum response length set to $8k$, learning rate to 2×10^{-6}, and temperature to 0.6. For main results

Table 1: Main evaluation results.

Model	Model Size	AIME 2024	AIME 2025	Average
≥ 100B Models and Proprietary Models				
Qwen3-235B-A22B (Yang et al., 2025a)	235B	85.7	81.5	83.6
DeepSeek-R1 (Guo et al., 2025)	671B	79.8	70.0	74.9
DeepSeek-R1-0528 (Guo et al., 2025)	671B	91.4	87.5	89.5
$o1_{medium}$ (Jaech et al., 2024)	-	83.3	79.0	81.2
$o3\text{-}mini_{medium}$ (OpenAI, 2025)	-	79.6	76.7	78.2
$o3\text{-}mini_{high}$ (OpenAI, 2025)	-	87.3	86.5	86.9
$Gemini\text{-}2.0_{Flash\text{-}Thinking}$ (Google, 2024)	-	73.3	53.5	63.4
$Gemini\text{-}2.5_{Flash\text{-}Thinking\text{-}0520}$ (Comanici et al., 2025)	-	82.3	72.0	77.2
10B - 100B Scale Models				
Qwen3-30B-A3B (Yang et al., 2025a)	30B	–	61.3	–
R1-Distill-Qwen-32B (Guo et al., 2025)	32B	72.6	54.9	63.8
QwQ-32B (Qwen, 2025)	32B	79.5	65.3	72.4
Qwen3-32B (Yang et al., 2025a)	32B	81.4	72.9	77.2
SkyWork-OR1-32B (He et al., 2025)	32B	82.2	73.3	77.8
$BP\text{-}Math\text{-}32B_{SFT}$	32B	84.4	78.1	81.3
$BP\text{-}Math\text{-}32B_{GRPO}$	32B	84.6	78.8	81.7
$BP\text{-}Math\text{-}32B_{BAPO}$	32B	**87.1**	**80.0**	**83.5**
≤ 10B Models				
R1-Distill-Qwen-7B (Guo et al., 2025)	7B	54.2	38.4	46.3
Light-R1-7B-DS (Wen et al., 2025)	7B	59.1	44.2	51.7
AReaL-boba-RL-7B (Fu et al., 2025)	7B	61.9	48.3	55.1
AceReason-Nemotron-7B (Chen et al., 2025)	7B	69.0	53.6	61.3
SkyWork-OR1-7B (He et al., 2025)	7B	70.2	54.6	62.4
$BP\text{-}Math\text{-}7B_{SFT}$	7B	66.9	59.0	62.9
$BP\text{-}Math\text{-}7B_{GRPO}$	7B	69.2	59.2	64.2
$BP\text{-}Math\text{-}7B_{BAPO}$	7B	**70.8**	**62.5**	**66.7**

on BP-Math models, we set the maximum response length to $64k$ to align with the SFT setting. To introduce staleness, we adopt multiple strategies, including experience reuse through ppo_epoch (Schulman et al., 2017) and the modern partial rollouts (Team et al., 2025; Fu et al., 2025). For BAPO, we set the target contribution $\rho_0 = 0.4$, the movable range $a^- = 0.6$, $b^- = 0.9$, $a^+ = 1.2$, $b^+ = 3.0$, and the step size $\delta_1 = 0.05$, $\delta_2 = 0.02$. These hyperparameters are not finely tuned, as they already demonstrate strong empirical performance. For evaluation, we report results averaged over 16 rollouts.

Baselines. We include a variety of commercial and open-source models of different scales as baselines, as shown in Table 1, and report their performance as extracted from prior work. In addition, we compare different training approaches, including SFT and GRPO.

5.2 Main Results

The main results are shown in Figure 1 and Table 1.

Significant performance improvements across models of varying sizes. For strong SFT models, GRPO provides only marginal benefits—for instance, it improves performance by just 0.2 and 0.7 points on AIME24 and AIME25 with the BP-Math-32B model. In contrast, BAPO delivers substantial gains across models of different scales. Specifically, with the BP-Math-32B model, BAPO outperforms SFT by 2.7 and 1.9 points on AIME24 and AIME25, respectively; with the BP-Math-7B model, it achieves even larger improvements of 3.9 and 3.5 points.

SOTA performance over open-source models of comparable sizes and competitive results against proprietary models. Compared to open-source models of similar sizes, our BAPO-trained models achieve state-of-the-art (SOTA) performance. For instance, among 32B models, $BP\text{-}Math\text{-}32B_{BAPO}$ outperforms Qwen3-32B by 5.7 and 7.1 points on AIME24 and AIME25, respectively, and surpasses SkyWork-OR1-32B by 4.9 and 6.7 points. Among 7B models, $BP\text{-}Math\text{-}7B_{BAPO}$ also delivers a notable 7.9-point improvement over SkyWork-OR1-7B on AIME25.

Moreover, $BP\text{-}Math\text{-}32B_{BAPO}$ even outperforms some larger-scale models—for example, it surpasses DeepSeek-R1 by 7.3 and 10.0 points on AIME24 and AIME25, respectively—while achieving performance comparable to o3-mini. Notably, even the smaller $BP\text{-}Math\text{-}7B_{BAPO}$ yields results on par with Gemini-2.0-Flash-Thinking, underscoring the competitiveness of our approach against commercial models.

