

Agentic Entropy-Balanced Policy Optimization

Guanting Dong^{1§}, Licheng Bao^{2§}, Zhongyuan Wang^{2§}, Kangzhi Zhao², Xiaoxi Li¹, Jiajie Jin¹, Jinghan Yang^{2§}, Hangyu Mao², Fuzheng Zhang², Kun Gai², Guorui Zhou^{2✉}, Yutao Zhu¹, Ji-Rong Wen¹, Zhicheng Dou^{1✉}

¹Renmin University of China ²Kuaishou Technology

{dongguanting, dou}@ruc.edu.cn

GitHub: <https://github.com/dongguanting/ARPO>

Abstract

Recently, Agentic Reinforcement Learning (Agentic RL) has made significant progress in incentivizing the multi-turn, long-horizon tool-use capabilities of web agents. While mainstream agentic RL algorithms autonomously explore high-uncertainty tool-call steps under the guidance of entropy, excessive reliance on entropy signals can impose further constraints, leading to the training collapse. In this paper, we delve into the challenges caused by entropy and propose the Agentic Entropy-Balanced Policy Optimization (AEPO), an agentic RL algorithm designed to balance entropy in both the rollout and policy update phases. AEPO comprises two core components: (1) a dynamic entropy-balanced rollout mechanism that adaptively allocate global and branch sampling budget through entropy pre-monitoring, while imposing a branch penalty on consecutive high-entropy tool-call steps to prevent over-branching issues; and (2) Entropy-Balanced Policy Optimization that inserts a stop-gradient operation into the high-entropy clipping term to preserve and properly rescale gradients on high-entropy tokens, while incorporating entropy-aware advantage estimation to prioritize learning on high-uncertainty tokens. Results across 14 challenging datasets show that AEPO consistently outperforms 7 mainstream RL algorithms. With just 1K RL samples, Qwen3-14B with AEPO achieves impressive results: **47.6% on GAIA**, **11.2% on Humanity's Last Exam**, and **43.0% on WebWalkerQA for Pass@1**; **65.0% on GAIA**, **26.0% on Humanity's Last Exam**, and **70.0% on WebWalkerQA for Pass@5**. Further analysis reveals that AEPO improves rollout sampling diversity while maintaining stable policy entropy, facilitating scalable web agent training.

Keywords

Agentic Reinforcement Learning, Agentic Search, Web Agent, Tool Learning, Large Language Model

1 Introduction

The emergence of large language models (LLMs) have profoundly revolutionized a wide range of natural language reasoning tasks [3,

§ Work done during internship at Kuaishou.

✉ Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW '26, Dubai, UAE

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YYY/MM

<https://doi.org/xxxxxxxx.xxxxxxx>

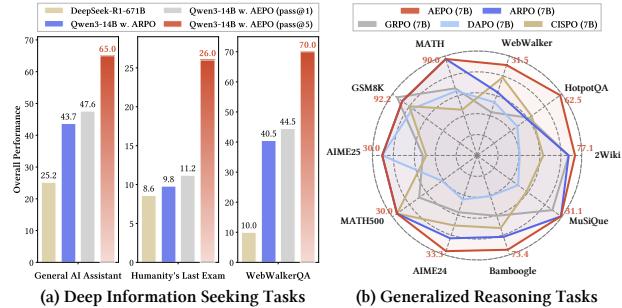


Figure 1: Performance overview of AEPO algorithm.

29, 64, 91–93, 113, 119]. Despite their impressive capabilities, the static nature of their internal knowledge often leads LLMs to experience hallucinations and information staleness in knowledge-intensive scenarios [123]. Retrieval-augmented generation (RAG) addresses these limitations by empowering LLMs to reason with retrieved relevant knowledge, thereby improving the reliability of generated answers [15, 17, 31, 42, 45, 52, 87]. However, with the explosive growth of web information, the static RAG workflow limits effective interaction between LLMs and search engines, resulting in significant bottlenecks in open-domain web exploration. To overcome these challenges, a series of LLM-based web agents have emerged [22, 50, 51]. These agents perform on-demand web searches during reasoning and strategically interact with external tool environments, achieving reliable, in-depth web information seeking [6, 16, 30, 40, 48, 82, 127].

To strive for efficient training of web agents, early implementations focus on distill tool-use trajectories from stronger models and guide weaker models through supervised fine-tuning (SFT) [23, 28, 46, 116]. However, relying solely on SFT struggles to discover autonomous and generalizable tool-use capability [10]. As large-scale reinforcement learning with verifiable rewards (RLVR) demonstrates the potential to unlock frontier LLM capabilities [29, 91, 94], several web-search agents adopt trajectory-level RL [76, 118, 124] combined with carefully designed reward functions to elicit agentic reasoning in LLMs [13, 105]. While effective to some extent, this line of work consistently overlooks the multi-turn interactive nature between LLMs and tool environments [122], making it difficult to discover step-level tool-use behaviors during RL training. To mitigate this limitation, recent efforts in agentic RL have shown that web agents often display high entropy in their output tokens due to uncertainty about the external tool-call results [14]. Drawing on this finding, they introduce a tree-structured rollout method that adaptively branches at high-entropy tool-call steps, effectively broadening sampling diversity and coverage [25, 38, 54, 57].

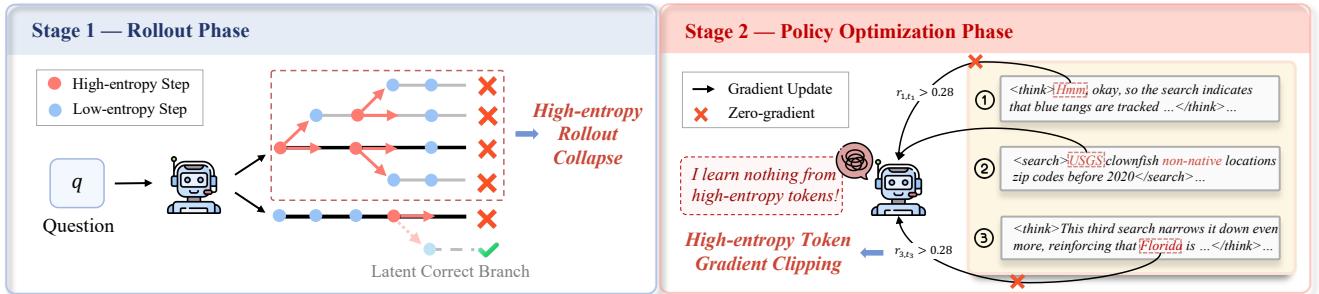


Figure 2: Two high-entropy challenges in agentic RL. (1) **High-Entropy Rollout Collapse:** Over-branching at high-entropy steps along specific paths, limiting exploration of other potential correct branches; (2) **High-Entropy Token Gradient Clipping:** Consistent clipping of high-entropy token gradients during policy updates hinders learning effective exploration behaviors.