5.3 Discussion

Partial rollout. To speed up rollouts in LLM reinforcement learning, modern AI infrastructures have introduced several techniques, with partial rollout being particularly noteworthy (Team et al., 2025; Fu et al., 2025). In this approach, long trajectories are split into segments: when a rollout exceeds a fixed token budget, the unfinished portion is stored in a replay buffer and resumed in later iterations instead of being regenerated from scratch. While this improves training efficiency, it also introduces off-policy learning, since different parts of the same trajectory may come from multiple outdated policies. We evaluate BAPO under this setting, as shown in Figure 12. Compared to the baseline GRPO, BAPO exhibits greater robustness to such off-policy infrastructures and achieves more stable optimization.

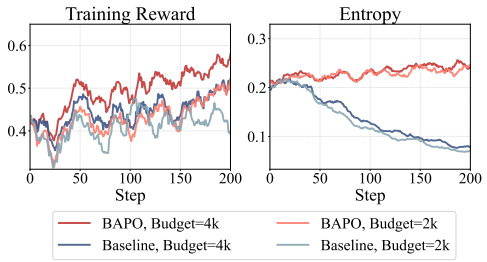

Figure 12: Training dynamics with partial rollout.

Results on OctoThinker-Llama3.2-3B-Long-Zero. In addition to the DeepSeek-R1-Distill-Qwen, we also conducted experiments on Llama-based models (Wang et al., 2025b). As shown in Table 2 and Figure 13 in Appendix B, our method achieves more competitive results and exhibits greater stability in training dynamics.

Table 2: Performance of Llama-based models.

Method	AIME 2024	AIME 2025	MATH
GRPO	2.5%	2.9%	58.4%
BAPO	5.4%	5.8%	66.0%

6 Related Work

Recent landmark models, like OpenAI o1 (Jaech et al., 2024), DeepSeek-R1 (Guo et al., 2025), Gemini 2.5 (Comanici et al., 2025), QwQ (Qwen, 2025), have demonstrated that reinforcement learning can effectively enable long chain-of-thought reasoning in LLMs (Shao et al., 2024; Zhang et al., 2025). Mainstream algorithms include PPO (Schulman et al., 2017) and GRPO (Shao et al., 2024): PPO constrains updates via a clipping-based surrogate objective, while GRPO enhances long-horizon reasoning through group-based rewards.

Despite the remarkable success of RL for LLMs, ensuring stability and efficiency in optimization remains a major challenge (Yu et al., 2025; Cui et al., 2025). Recent studies have sought to better understand the underlying mechanisms of RL and proposed new methods to achieve a balance (Cui et al., 2025; Zheng et al., 2025; Wang et al., 2025a; Yang et al., 2025b). For example, DAPO (Yu et al., 2025) introduces techniques such as Clip-Higher and dynamic sampling to raise the performance ceiling; Wang et al. (2025a) explore optimizing only a small subset of high-entropy tokens for improved efficiency. He et al. (2025), Cui et al. (2025), and other works (Zheng et al., 2025; Cheng et al., 2025; Liu et al., 2025) systematically investigate how to maintain entropy stability during training, thereby preserving the model's exploration ability. For off-policy RL, Roux et al. (2025) and Arnal et al. (2025) introduce asymmetric clipping mechanisms. The most similar to our work is DCPO (Yang et al., 2025b), which adjusts token-level clipping based on token prior probabilities. However, our approach takes a holistic optimization perspective: we observe the imbalance in loss contributions and derive the Entropy-Clip Rule for the PPO objective, enabling dynamic control over global clipping bounds. We further validate the effectiveness of our method through larger-scale experiments.

7 Conclusion

In this paper, we begin by analyzing the impact of data staleness on model training through both empirical and theoretical studies. We reveal the imbalance between positive and negative samples in RL optimization, and derive as well as empirically validate the Entropy-Clip Rule for PPO-like objectives. Building on these insights, we propose BAPO, which dynamically adjusts the clipping bounds to balance positive and negative samples while preserving the model's exploratory capability during training. We conduct extensive experiments across different models and settings to validate our method. We hope our work provides key insights for the LLM RL community.

Ethics Statement

This research introduces an RL methodology designed to augment reasoning capabilities. However, we recognize that it may inadvertently strengthen other capabilities, including those with potential for malicious use. We firmly state that this work is intended for ethical and constructive purposes. Users of this method bear the full responsibility for ensuring it is applied in a safe, fair, and harmless manner. Any misuse of this method is strictly against the intent of the authors.

Reproducibility Statement

We have describe our method and the hyperparameters in §4 and §5. To support reproducibility, we will open-source our code. The datasets used for RL experiments are already publicly available.

References

AIME. Aime problems and solution, 2025. URL https://artofproblemsolving.com/wiki/index.php/AIME_Problems_and_Solutions.

Anthropic. Claude code, 2025. URL https://docs.anthropic.com/en/docs/claude-code.

Charles Arnal, Gaëtan Narozniak, Vivien Cabannes, Yunhao Tang, Julia Kempe, and Rémi Munos. Asymmetric REINFORCE for off-policy reinforcement learning: Balancing positive and negative rewards. *CoRR*, abs/2506.20520, 2025. doi: 10.48550/ARXIV.2506.20520. URL https://doi.org/10.48550/arXiv.2506.20520.

Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, and Haiming Wang. Kimi K2: open agentic intelligence. *CoRR*, abs/2507.20534, 2025. doi: 10.48550/ARXIV.2507.20534. URL https://doi.org/10.48550/arXiv.2507.20534.

Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Acereason-nemotron: Advancing math and code reasoning through reinforcement learning. *CoRR*, abs/2505.16400, 2025. doi: 10.48550/ARXIV.2505.16400. URL https://doi.org/10.48550/arXiv.2505.16400.

Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. Reasoning with exploration: An entropy perspective. *CoRR*, abs/2506.14758, 2025. doi: 10.48550/ARXIV.2506.14758. URL https://doi.org/10.48550/arXiv.2506.14758.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit S. Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilaï Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell Sangnie Bhardwaj, Pawel

Janus, Tero Rissa, Dan Horgan, Sharon Silver, Ayzaan Wahid, Sergey Brin, Yves Raimond, Klemen Kloboves, Cindy Wang, Nitesh Bharadwaj Gundavarapu, Ilia Shumailov, Bo Wang, Mantas Pajarskas, Joe Heyward, Martin Nikoltchev, Maciej Kula, Hao Zhou, Zachary Garrett, Sushant Kafle, Sercan Arik, Ankita Goel, Mingyao Yang, Jiho Park, Koji Kojima, Parsa Mahmoudieh, Koray Kavukcuoglu, Grace Chen, Doug Fritz, Anton Bulyenov, Sudeshna Roy, Dimitris Paparas, Hadar Shemtov, Bo-Juen Chen, Robin Strudel, David Reitter, Aurko Roy, Andrey Vlasov, Changwan Ryu, Chas Leichner, Haichuan Yang, Zelda Mariet, Denis Vnukov, Tim Sohn, Amy Stuart, Wei Liang, Minmin Chen, Praynaa Rawlani, Christy Koh, JD Co-Reyes, Guangda Lai, Praseem Banzal, Dimitrios Vytiniotis, Jieru Mei, and Mu Cai. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025. doi: 10.48550/ARXIV.2507.06261. URL https://doi.org/10.48550/arXiv.2507.06261.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng, Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *CoRR*, abs/2505.22617, 2025. doi: 10.48550/ARXIV.2505.22617. URL https://doi.org/10.48550/arXiv.2505.22617.

Yiwen Ding, Zhiheng Xi, Wei He, Lizhuoyuan Lizhuoyuan, Yitao Zhai, Shi Xiaowei, Xunliang Cai, Tao Gui, Qi Zhang, and Xuanjing Huang. Mitigating tail narrowing in LLM self-improvement via socratic-guided sampling. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 10627–10646. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.533. URL https://doi.org/10.18653/v1/2025.naacl-long.533.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *CoRR*, abs/2505.24298, 2025. doi: 10.48550/ARXIV.2505.24298. URL https://doi.org/10.48550/arXiv.2505.24298.

Google. Introducing gemini 2.0: our new ai model for the agentic era, December 2024. URL https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/.

Çaglar Gülçehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, Wolfgang Macherey, Arnaud Doucet, Orhan Firat, and Nando de Freitas. Reinforced self-training (rest) for language modeling. *CoRR*, abs/2308.08998, 2023. doi: 10.48550/ARXIV.2308.08998. URL https://doi.org/10.48550/arXiv.2308.08998.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou,

Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Jujie He, Jiacai Liu, Chris Yuhao Liu, Rui Yan, Chaojie Wang, Peng Cheng, Xiaoyu Zhang, Fuxiang Zhang, Jiacheng Xu, Wei Shen, Siyuan Li, Liang Zeng, Tianwen Wei, Cheng Cheng, Bo An, Yang Liu, and Yahui Zhou. Skywork open reasoner 1 technical report. *CoRR*, abs/2505.22312, 2025. doi: 10.48550/ARXIV.2505.22312. URL https://doi.org/10.48550/arXiv.2505.22312.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. *CoRR*, abs/2412.16720, 2024. doi: 10.48550/ARXIV.2412.16720. URL https://doi.org/10.48550/arXiv.2412.16720.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *CoRR*, abs/2503.20783, 2025. doi: 10.48550/ARXIV.2503.20783. URL https://doi.org/10.48550/arXiv.2503.20783.

OpenAI. Openai o3-mini system card, 2025. URL https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

Qwen. Qwq-32b: Embracing the power of reinforcement learning, March 2025. URL https://qwenlm.github.io/blog/qwq-32b/.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alexandre Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Tóth, and Sam Work. Tapered off-policy REINFORCE: stable and efficient reinforcement learning for llms. *CoRR*, abs/2503.14286, 2025. doi: 10.48550/ARXIV.2503.14286. URL https://doi.org/10.48550/arXiv.2503.14286.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open

language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, and Will Dabney. Understanding the performance gap between online and offline alignment algorithms. *CoRR*, abs/2405.08448, 2024. doi: 10.48550/ARXIV.2405.08448. URL https://doi.org/10.48550/arXiv.2405.08448.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms. *CoRR*, abs/2501.12599, 2025. doi: 10.48550/ARXIV.2501.12599. URL https://doi.org/10.48550/arXiv.2501.12599.

Luong Quoc Trung, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. Reft: Reasoning with reinforced fine-tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 7601–7614. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.410. URL https://doi.org/10.18653/v1/2024.acl-long.410.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui 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. *CoRR*, abs/2506.01939, 2025a. doi: 10.48550/ARXIV.2506.01939. URL https://doi.org/10.48550/arXiv.2506.01939.