Although these entropy-driven agentic RL algorithms stimulate exploration of latent tool-use behaviors, such high-entropy signals further pose two extra challenges for web agent training:

- (1) **High-entropy Rollup Collapse:** During the rollout phase, high-entropy tool-call steps often occur consecutively, leading the LLM to over-branch along a single trajectory under high-entropy guidance. This situation depletes the branching budget for other trajectories at high-entropy steps, ultimately limiting the diversity and scope of rollout sampling (see Figure 2 (left)).
- (2) **High-entropy Token Gradient Clipping:** The tree-structured rollout strategy encourages LLMs to explore step-level tool-use behaviors, thus preserving valuable high-entropy tokens. However, vanilla RL algorithms aggressively clip high-entropy token's gradient during policy update phase, leading to the premature termination of the LLM's exploration (see Figure 2 (right)).

Consequently, efficiently balancing entropy in agentic RL remains a fundamental challenge in the pursuit of generalized agent training.

To address these challenges, we propose **Agentic Entropy-Balanced Policy Optimization (AEPO)**, an entropy-balanced agentic RL algorithm designed for training multi-turn web agents. Unlike traditional entropy-driven RL approaches [14, 55], AEPO focuses on balancing and rationalizing rollout branching and policy updates under the guidance of high-entropy tool calls, thereby achieving more stable RL training. Specifically, we pioneer the quantification of two inherent entropy-driven challenges on agentic RL.

Building on these insights, AEPO introduces two key algorithmic optimizations: (1) **Dynamic Entropy-balanced Rollout Mechanism:** To mitigate “High-entropy Rollout Collapse” issue, AEPO initially proposes the entropy pre-monitoring to adaptively allocate global and branch sampling budget, ensuring balanced exploration across the tree-structured rollout. Moreover, it incorporates a branch penalty strategy for consecutive high-entropy tool-call steps to effectively address over-branching issues in specific chains. (2) **Entropy-Balanced Policy Optimization:** Draw inspiration from recent clipping-optimized works [3, 84], we intuitively integrate a *stop-gradient* operation into the high-entropy clipping term during policy updates to tackle the “High-Entropy Token Gradient Clipping”. This preserves and properly rescales gradients of high-entropy tokens during backpropagation while leaving the forward pass unchanged. Furthermore, AEPO proposes entropy-aware advantage estimation, integrating entropy advantage into vanilla

advantage estimation, enabling the model to prioritize learning on high-uncertainty tokens.

We conduct comprehensive evaluations across 14 datasets covering deep information seeking, knowledge-intensive reasoning, and computational reasoning. As shown in Figure 1, the results show that AEPO consistently outperforms mainstream RL algorithms in generalized reasoning tasks. Remarkably, with only 1k RL training samples, Owen3-14B with AEPO achieves impressive results: **47.6% on GAIA, 11.2% on HLE and 43.0% on WebWalkerQA for Pass@1;** and **65.0% on GAIA, 26.0% on Humanity’s Last Exam and 70.0% on WebWalkerQA for Pass@5.** Further analysis confirms that AEPO effectively broadens sampling diversity during rollouts while maintaining high and stable policy entropy throughout RL training, providing a promising solution for developing general web agents.

In summary, the key contributions are as follows:

- We systematically reveal two entropy-driven issues inherent to agentic RL: “High-Entropy Rollout Collapse” and “High-Entropy Token Gradient Clipping”. Through preliminary experiments, we quantify their impact on multi-turn web-agent training, offering empirical evidence for further research into entropy balancing.
- We propose a **Dynamic Entropy-Balanced Rollout mechanism**, which adaptively allocates rollout sampling budgets via entropy pre-monitoring, while imposing a branch penalty on consecutive high-entropy steps to prevent over-branching issues.
- We introduce **Entropy-Balanced Policy Optimization**, which intuitively integrates a *stop-gradient* operation into the high-entropy clipping term to preserve and rescale gradients on high-entropy tokens, while incorporating entropy-aware advantage estimation to prioritize learning on high-uncertainty tokens.
- Experiments on 14 challenging benchmarks demonstrate that AEPO consistently outperforms mainstream RL algorithms in web agent training. Quantitative analyses across dimensions such as *Pass@k sampling*, *rollout diversity*, *tool-call efficiency* and *entropy dynamics* verify AEPO’s strong scalability and stability, offering valuable insights for developing general web agents.

2 Preliminary

Before introducing the AEPO algorithm, we will briefly outline key task definitions and illustrate preliminary entropy-based experiments to reveal key limitations of web agent RL training.

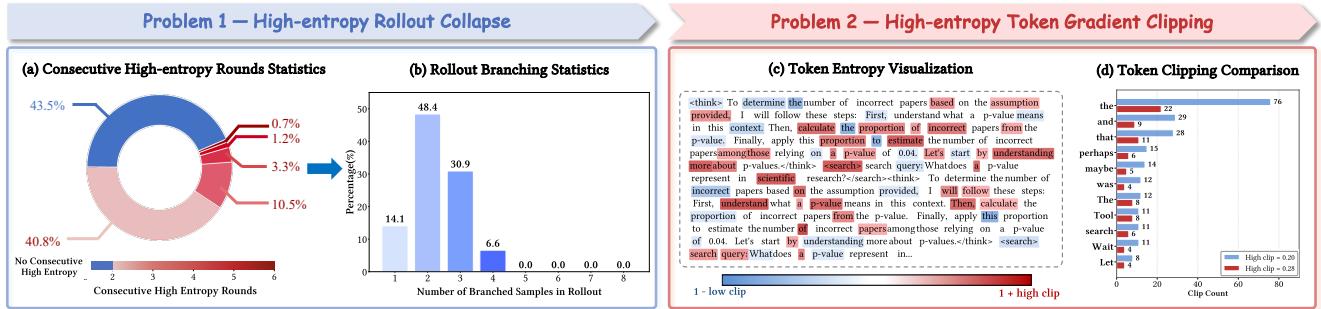


Figure 3: Quantitative statistics of two entropy-based challenges in web agent RL training.

2.1 Problem Definition

2.1.1 Agentic Reinforcement Learning. In this section, we define the training objective for agentic reinforcement learning as follows:

$$\max_{\pi_\theta} \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot | x; T)} [r_\phi(x, y)] - \beta \mathbb{D}_{\text{KL}} [\pi_\theta(y | x; T) \| \pi_{\text{ref}}(y | x; T)], \quad (1)$$

Here, T is the available tool set, π_θ and π_{ref} denote the policy and the reference LLM. The symbols r_ϕ represent the reward function. The input x is drawn from the dataset \mathcal{D} , and y is the corresponding output containing tool-call results.