Zengzhi Wang, Fan Zhou, Xuefeng Li, and Pengfei Liu. Octothinker: Mid-training incentivizes reinforcement learning scaling. *CoRR*, abs/2506.20512, 2025b. doi: 10.48550/ARXIV.2506.20512. URL https://doi.org/10.48550/arXiv.2506.20512.

Liang Wen, Yunke Cai, Fenrui Xiao, Xin He, Qi An, Zhenyu Duan, Yimin Du, Junchen Liu, Lifu Tang, Xiaowei Lv, Haosheng Zou, Yongchao Deng, Shousheng Jia, and Xiangzheng Zhang. Light-r1: Curriculum sft, DPO and RL for long COT from scratch and beyond. *CoRR*, abs/2503.10460, 2025. doi: 10.48550/ARXIV.2503.10460. URL https://doi.org/10.48550/arXiv.2503.10460.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8:229–256, 1992. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

Zhiheng Xi, Wenxiang Chen, Boyang Hong, Senjie Jin, Rui Zheng, Wei He, Yiwen Ding, Shichun Liu, Xin Guo, Junzhe Wang, Honglin Guo, Wei Shen, Xiaoran Fan, Yuhao Zhou, Shihan Dou, Xiao Wang, Xinbo Zhang, Peng Sun, Tao Gui, Qi Zhang, and Xuanjing Huang. Training large language models for reasoning through reverse curriculum reinforcement learning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=t82Y3fmRtk.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu,

Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024. doi: 10.48550/ARXIV.2409.12122. URL https://doi.org/10.48550/arXiv.2409.12122.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025a. doi: 10.48550/ARXIV.2505.09388. URL https://doi.org/10.48550/arXiv.2505.09388.

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

Zhihe Yang, Xufang Luo, Zilong Wang, Dongqi Han, Zhiyuan He, Dongsheng Li, and Yunjian Xu. Do not let low-probability tokens over-dominate in RL for llms. *CoRR*, abs/2505.12929, 2025c. doi: 10.48550/ARXIV.2505.12929. URL https://doi.org/10.48550/arXiv.2505.12929.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476, 2025. doi: 10.48550/ARXIV.2503.14476. URL https://doi.org/10.48550/arXiv.2503.14476.

Ruifeng Yuan, Chenghao Xiao, Sicong Leng, Jianyu Wang, Long Li, Weiwen Xu, Hou Pong Chan, Deli Zhao, Tingyang Xu, Zhongyu Wei, Hao Zhang, and Yu Rong. Vl-cogito: Progressive curriculum reinforcement learning for advanced multimodal reasoning. *CoRR*, abs/2507.22607, 2025. doi: 10.48550/ARXIV.2507.22607. URL https://doi.org/10.48550/arXiv.2507.22607.

Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization. *CoRR*, abs/2507.18071, 2025. doi: 10.48550/ARXIV.2507.18071. URL https://doi.org/10.48550/arXiv.2507.18071.

Xuekai Zhu, Daixuan Cheng, Dinghuai Zhang, Hengli Li, Kaiyan Zhang, Che Jiang, Youbang Sun, Ermo Hua, Yuxin Zuo, Xingtai Lv, et al. Flowrl: Matching reward distributions for llm reasoning. *arXiv preprint arXiv:2509.15207*, 2025.

A The Use of Large Language Models

LLMs are utilized in this manuscript for partial grammatical checks and language polishing. The authors are fully responsible for the final content.

B Performance on OctoThinker-Llama

We illustrate the training dynamics on OctoThinker-Llama in Figure 13. Since Llama family models behave badly in RL training, we choose the model after mid-training (Wang et al., 2025b) to show

the robustness of BAPO. We can find that BAPO provides consistent and significant improvement in training. For training details, we set the low bound as 0.8-0.9, high bound as 1.2-2.0, and target positive loss contribution as 0.45.

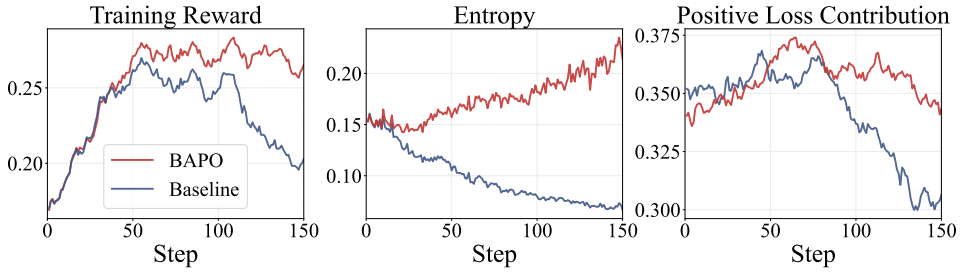

Figure 13: Training dynamics of OctoThinker-Llama-3B-Long-Zero.

C Proofs of Equation 6

C.1 Explanations for all variables and expressions

All notation used in the following justification, including variables and expressions, is provided with detailed explanations in Table 3.

C.2 Preparation: Rewrite the PPO derivatives

To facilitate the justification of the propositions below, we rewrite the PPO loss function in the following form:

$$\nabla J^{\text{PPO}} = \underbrace{\sum_{A(y_t)>0} \pi_\theta(y_t) \cdot \mathbb{I}\{r(y_t) < 1+\varepsilon\} \cdot A(y_t) \cdot \nabla \log \pi_\theta(y_t)}_{\text{positive tokens}} + \underbrace{\sum_{A(y_t)<0} \pi_\theta(y_t) \cdot \mathbb{I}\{r(y_t) > 1-\varepsilon\} \cdot A(y_t) \cdot \nabla \log \pi_\theta(y_t)}_{\text{negative tokens}}$$

where

$$\pi_\theta(y_t) = \pi_\theta(y_t|\boldsymbol{x}, \boldsymbol{y_{<t}}) \ , \ r(y_t) = \frac{\pi_\theta(y_t \mid \boldsymbol{x}, \boldsymbol{y_{<t}})}{\pi_{\theta_{\text{rollout}}}(y_t \mid \boldsymbol{x}, \boldsymbol{y_{<t}})} \ , \ A(y_t) = A(y_t|\boldsymbol{x}, \boldsymbol{y_{<t}}) \ .$$

C.3 Proofs of the main Propositions

The following derivation is inspired by the proof framework in Cui et al. (2025). While the original work focuses mainly on the basic gradient formulation of naive REINFORCE to provide a heuristic explanation, our study advances this approach by deriving the gradient expression **specific to the PPO objective**. This refinement offers a specific, **intuitive yet theoretical** account of how policy entropy is intrinsically shaped by the interaction between token-level advantages and their sampling probabilities.

C.3.1 Preclaims

Proofs of these three lemmas below are available in Cui et al. (2025).

Lemma 1. *Let the actor policy π_θ be a tabular softmax policy, the difference of information entropy given prompt x between two consecutive steps k and $k+1$ satisfies*

$$\mathcal{H}(\pi_\theta^{k+1}|\boldsymbol{x}, \boldsymbol{y_{<t}}) - \mathcal{H}(\pi_\theta^{k}|\boldsymbol{x}, \boldsymbol{y_{<t}}) \approx -\operatorname{Cov}_{y_t \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})} \left(\log \pi_\theta^k(y_t),\ z_{\boldsymbol{y},\boldsymbol{x}}^{k+1} - z_{\boldsymbol{y},\boldsymbol{x}}^{k}\right).$$

Table 3: Notation used in justification below.

Category	Symbol	Meaning
Variables	π_θ	The policy parameterized by θ
	$\pi_{\theta_{rollout}}$	The standard sampling policy
	$\boldsymbol{x}$	Given prompt
	$\boldsymbol{y}$	A T-token response generated by π_θ when given $\boldsymbol{x}$
	y_t	The t-th token of y
	η	Learning rate
Expressions	$\pi_\theta(\cdot\|\boldsymbol{x}, \boldsymbol{y}_{<t})$	Probability of generating token $\cdot$ under policy π_θ given input $\boldsymbol{x}$ and previous tokens $\boldsymbol{y}_{<t}$
	$\pi_{\theta_{rollout}}(y_t\|\boldsymbol{x}, \boldsymbol{y}_{<t})$	Probability of generating token $\cdot$ under standard sampling policy $\pi_{\theta_{rollout}}$ given input $\boldsymbol{x}$ and previous tokens $\boldsymbol{y}_{<t}$
	$A(\cdot\|\boldsymbol{x}, \boldsymbol{y_{<t}})$	The measurement of how much better(or worse) selecting token $\cdot$ is compared to the expected value under the current policy, given $\boldsymbol{x}$ and $\boldsymbol{y_{<t}}$
	$\mathcal{H}(\cdot\|\boldsymbol{x}, \boldsymbol{y}_{<t})$	The information entropy of policy $\cdot$ given $\boldsymbol{x}$ and $\boldsymbol{y}_{<t}$
	$Cov_{y_t\sim\pi_\theta(\cdot\|\boldsymbol{x},\boldsymbol{y}_{<t})}(a(y_t), b(y_t))$	The expected covariance of $a(y_t)$ and $b(y_t)$ over y_t sampled from the policy π_θ, given $\boldsymbol{x}$ and $\boldsymbol{y}_{<t}$
	$\mathbb{I}(a = b)$	Indicator function that equals 1 if $a = b$ and 0 otherwise
	$Q^{(\pi_\theta)}(\cdot, \boldsymbol{x})$	The expected cumulative reward obtained by taking token $\cdot$ given input $\boldsymbol{x}$ and previous tokens under policy π_θ
	$V^{(\pi_\theta)}(\boldsymbol{x})$	The expected return of the new taking token given input $\boldsymbol{x}$ and previous tokens under policy π_θ
	$z_{\boldsymbol{y},\boldsymbol{x}}$	A quantity representing the cumulative weight of sequence $\boldsymbol{y}$ given input $\boldsymbol{x}$ under policy π_θ, reflecting its contribution to the policy taken at the current optimization step
	$\nabla_{\theta_{y_t,\boldsymbol{x}}} J(\theta)$	The gradient of the policy taken with respect to the logit parameter $\theta_{y_t,\boldsymbol{x}}$, representing how the policy π_θ should be adjusted for token y_t given input $\boldsymbol{x}$

Lemma 2 (Derivative of softmax function)**.**

$$\frac{\partial \log \pi_\theta(y_t)}{\partial \theta_{y_t',\boldsymbol{x}}} = \mathbb{I}\{y_t = y_t'\} - \pi_\theta(y_t')$$