2.1.2 Token Entropy Calculation. Building on recent studies in entropy-based RL efforts [9, 101, 103, 125], we determine the entropy of token generation at step t using the formula:

$$H_t = - \sum_{j=1}^V p_{t,j} \log p_{t,j}, \quad \text{where } p_t = \pi_\theta(\cdot | \mathcal{R}_{<t}, x; T) = \text{Softmax}\left(\frac{\mathbf{z}_t}{\tau}\right). \quad (2)$$

In this context, V represents the size of the vocabulary, $\mathbf{z}_t \in \mathbb{R}^V$ are the logits before applying the softmax function, and τ is the temperature parameter for decoding. This entropy quantifies the level of uncertainty in the distribution of token generation.

2.2 Entropy-based Pilot Experiments

In this section, we delve into two high-entropy challenges in web agent training and quantify their limitations.

2.2.1 Problem-1: High-Entropy Rollout Collapse. We select ARPO [14] as the backbone algorithm, a representative entropy-guided agentic RL method, and train with its default 1k training samples. We further quantify: (i) the steps in each sampled trajectory that exhibit consecutive high-entropy tool usage; (ii) within each rollout batch (branching budget is 8), the number and probability of trajectories that contain high-entropy branches.

As shown in Figure 3 (left), our key findings are: (i) **High-entropy tool-call turns exhibit transitivity**: the proportion of consecutive high-entropy tool-call turns (56.5%) exceeds isolated high-entropy turns (43.5%), with trajectories reaching up to 6 consecutive high-entropy turns. This indicates that high-entropy tool-call rounds often occur consecutively. (ii) **Rollout branch collapse**: 93.4% of branches concentrate on 1–3 trajectories, while the remaining trajectories receive virtually no budget for high-entropy branch sampling. This shows an imbalanced allocation of rollout branching resources.

We argue two observations are tightly coupled: **Due to an excessive number of consecutive high-entropy rounds in specific samples, the model tends to over-branch on a few trajectories during the rollout phase.** We define this issue as the “*High-Entropy Rollout Collapse*”.

2.2.2 Problem-2: High-Entropy Token Gradient Clipping. Under the same setup as previous experiment, we further visualize the policy update phase of ARPO, including (i) the importance sampling ratio of tokens in trajectories;¹ (ii) a comparison of the Top-10 gradient-clipped tokens between ARPO and DAPO during a training step, with clipping thresholds of 0.2 and 0.28.

As illustrated in Figure 3 (right), we identify the following insights: (i) Consistent with findings in single-turn RL efforts [9, 58], tokens related to logical transitions, connections, and reflections typically exhibit high entropy. Beyond this, specific tool-call tokens also show high entropy. These tokens are highly functional and have low contextual dependency, incentivizing the model to explore diverse reasoning paths and tool-use patterns. (ii) Vanilla RL method uniformly clip the gradients of high-entropy tokens without distinguishing whether they include valuable exploratory behaviors. Although DAPO adopt clip-higher strategy [118] to alleviates this by increasing the threshold, the clipping distribution remains similar and the clipped token count is still substantial.

Moreover, we empirically find that significant gradient clipping emerges in the very first policy update, resulting in a lack of gradient support for high-entropy exploratory tokens in early training. This leads to fixed paradigmatic reasoning, hindering the LLM to explore tool-use patterns. We define this issue as the “*High-Entropy Token Gradient Clipping*”.

2.3 Agentic Tool Design

In this paper, we focus on exploring entropy-balanced optimization for web agent RL algorithms. To this end, we align with existing work on web agent RL training [14, 48, 105] and select three of the most representative tools to evaluate the effectiveness of AEPO: (1) **Web Search Engine**: Provides retrieved source text and corresponding URL information from the web in response to user queries. (2) **Web Browser**: Accesses and parses URL information returned by the search engine, then summarizing the content. (3) **Code Executor**: Executes code generated by models, returning the execution results or error messages.

3 Methodology

This section introduces Agentic Entropy-Balanced Policy Optimization (AEPO), an agentic RL algorithm proposed to balance entropy during both the rollout and policy update phases. As shown in Figure 4, AEPO comprises two core components:

- (1) **Dynamic Entropy-Balanced Rollout:** To mitigate the “*High-Entropy Rollout Collapse*” identified in pilot experiments (§2.2), we adaptively allocate the sampling budget between global and branch sampling via

¹The token-level importance sampling ratio correlate positively with entropy in RL

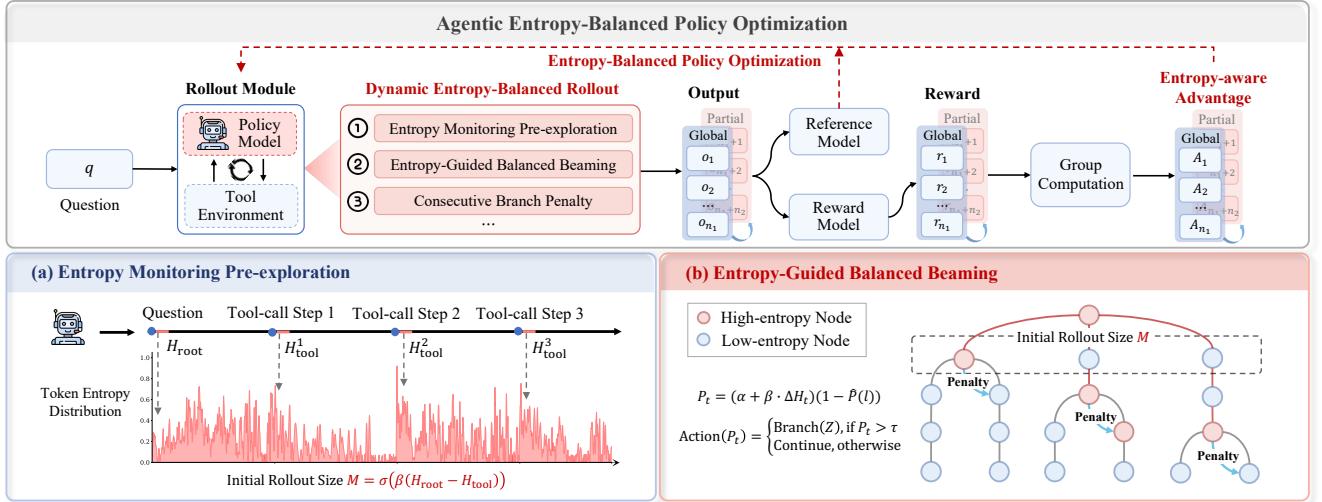


Figure 4: The overview of Agentic Entropy-Balanced Policy Optimization.

- entropy pre-monitoring (§3.1.1), and penalize consecutive high-entropy tool-call steps during rollout to avoid over-branching (§3.1.2).
- (2) **Entropy-Balanced Policy Optimization:** To further address “High-Entropy Token Gradient Clipping”, we insert a stop-gradient operation into the clipping term to preserve and properly rescale gradients on high-entropy tokens (§3.2.1), while incorporating entropy-aware advantage estimation to prioritize learning on high-uncertainty tokens (§3.2.2).