Lemma 3 (Expectation of Advantage function given prompt x)**.**

$$\begin{aligned}\mathbb{E}_{y_t\sim\pi_\theta(\cdot|x,\boldsymbol{y_{<t}})}\big[A^{\pi_\theta}(y_t)\big] &= \mathbb{E}_{y_t\sim\pi_\theta(\cdot|x,\boldsymbol{y_{<t}})}\big[Q^{\pi_\theta}(y_t, \boldsymbol{x}) - V^{\pi_\theta}(\boldsymbol{x})\big] \\ &= \mathbb{E}_{y_t\sim\pi_\theta(\cdot|x,\boldsymbol{y_{<t}})}\big[Q(y_t, \boldsymbol{x})\big] - \mathbb{E}_{y_t\sim\pi_\theta(\cdot|x,\boldsymbol{y_{<t}})}\big[V(\boldsymbol{x})\big] \\ &= V(\boldsymbol{x}) - V(\boldsymbol{x}) \\ &= 0\end{aligned}$$

C.3.2 Principle Propositions

Proposition 1: Assume the actor policy π_θ follows a tabular softmax policy and is optimized via the PPO objective, the difference of $z_{\boldsymbol{y},\boldsymbol{x}}$ between two consecutive steps k and k+1 satisfies

$$z^{k+1}_{\boldsymbol{y},\boldsymbol{x}} - z^{k}_{\boldsymbol{y},\boldsymbol{x}} = \eta \cdot \pi_\theta(y_t) \cdot [A(y_t) \cdot \mathcal{X}(y_t) + C],$$

where

$$\mathcal{X}(y_t) = \begin{cases} 1, & \textbf{\textit{if}}\ A(y_t) > 0\ \&\ r(y_t) < 1+\epsilon \\ & \textbf{\textit{or}}\ A(y_t) < 0\ \&\ r(y_t) > 1-\epsilon \\ 0, & \text{otherwise} \end{cases}$$

and C includes all clauses irrelevant to y_t.

It is worth noting that $\boldsymbol{\mathcal{X}(y_t) = 0}$ if and only if $\boldsymbol{y_t}$ is clipped.

Proof. In tabular softmax policy, each trajectory-prompt pair $(\boldsymbol{y}, \boldsymbol{x})$ is associated with an individual logit parameter $z_{\boldsymbol{y},\boldsymbol{x}} = \theta_{y_t,\boldsymbol{x}}$. Through gradient backtracking, $z_{\boldsymbol{y},\boldsymbol{x}}$ is updated via $z^{k+1}_{\boldsymbol{y},\boldsymbol{x}} = z^{k}_{\boldsymbol{y},\boldsymbol{x}} + \eta \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} J(\theta)$. According to the loss function of PPO, we have

$$\begin{aligned}
z^{k+1}_{\boldsymbol{y},\boldsymbol{x}} - z^{k}_{\boldsymbol{y},\boldsymbol{x}} &= \eta \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} J_{PPO}(\theta) \\
&= \eta \cdot \mathbb{E}_{\substack{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}}) \\ A(y'_t)>0}} \left[\mathbb{I}\{r(y'_t) < 1+\varepsilon\} \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} \log \pi_\theta(y'_t) \cdot A(y'_t)\right] \\
&+ \eta \cdot \mathbb{E}_{\substack{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}}) \\ A(y'_t)<0}} \left[\mathbb{I}\{r(y'_t) > 1-\varepsilon\} \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} \log \pi_\theta(y'_t) \cdot A(y'_t)\right] \\
&= \underbrace{\eta \cdot \mathbb{E}_{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}})} \left[\nabla_{\theta_{y_t,\boldsymbol{x}}} \log \pi_\theta(y'_t) \cdot A(y'_t)\right]}_{①} \\
&\underbrace{- \eta \cdot \mathbb{E}_{\substack{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}}) \\ A(y'_t)>0}} \left[\mathbb{I}\{r(y'_t) > 1+\varepsilon\} \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} \log \pi_\theta(y'_t) \cdot A(y'_t)\right]}_{②} \\
&\underbrace{- \eta \cdot \mathbb{E}_{\substack{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}}) \\ A(y'_t)<0}} \left[\mathbb{I}\{r(y'_t) < 1-\varepsilon\} \cdot \nabla_{\theta_{y_t,\boldsymbol{x}}} \log \pi_\theta(y'_t) \cdot A(y'_t)\right]}_{③} \\
&= ① - (② + ③)
\end{aligned} \tag{8}$$

We first perform the derivation on the term marked as ①:

$$\begin{aligned}
① &= \eta \cdot \mathbb{E}_{y'_t \sim \pi_\theta(\cdot|\boldsymbol{x},\boldsymbol{y_{<t}})} \left[\frac{\partial \log \pi_\theta(y'_t)}{\partial \theta_{y_t,\boldsymbol{x}}} \cdot A(y'_t)\right] \\
&\overset{\text{Lemma 2}}{=} \eta \cdot \sum_{y'_t} \left[\pi_\theta(y'_t) \cdot (\mathbb{I}\{y'_t = y_t\} - \pi_\theta(y_t)) \cdot A(y'_t)\right] \\
&= \eta \cdot \pi_\theta(y_t) \cdot \left[(1 - \pi_\theta(y_t)) \cdot A(y_t) - \sum_{y'_t \neq y_t} \pi_\theta(y'_t) \cdot A(y'_t)\right] \\
&= \eta \cdot \pi_\theta(y_t) \cdot \left[A(y_t) - \sum_{y'_t} \pi_\theta(y'_t) \cdot A(y'_t)\right] \\
&\overset{\text{Lemma 3}}{=} \eta \cdot \pi_\theta(y_t) \cdot [A(y_t) - 0] \\
&= \eta \cdot \pi_\theta(y_t) \cdot A(y_t)
\end{aligned}$$