Below, we will delve into the specifics of our approach.

3.1 Dynamic Entropy-Balanced Rollout

In this section, we address the “High-Entropy Rollout Collapse” by naturally breaking it down into two sub-goals: (1) Providing more reasonable resource allocation for global and branch sampling; (2) Penalizing continuous high-entropy branch sampling within single trajectories. Consequently, we propose the following two algorithmic solutions.

3.1.1 Entropy Pre-Monitoring. Traditional tree-based rollout empirically allocate resources for global and branch exploration without theoretical support [14, 38, 55, 114]. Inspired by information bottleneck theory [96], we advocate the allocation of global and partial branching exploration resources from the perspective of maximizing information gain. Specifically, given a total rollout sampling budget of k , which includes m global samples and $k - m$ high-entropy partial branch samples, we simply model the sampling information gain I_{Gain} per rollout step as:

$$I_{\text{Gain}} = \underbrace{m \cdot I_{\text{root}}}_{\text{Global}} + \underbrace{(k - m) \cdot I_{\text{tool}}}_{\text{Partial}}. \quad (3)$$

Here, I_{root} and I_{tool} represent the information gain from the input question and external tool-call result. In the autoregressive decoding process of a language model, the information gain of the question is typically measured by the token entropy decoded by the model, with informative questions generally leading to greater uncertainty [8, 130]. Therefore, we derive the following positive correlation:

$$I_{\text{Gain}} \propto \underbrace{m \cdot H_{\text{root}}}_{\text{Global}} + \underbrace{(k - m) \cdot H_{\text{tool}}^{\text{avg}}}_{\text{Partial}}, \quad \text{where } H_{\text{tool}}^{\text{avg}} = \frac{1}{N} \sum_{i=1}^N H_{\text{tool}}^i, \quad (4)$$

where H_{root} and H_{tool}^i represent the entropy of the question and the entropy introduced by the i -th tool call, respectively. Based on the formula, we

reveal that: (1) When $H_{\text{root}} - H_{\text{tool}}^{\text{avg}} > 0$, the uncertainty from the initial question surpasses that from the subsequent tools. In this case, we should increase m to enhance global exploration, thereby boosting the information gain I_{Gain} . (2) Conversely, when $H_{\text{root}} - H_{\text{tool}}^{\text{avg}} < 0$, m should be decreased to allocate more budget to branch exploration via tool calls.

Based on the above theoretical analysis, we propose the entropy pre-monitoring phase. As shown in Figure 4(a), we first allow the LLM to generate a complete tool-integrated trajectory for the input q . Following ARPO’s entropy calculation [14], we compute the question and tool average entropies for each trajectory according to Equation (4), forming the entropy matrices H_{root} and $H_{\text{tool}}^{\text{avg}} \in \mathbb{R}^{1 \times k}$. Subsequently, by comparing the H_{root} and H_{tool} , we dynamically determine the global sampling count m as:

$$m = k \cdot \sigma \left(\beta \cdot (H_{\text{root}} - H_{\text{tool}}^{\text{avg}}) \right), \quad (5)$$

where $\sigma(x)$ is the sigmoid function, and β controls sensitivity. The value of m is positively correlated with the entropy gap between the question and the tools. As a result, AEPO dynamically allocates rollout sampling resources, thereby enabling efficient sampling.

3.1.2 Entropy-Balanced Adaptive Rollout. After entropy pre-monitoring, we introduce the main entropy-balanced adaptive rollout phase to penalize consecutive high-entropy branch sampling, which comprises three core steps:

(1) **Entropy Variation Monitoring:** Following resource allocation from the pre-monitoring phase, the LLM generates m global trajectory-level samples for the query q , recording the initial entropy matrix H_{root} for each trajectory. After each tool-call step t , the real-time entropy of the model’s output is continuously monitored and represented as a step-level entropy matrix $H_t \in \mathbb{R}^{1 \times k}$. The standardized entropy variation relative to the initial entropy is then computed as $\Delta H_t = \text{Normalize}(H_t - H_{\text{root}})$, where the normalization involves dividing the sum of all values in ΔH by the vocabulary size V .

(2) **Entropy-Balanced Beaming:** Unlike traditional entropy-guided branch sampling [14, 126], AEPO promotes adaptive exploration that showcases beneficial entropy changes in tool-call steps while also constraining consecutive high-entropy branch sampling in specific chains. As shown in Figure 4(b), we introduce a consecutive branch penalty strategy. Given a tool-call step t , the number of consecutive high-entropy branches l prior to

step t for each chain is tracked, then defining the branch sampling probability at step t as follows:

$$P_t = (\alpha + \gamma \cdot \Delta H_t)(1 - \hat{P}(l)), \quad (6)$$

where α is the base sampling probability and γ is the entropy stabilization factor. $\hat{P}(l)$ is a linear function related to l . P_t decreases as the number of consecutive branching steps l increases, implementing a consecutive branching penalty. **This design makes the tree-structured rollout sampling more diverse, allowing for a more comprehensive coverage of the problem-solving space.** We then define the rollout branching action at step t as:

$$\text{Action}(P_t) = \begin{cases} \text{Branch}(Z), & \text{if } P_t > \tau, \\ \text{Continue}, & \text{otherwise.} \end{cases} \quad (7)$$

When P_t exceeds the predetermined threshold τ , $\text{Branch}(Z)$ is initiated, creating Z partial branching reasoning paths from the current node; otherwise, the current trajectory continues.

(3) Termination Conditions: Finally, our iterative rollout process terminates when one of the following conditions is met: (a) If the total number of branch paths Z^* reaches the partial sampling budget $k - m$, branching stops and sampling continues until the final answer is generated; (b) If all paths terminate before reaching $k - m$, then $k - m - Z^*$ additional trajectory-level samples are added to satisfy condition (a).

Through the dynamic entropy-balanced rollout, AEPO ensures the diversity of sampling branches while adaptively allocating exploration resources, thus addressing the “High-entropy Rollout Collapse” issue. The algorithm of dynamic entropy-balanced rollout is detailed in Algorithm 1.

3.2 Entropy-Balanced Policy Optimization

AEPO preserves a considerable number of exploratory tokens via entropy-balanced rollouts, presenting a challenge in effectively updating these tokens’ gradients. This section aims to improve targeted learning for these tokens by implementing the following designs:

3.2.1 Entropy Clipping-Balanced Mechanism. Unlike traditional methods that entirely discard gradients outside the clipping range [14, 76], AEPO introduces an innovative high-entropy clipping-balanced mechanism. The core idea is to retain high-entropy gradients that exceed the clipping interval, allowing the model to learn valuable exploratory token signals.

Motivated by GPPO [84], we integrate a *stop-gradient* operation into the high-entropy clipping term of the policy update phrase, decoupling forward and backward propagation. Our mechanism ensures that forward propagation remains unchanged, while protecting the gradient backward of high-entropy tokens from clipping constraints. For instance, in GRPO [76], given an input question x and a policy model y , GRPO enables the reference policy π_{ref} to generate a group of G outputs $\{y_1, y_2, \dots, y_G\}$ and optimizes the policy by maximizing:

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \min \left(\delta \tilde{A}^{(t)}, \text{clip} \left(\delta, 1 - \epsilon_l, \frac{1 + \epsilon_h}{\text{sg}(\delta)} \delta \right) \tilde{A}^{(t)} \right) \right], \quad (8)$$

where $\delta = r_t^{(j)}(\theta)$ represents the importance sampling ratio, and $\text{sg}(\cdot)$ denotes the *stop-gradient* operation. It is noteworthy that **the value of the term $\delta \cdot \text{sg}(\delta)$ always equals 1, ensuring that AEPO’s forward computation remains unchanged.** For the backpropagation, AEPO’s gradient update process is formulated as:

$$\nabla_{\theta} \mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \mathcal{F}_{j,t}(\theta) \cdot \phi_{\theta}(a_{j,t}, s_{j,t}) \cdot \tilde{A}^{(t)} \right], \quad (9)$$

$$\text{where } \mathcal{F}_{j,t}(\theta) = \begin{cases} 1 + \epsilon_h, & \text{if } \delta > 1 + \epsilon_h \text{ and } \tilde{A}^{(t)} > 0, \\ 0, & \text{if } \delta < 1 - \epsilon_h \text{ and } \tilde{A}^{(t)} < 0, \\ \delta, & \text{otherwise.} \end{cases}$$

Algorithm 1 Dynamic Entropy-Balanced Rollout

```

Require: Reasoning model  $\pi_{\theta}$ , external tools  $T$ , total rollout size  $k$ , entropy sensitivity  $\beta$ , branch penalty slope  $\gamma$ 
1: Input: Dataset  $D$ 
2: Initialize reference model:  $\pi_{\theta}^{\text{old}} \leftarrow \pi_{\theta}$ 
3: for  $i = 1$  to  $C$  do
4:   Sample mini-batch  $D_b \subset D$ 
5:   for each query  $q \in D_b$  do
6:     // Entropy Pre-Monitoring
7:     Generate 1 complete trajectory  $r$  to obtain  $H_{\text{root}}$  and  $H_{\text{tool}}^{\text{avg}}$ 
8:     Global rollout size  $m \leftarrow k \cdot \sigma(\beta(H_{\text{root}} - H_{\text{tool}}^{\text{avg}}))$ 
9:     Branch rollout size  $b \leftarrow k - m$ 
10:    // Entropy-Balanced Adaptive Rollout
11:    Initialize rollout pool  $\mathcal{P} \leftarrow \emptyset$ 
12:    Consecutive-high-entropy counter  $l \leftarrow 0$ 
13:    while  $|\mathcal{P}| < m$  do
14:      Sample trajectory  $r$ ; add to  $\mathcal{P}$ 
15:    end while
16:    while  $b > 0$  and  $\exists r_j \in \mathcal{P}$  not terminated do
17:      // (1) Entropy Variation Monitoring
18:      Select a trajectory  $r \in \mathcal{P}$  at tool-call step  $t$ 
19:       $\Delta H_t \leftarrow \text{Normalize}(H_t - H_{\text{initial}})$ 
20:      // (2) Entropy-Balanced Beaming
21:      Consecutive penalty  $\hat{P}(l) \leftarrow \gamma \cdot l$ 
22:      Branch probability  $P_t \leftarrow (\alpha + \beta \Delta H_t)(1 - \hat{P}(l))$ 
23:      // (3) Termination Conditions
24:      if  $P_t > \tau$  then
25:        Branch  $Z$  sub-trajectories;  $b \leftarrow b - Z$ 
26:      else
27:         $l \leftarrow l + 1$  if  $\Delta H_t > 0$ 
28:      end if
29:    end while
30:    if  $b > 0$  then
31:      Sample  $b$  additional trajectories and add to  $\mathcal{P}$ 
32:    end if
33:  end for
34: end for
35: Output: rollout trajectory set  $\mathcal{P}$ 

```

During backpropagation, the gradient of a high-entropy token is retained and appropriately rescaled to $1 + \epsilon_h$ only when $\delta > 1 + \epsilon_h$ and $\tilde{A}^{(t)} > 0$. In other cases, the gradient update rule aligns with vanilla clipping mechanism of GRPO. This controlled rescaling ensures that the model learns a balanced exploratory behavior without completely ignoring high-entropy tokens. To more clearly articulate the theoretical aspects of AEPO compared to clipping-optimized RL methods [3, 85], we discuss their differences in Appendix B².

3.2.2 Entropy-aware Advantage Estimation. Owing to the clipping-balanced mechanism, we retain the gradients of high-entropy tokens. However, a challenge arises in training the model to better distinguish between exploratory and non-exploratory tokens. Traditional outcome-based RL algorithms assign the same advantage to all tokens in a sequence based on the answer correctness, neglecting the model’s confidence levels across different tokens [35, 76].

To this end, we propose an **entropy-aware advantage estimation** that incorporates token entropy calculation into advantage shaping. This approach allows the model to assign greater rewards to exploratory tokens

²For detailed proof of the gradient form of AEPO, please refer to Appendix A

that are correct but exhibit high uncertainty. A natural way is to calculate an accuracy-based advantage while integrating an entropy-based advantage term, defined as follows:

$$\tilde{A}_{\text{Acc}}^{(t)} = \frac{r_t - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}, \quad \tilde{A}_{\Delta H}^{(t)} = \frac{H_t - \text{mean}(\{H_t\}_{t=1}^T)}{\text{std}(\{H_t\}_{t=1}^T)}, \quad (10)$$

where H_t represents the t -th token entropy, and T is the total number of tokens across all trajectories in the group. We estimate the entropy advantage for each token based on the average token entropy within the same trajectory. Furthermore, we treat the entropy advantage as a regularization term in the advantage estimation to reshape A_{acc} as:

$$\tilde{A}^{(t)} = \tilde{A}_{\text{Acc}}^{(t)} * (1 + \alpha \cdot \tilde{A}_{\Delta H}^{(t)}). \quad (11)$$

This step is computed before the policy update. Notably, our entropy-aware advantage estimation can be seamlessly integrated with existing agentic RL algorithms to further enhance the model's emphasis on learning exploratory tokens during training. The full algorithm of AEPO is detailed in Algorithm 2.