To keep the presentation concise, we provide only the resulting derivations of Term ② and ③, as the detailed steps follow similarly to those for Term ①.

$$
\begin{aligned}
②+③ &= \eta \cdot \pi_\theta(y_t) \cdot A(y_t) \cdot (1 - \mathcal{X}(y_t)) \\
&\quad - \eta \cdot \pi_\theta(y_t) \cdot \sum_{A(y_t')>0} [\mathbb{I}\{r(y_t') > 1+\varepsilon\} \cdot \pi_\theta(y_t') \cdot A(y_t')] \\
&\quad - \eta \cdot \pi_\theta(y_t) \cdot \sum_{A(y_t')<0} [\mathbb{I}\{r(y_t') < 1-\varepsilon\} \cdot \pi_\theta(y_t') \cdot A(y_t')]
\end{aligned}
$$

By substituting the results of the above derivation into Clause (8), we observe that:

$$
\begin{aligned}
(8) &= ① - (②+③) \\
&= \eta \cdot \pi_\theta(y_t) \cdot \Big\{ A(y_t) \cdot \mathcal{X}(y_t) \\
&\quad + \sum_{A(y_t')>0} [\mathbb{I}\{r(y_t') > 1+\varepsilon\} \cdot \pi_\theta(y_t') \cdot A(y_t')] \\
&\quad + \sum_{A(y_t')<0} [\mathbb{I}\{r(y_t') < 1-\varepsilon\} \cdot \pi_\theta(y_t') \cdot A(y_t')] \Big\}
\end{aligned}
$$

By grouping all elements unrelated to y_t into C, we are able to successfully establish our proposition.

□

Building on Proposition 1, we establish the relationship between policy entropy and the covariance of specific tokens, which is stated as Proposition 2 below.

Proposition 2 (Equation 6): Let the actor policy π_θ be tabular softmax policy, and π_θ is updated via PPO objective, the difference of information entropy given prompt x and trajectory part $y_{<t}$ between two consecutive steps k and k+1 satisfies

$$
\mathcal{H}(\pi_\theta^{k+1}|\boldsymbol{x}, \boldsymbol{y_{<t}}) - \mathcal{H}(\pi_\theta^{k}|\boldsymbol{x}, \boldsymbol{y_{<t}}) \approx -\eta \cdot \mathrm{Cov}_{y_t \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})} \left(\log \pi_\theta^k(y_t), A(y_t) \cdot \mathcal{X}(y_t) + C \right).
$$

Proof. Leveraging the conclusions of Lemma 1 and Proposition 1, we find that, under policy optimization and iteration via the PPO algorithm, the following relationship is satisfied:

$$
z_{\boldsymbol{y},\boldsymbol{x}}^{k+1} - z_{\boldsymbol{y},\boldsymbol{x}}^{k} = \eta \cdot (A(y_t) \cdot \mathcal{X}(y_t) + C).
$$

Applying this into Lemma 1, we have

$$
\mathcal{H}(\pi_\theta^{k+1}|\boldsymbol{x}, \boldsymbol{y_{<t}}) - \mathcal{H}(\pi_\theta^{k}|\boldsymbol{x}, \boldsymbol{y_{<t}}) \approx -\eta \cdot \mathrm{Cov}_{y_t \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})} \left(\log \pi_\theta^k(y_t), A(y_t) \cdot \mathcal{X}(y_t) + C \right).
$$

□

C.4 Analysis

C.4.1 Direct Analysis: Why Varying ε Alters Entropy?

We begin by examining the covariance of the clipped token, denoted as α.

Based on the observation stated above, the contribution of α to the entropy can be expressed as:

$$
-\eta \cdot \pi_\theta^k(\alpha) \cdot \mathrm{Cov}\big(\log \pi_\theta^k(\alpha), C\big) = 0,
$$

which indicates that only the retained tokens contribute to the overall entropy.

In other words, we manipulate the number of tokens that can contribute to the entropy by altering the parameter ε.

C.4.2 Advanced Analysis : Which Type of Tokens Matter Most for Entropy?

To understand how individual tokens contribute to the overall entropy, we first revisit the Proposition C.3.2 established above. In this section, we provide a more precise definition of tokens with low/high probabilities and advantages. It should be noted that in the analysis experiment (Figure 5), we adopt the naive REINFORCE algorithm without clipping. Consequently, tokens with high or low advantages are defined according to the sign of their advantage values, i.e., > 0 for high advantage and < 0 for low advantage.

$$\begin{aligned}\mathcal{H}(\pi_\theta^{k+1}|\boldsymbol{x}, \boldsymbol{y_{<t}}) - \mathcal{H}(\pi_\theta^{k}|\boldsymbol{x}, \boldsymbol{y_{<t}}) &\approx -\eta \cdot \mathrm{Cov}_{y_t \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})} \left(\log \pi_\theta^k(y_t), A(y_t) \cdot \mathcal{X}(y_t) + C\right) \\ &= -\eta \cdot \sum_{p=1}^{T} \pi_\theta^k(y_p|\boldsymbol{x}, \boldsymbol{y_{<t}}) \cdot \left(\log \pi_\theta^k(y_p) - \mathbb{E}_{y_i \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})}[\log \pi_\theta^k(y_i)]\right) \\ &\quad \cdot \left(A(y_p) \cdot \mathcal{X}(y_p) - \mathbb{E}_{y_i \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})}[A(y_i) \cdot \mathcal{X}(y_i)]\right).\end{aligned}$$

where T is the size of the dictionary.