4 Experiment Settings

4.1 Datasets

We assess the effectiveness of AEPO in web agents RL training through three long-term reasoning tasks:

- (1) **Deep Information Seeking Tasks:** This includes challenging evaluations for web agents: General AI Assistant (GAIA) [60] and the Human Last Exam (HLE) [66], as well as deep information seeking: WebWalkerQA [107], XBenchmark [5], and Frames [43].
- (2) **Knowledge-Intensive Reasoning:** This covers 3 multi-hop complex open-domain question-answering tasks: 2WikiMultihopQA [33], Musique [97], and Bamboogle [67], along with the web multi-hop task WebWalkerQA [108].
- (3) **Computational Reasoning:** This includes simple math reasoning tasks like GSM8K [11], MATH [32], and competition-level math challenges: MATH500 [56], AIME2024, and AIME2025.³ All dataset splits align with the standard settings established by previous works [14, 41, 51].

4.2 Baselines

We consider the following strong baseline methods:

- (1) **Advanced RL Algorithms:** We select three categories of RL algorithms: (1) Vanilla RL: GRPO [76] and Reinforce++ [35]; (2) Clipping-optimized RL: DAPO [118], CISPO [61] and GPPO [84]; and (3) Agentic RL: GIGPO [26] and ARPO [14].
- (2) **Advanced Backbone Models:** For challenging reasoning benchmarks, we evaluate the instruction-tuned versions of Qwen2.5 [74] and Llama3.1 [18]. For deep information seeking tasks, we also report results for QwQ [95], DeepSeek-R1 [29], GPT-4o [37], and o1-preview [37], using Qwen3-32B [113] as a reference.
- (3) **Advanced Web Agents:** We introduce a series of open-source workflow-based search agents as references, including vanilla RAG [44], Search o1 [50], Webthinker [51], and ReAct [115]. The detailed introduction of baselines are listed in Appendix C

4.3 Evaluation Metric

Consistent with previous work, we use the F1 score to evaluate four question-answering tasks that require intensive knowledge reasoning. For other tasks, we employ the VLLM framework to serve Qwen2.5-72B-instruct, using LLM-as-Judge to assess the answers. In all tasks, answers are extracted from the box in the response. By default, the temperature is set to 0.6 and top-p to 0.95, and we evaluate using the Pass@1 score.

³<https://huggingface.co/datasets/AI-MO/aimo-validation-aime>

4.4 Implementation Details

In the AEPO phase, we implement the AEPO algorithm using the VERL framework [77], excluding tool-call results from the loss calculation to avoid bias. Our setup includes a training batch size of 128, a PPO mini-batch size of 16, and a context length of 20K. For AEPO rollout, the global rollout size is 16, with a, β set to 0.2. Resource allocation follows Equation 5, with a consecutive branch penalty probability of $P(l) = 0.2 \cdot l$. Other settings align with ARPO for fair comparison. To stabilize RL training, the KL divergence coefficient in GRPO is set to 0. The RL for reasoning and deep information seeking is 2 and 5 epochs. All experiments use 16 NVIDIA H800 GPUs.

During training and evaluation, we use the Bing Search API (US-EN region) as the search engine. Following RAG-related work [41, 51], we retrieve 10 web pages per query. For reasoning tasks, we use the top 10 snippets; for deep information seeking, we extract up to 6000 tokens per page and use a same size model as a browser agent.

5 Experiment Results

5.1 Main Result on Deep Information Seeking

To validate the effectiveness of AEPO in challenging deep web information seeking tasks, we train the Qwen3 series models combined with AEPO using 1K open-source samples and compared them with advanced web agents and RL algorithms. As shown in Table 1, we derived the following insights:

(1) **Limitations of Advanced Large Models:** Both advanced closed-source LLMs and large-parameter open-source LLMs (e.g. GPT-4o and DeepSeek-R1-671B) perform poorly in challenging deep information seeking scenarios, particularly on the GAIA (<30%) and HLE (<10%) benchmarks. This indicates that relying solely on model internal knowledge is insufficient for complex agentic search tasks.

(2) **Strong Generalization Ability of AEPO in Deep Information Seeking:** Compared to robust web agents and advanced RL algorithms, the Qwen3-8B and 14B models combined with AEPO demonstrate exceptional performance, achieving pass@1 scores of 11.2%, 47.6% and 43% on the HLE, GAIA and WebWalkerQA benchmarks, respectively. Notably, our model is trained solely on 1k samples from an open-source web search dataset, without any data synthesis or filtering, showcasing its efficiency in training web agents.

(3) **Importance of Dual Entropy Balancing Optimization:** AEPO consistently outperforms ARPO in both average performance and individual benchmarks, with Qwen3-8B showing a significant 6% improvement on the GAIA benchmark and WebWalkerQA. This highlights the importance of AEPO's algorithmic design, which implements dual entropy balancing in both the Rollout and policy update phases, effectively facilitating LLMs' exploratory tool behavior and addressing two high-entropy challenges. This is crucial for deep information seeking scenarios involving frequent tool invocation.

5.2 Main Result on Generalized Reasoning

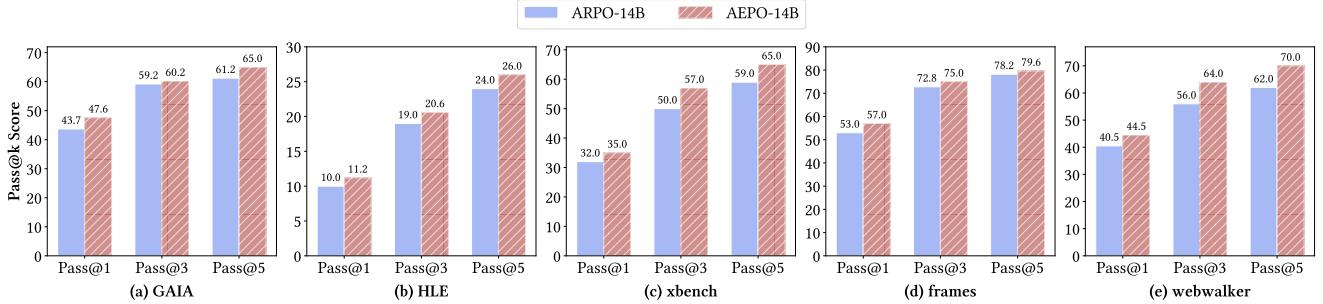
To further validate the effectiveness of AEPO in web agent training, we conduct a comparison of AEPO with 7 RL algorithms across 10 challenging reasoning tasks. As shown in Table 2, our key insights are as follows:

(1) **Instability of Clipping RL Algorithms in Web Agent Training:** Using GRPO as a baseline, clipping-optimized RL algorithms perform well on Qwen 2.5-7B-instruct, with GPPO and CISPO achieving average scores above 57%. However, in Llama3-8B, they do not show significant improvement over GRPO. Furthermore, practical experiments reveal that clipping-optimized RL algorithms often lead to entropy collapse, disrupting training performance. This indicates that clipping-optimized RL algorithms are sensitive to the architecture of the backbone model and often show instability during web agent training.

(2) **Generalization Ability of Agentic RL Algorithms:** Agentic RL algorithms, represented by ARPO and GIGPO, demonstrate stable and robust

Table 1: Overall performance on deep information seeking tasks. The best results are indicated in bold, and the second-best results are underlined. Results from larger or closed-source models are presented in gray for reference.

Method	General AI Assistant				WebWalkerQA				Humanity's Last Exam				XBench-DR	FRAMES
	Lv.1	Lv.2	Lv.3	Avg.	Easy	Med.	Hard	Avg.	NS	CE	SF	Avg.	Avg.	Avg.
Direct Reasoning (>=32B)														
Qwen3-32B-thinking	26.2	12.1	0	14.9	6.9	1.1	2.9	3.1	14.6	9.8	8.4	12.6	14.0	26.0
DeepSeek-R1-32B	21.5	13.6	0.0	14.2	7.5	1.4	4.2	3.8	6.6	5.1	6.5	6.4	10.0	23.8
QwQ-32B	30.9	6.5	5.2	18.9	7.5	2.1	4.6	4.3	11.5	7.3	5.2	9.6	10.7	28.8
GPT-4o	23.1	15.4	8.3	17.5	6.7	6.0	4.2	5.5	2.7	1.2	3.2	2.6	18.0	44.6
DeepSeek-R1-671B	40.5	21.2	5.2	25.2	5.0	11.8	11.3	10.0	8.5	8.1	9.3	8.6	32.7	45.6
o1-preview [†]	-	-	-	-	11.9	10.4	7.9	9.9	12.9	8.1	6.6	11.1	-	-
Single-Enhanced Method (Qwen3-8B)														
Vanilla RAG	28.2	15.4	16.7	20.4	8.9	10.7	9.9	10.0	5.1	1.6	<u>12.9</u>	5.8	8.0	18.8
Search-o1	35.9	15.4	0.0	21.4	6.7	15.5	9.7	11.5	7.6	2.7	5.3	6.4	10.0	19.2
WebThinker	43.6	11.5	0.0	22.3	6.7	13.1	16.9	13.0	7.3	4.0	6.3	6.6	13.0	21.4
ReAct	35.9	17.3	<u>8.3</u>	23.3	8.9	16.7	18.3	15.5	4.2	4.0	6.3	4.6	16.0	21.1
RL-based Method (Qwen3-8B)														
Qwen3-8B	28.1	15.4	16.7	20.4	0.0	2.4	2.8	2.0	3.9	2.7	8.4	4.6	9.0	19.0
+ GRPO	48.7	25.0	<u>8.3</u>	32.0	28.9	32.1	<u>28.2</u>	30.0	<u>7.9</u>	4.0	10.5	7.8	20.0	46.2
+ ARPO	<u>53.9</u>	<u>32.7</u>	16.7	<u>38.8</u>	<u>31.1</u>	<u>35.7</u>	<u>28.2</u>	<u>32.0</u>	7.3	6.7	15.8	<u>8.8</u>	<u>25.0</u>	<u>47.8</u>
+ AEPO (Ours)	61.5	42.3	<u>8.3</u>	45.6	40.0	39.3	35.2	38.0	12.1	<u>5.3</u>	11.6	11.0	28.0	50.2
Single-Enhanced Method (Qwen3-14B)														
Vanilla RAG	38.5	19.2	<u>8.3</u>	25.2	17.8	13.1	11.3	13.5	5.5	6.3	9.4	6.0	15.0	31.4
Search-o1	48.7	23.1	0.0	30.1	11.1	21.4	16.9	17.5	6.4	4.0	10.5	6.8	21.0	39.8
WebThinker	48.7	26.9	<u>8.3</u>	33.0	13.3	23.8	18.3	19.5	7.0	4.0	9.5	7.0	23.0	40.8
ReAct	48.7	25.0	<u>8.3</u>	32.0	11.1	20.2	12.7	15.5	5.8	5.3	10.5	6.6	20.0	37.6
RL-based Method (Qwen3-14B)														
Qwen3-14B	33.3	13.5	0.0	19.4	6.7	2.4	4.2	4.0	5.5	6.7	11.6	6.8	14.0	23.8
+ GRPO	51.3	34.6	0.0	36.9	<u>35.6</u>	42.9	35.2	38.5	7.9	6.7	<u>12.6</u>	8.6	<u>27.0</u>	54.6
+ ARPO	<u>56.4</u>	<u>40.4</u>	16.7	<u>43.7</u>	40.0	44.1	<u>36.6</u>	<u>40.5</u>	<u>10.3</u>	<u>10.7</u>	<u>13.7</u>	<u>10.0</u>	32.0	<u>55.4</u>
+ AEPO (Ours)	61.5	44.2	16.7	47.6	40.0	50.0	40.9	44.5	10.6	14.7	10.5	11.2	35.0	58.8

**Figure 5: The comparison analysis of Qwen3-14B using ARPO and AEPO across Pass@1 to Pass@5 metrics.**

performance across both backbone models, with ARPO achieving average performance consistency above 55%. Notably, these methods attempt tree-structured rollout during the rollout phase, further confirming the effectiveness of branching exploration in high-entropy tool-call steps.

(3) **Effectiveness of AEPO:** AEPO consistently outperforms other reinforcement learning algorithms across 10 datasets and backbone models, achieving an average accuracy improvement of nearly 5% over GRPO while maintaining competitiveness across fine-grained domains. These results highlight AEPO’s efficiency and strong adaptability across different model architectures and tasks, making it more suitable than other RL algorithms for training multi-turn web agents.

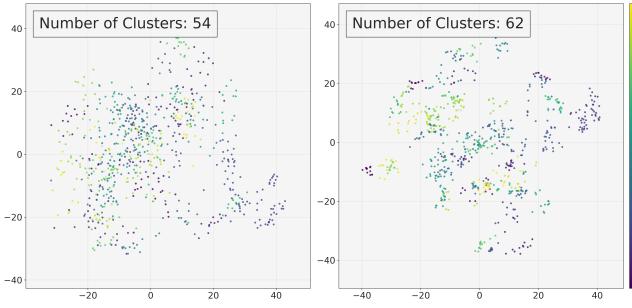
5.3 Pass@K Sampling Analysis

Due to the dynamic multi-turn interactions and complexity of tool environments in web agent training, we conduct a sampling analysis of the model’s Pass@3 and Pass@5 to accurately assess its true problem-solving abilities.

As illustrated in Figure 5, AEPO demonstrates significant performance improvements with larger-scale sampling. Notably, the Qwen3-14B model combined with AEPO achieves remarkable results: GAIA at 65%, HLE at 26%, and XBench-DR at 65%. Compared to the robust agentic RL algorithm ARPO, AEPO consistently excels across five datasets. This stable improvement in Pass@K can be primarily attributed to AEPO’s entropy balancing optimizations, which allows the model to explore fine-grained tool usage behaviors more efficiently, thereby enhancing reasoning and sampling efficiency.

Table 2: Overall performance on ten challenging reasoning tasks are presented. The top two outcomes are bolded and underlined.