For convenience, we denote $\mathbb{E}_{y_i}$ as $\mathbb{E}_{y_i \sim \pi_\theta^k(\cdot|\boldsymbol{x}, \boldsymbol{y_{<t}})}$. As only retained tokens contribute to the entropy, we focus only on tokens that are not clipped. We begin by making the following simplification:

$$\mathbb{E}_{y_i}(A(y_i) \cdot \mathcal{X}(y_i)) = \mathbb{E}_{y_{\text{clipped}}}(A(y_i) \cdot 0) + \mathbb{E}_{y_{\text{retained}}}(A(y_i) \cdot 1) = \mathbb{E}_{y_{\text{retained}}}(A(y_i)).$$

So for a selected token y_s, its contribution to the overall entropy can be expressed as:

$$-\eta \cdot \pi_\theta(y_s) \cdot (\log \pi_\theta(y_s) - \mathbb{E}_{y_i}(\log \pi_\theta(y_i))) \cdot (A(y_s) - \mathbb{E}_{y_{\text{retained}}} A(y_{\text{retained}})).$$

Next, we analyze how different types of tokens contribute to the overall entropy. To avoid ambiguity, we first give strict definitions that distinguish between tokens with high/low probabilities and tokens with high/low advantages.

Definition 1. *For a token y_s, we classify it as follows:*

- ***High advantage:*** *if*
$$A(y_s) > \mathbb{E}_{y_{\text{retained}}} A(y_{\text{retained}})$$
Otherwise, it is called low advantage.
- ***High probability:*** *if*
$$\pi_\theta(y_s) > \exp(\mathbb{E}_{y_i}(\log \pi_\theta(y_i)))$$
Otherwise, it is called low probability.

Secondly, we present two propositions that directly follow from the above definitions.

Proposition 3. For a token y_s, we have

$$A(y_s) - \mathbb{E}_{y_{\text{retained}}} A(y_{\text{retained}}) \begin{cases} > 0, & \text{if } y_s \text{ is a high-advantage token,} \\ < 0, & \text{if } y_s \text{ is a low-advantage token.} \end{cases}$$

Proposition 4. For a token y_s, we have

$$\pi_\theta(y_s) \cdot (\log \pi_\theta(y_s) - \mathbb{E}_{y_i}(\log \pi_\theta(y_i))) \begin{cases} > 0, & \text{if } y_s \text{ is a high-probability token,} \\ < 0, & \text{if } y_s \text{ is a low-probability token.} \end{cases}$$

Proof. Let us denote

$$C = \mathbb{E}_{y_i}(\log \pi_\theta(y_i))),$$

which is independent of y_s, and let $x = \pi_\theta(y_s)$. As $\pi_\theta(y) < 1$ for every y, $C < 0$.
Consider the function

$$f(x) = x \cdot (\log(x) - C).$$

Figure 14 illustrates the behavior of this function.

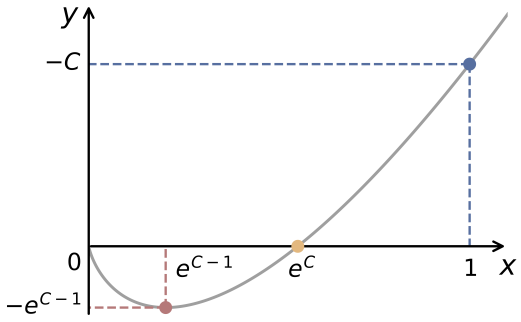

Figure 14: Graph of the function $f(x) = x(\log x - C)$.

The proposition follows directly from the properties of $f(x)$ as observed in the figure. □

Due to the propositions given above, we have the table below:

$$\Delta\mathcal{H}(y_s) \approx -\eta \cdot \underbrace{\pi_\theta(y_s) \cdot (\log \pi_\theta(y_s) - \mathbb{E}_{y_i}(\log \pi_\theta(y_i)))}_{④} \cdot \underbrace{(A(y_s) - \mathbb{E}_{y_{\text{retained}}} A(y_{\text{retained}})}_{⑤}$$

Table 4: Influence of token characteristics on $\Delta\mathcal{H}(y_s)$. The "prob" denotes the probability $\pi_\theta(y_s)$, and the "adv" represents the advantage $A(y_s)$.

Token properties	④	⑤	$\boldsymbol{\Delta\mathcal{H}(y_s)}$ $\boldsymbol{(-\eta \cdot ④ \cdot ⑤)}$
high prob, high adv	> 0	> 0	< 0
high prob, low adv	> 0	< 0	> 0
low prob, high adv	< 0	> 0	> 0
low prob, low adv	< 0	< 0	< 0

It should be noted that a token y_s **decreases** the entropy if $\Delta\mathcal{H}(y_s) < 0$, and **increases** it otherwise.

Therefore, we observe that tokens which are positive with high probabilities and high advantages, or negative with low probabilities and low advantages, contribute to a reduction in the overall entropy. Conversely, positive tokens with high probabilities but low advantages, and negative tokens with high probabilities but low advantages, contribute to an increase in the overall entropy. This observation justifies the statement made in the main part of the thesis.