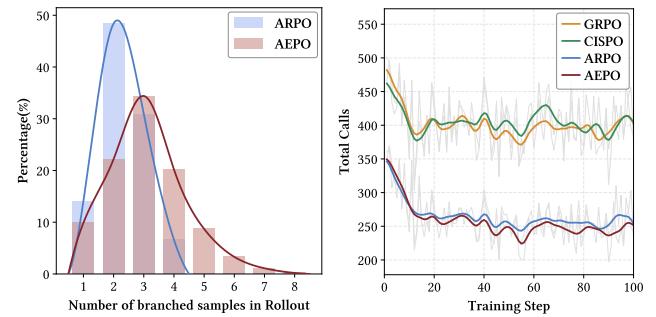
Method	Mathematical Reasoning					Knowledge-Intensive Reasoning					Avg.	
	AIME24	AIME25	MATH500	GSM8K	MATH	WebWalker	HotpotQA	2Wiki.	MuSiQue.	Bamboogle		
<i>Backbone Model: Llama3.1-8B-Instruct</i>												
Classical RL Method												
+ GRPO	13.3	<u>13.3</u>	62.4	87.4	79.2	26.5	57.8	71.8	31.0	68.2	51.1	
+ Reinforce ++	13.3	16.7	61.4	87.0	77.2	27.5	57.1	71.6	29.9	69.1	51.1	
Clipping-optimized RL Method												
+ DAPO	16.7	<u>13.3</u>	61.2	87.4	76.4	25.5	56.6	70.3	29.2	67.3	50.4	
+ GPPPO	16.7	6.7	61.8	86.6	79.4	27.5	61.8	72.8	29.8	71.9	51.5	
+ CISPO	13.3	6.7	62.2	87.0	78.2	26.0	57.3	<u>75.6</u>	32.2	71.8	51.0	
Agentic RL Method												
+ GIGPO	20.0	13.3	62.4	87.4	77.2	<u>31.5</u>	61.8	74.6	31.8	72.1	53.2	
+ ARPO	<u>23.3</u>	16.7	<u>64.6</u>	88.0	<u>80.2</u>	30.5	65.4	75.5	34.8	<u>73.8</u>	<u>55.3</u>	
+ AEPO (Ours)	26.7	16.7	65.8	<u>87.6</u>	80.6	33.5	<u>64.7</u>	79.0	<u>33.0</u>	75.8	56.3	
<i>Backbone Model: Qwen2.5-7B-Instruct</i>												
Classical RL Method												
+ GRPO	23.3	<u>26.7</u>	78.0	92.8	87.8	22.0	59.0	<u>76.1</u>	30.6	68.4	56.5	
+ Reinforce ++	26.7	23.3	78.0	92.2	<u>88.8</u>	26.0	55.1	68.9	25.2	64.9	54.9	
Clipping-optimized RL Method												
+ DAPO	20.0	23.3	80.4	91.0	<u>88.8</u>	24.0	57.7	68.4	28.6	65.5	54.8	
+ GPPPO	26.7	23.3	76.2	91.6	87.6	<u>31.0</u>	60.7	74.2	31.5	<u>72.4</u>	57.5	
+ CISPO	26.7	30.0	77.8	91.4	86.2	29.0	59.3	72.1	29.1	70.2	57.2	
Agentic RL Method												
+ GIGPO	<u>30.0</u>	20.0	78.4	91.6	87.6	30.5	58.1	73.5	<u>31.1</u>	70.1	57.1	
+ ARPO	<u>30.0</u>	30.0	<u>78.8</u>	<u>92.2</u>	<u>88.8</u>	26.0	58.8	<u>76.1</u>	<u>31.1</u>	71.5	<u>58.3</u>	
+ AEPO (Ours)	33.3	30.0	80.4	<u>92.2</u>	90.0	31.5	62.5	77.1	<u>31.1</u>	73.4	60.1	

**Figure 6: Visualization of Rollout diversity: ARPO (left) and AEPO (right)**

5.4 Does AEPO Mitigate Rollout Collapse?

(1) **Diversity Analysis.** To investigate whether AEPO’s dynamic entropy balanced rollout improves sampling diversity, we follow the setup of the preliminary experiment (§2.2) and randomly selected samples from 10 rollout steps, encompassing 640 distinct problems and approximately 7.6k trajectories. We further employ BGEM3 [4] as the semantic embedding model, applied the PCA method for dimensionality reduction, and used DBSCAN [19] for clustering to visualize the representation of rollout sampling.

As shown in Figure 6, the results indicate that compared to ARPO, AEPO’s sampling trajectories form more distinct cluster centers (54 vs. 62) and exhibit tighter intra-cluster distances with larger inter-cluster gaps. This demonstrates that AEPO improves the scope of rollout diversity and provides clearer differentiation in the sampling path distribution. We attribute this to AEPO’s entropy pre-monitoring and continuous entropy

**Figure 7: The comparison of branch sampling distribution in rollout (left); The comparison of tool-call efficiency across four RL algorithms (right).**

penalty branches, which effectively address the continuity of high-entropy branches to achieve comprehensive coverage of the problem-solving space.

(2) **Statistics Analysis.** To quantitatively analyze AEPO’s effectiveness in addressing rollout collapse, we measure the branch distribution of ARPO and AEPO over 10 steps during rollout. As shown in Figure 7 (left), with both the global and partial branch sampling budgets set to 8, ARPO typically branches into 2-3 trajectories. In contrast, AEPO exhibits a more diverse branching pattern, potentially covering all 8 paths with different branches. This highlights AEPO’s dynamic resource allocation and continuous branch penalty mechanism enable the model to explore potential high-entropy tool-call steps across different trajectories, effectively mitigating bias in specific path branches.

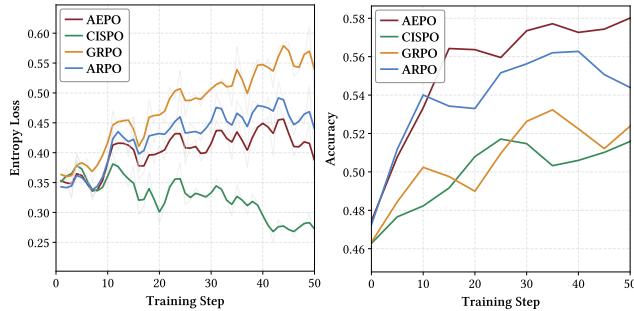


Figure 8: Visualization of training dynamics, including entropy loss(left) and accuracy (right) across training steps

5.5 Does AEPO Achieve Entropy-Balanced and Efficient RL Training?

(1) *Tool-call Efficiency Analysis.* In agentic RL training, effectively controlling the frequency of tool usage can significantly reduce financial costs. To confirm the efficiency of AEPO’s tool usage, we quantify the tool consumption of AEPO compared to other RL algorithms in the deep information seeking task. As shown in Figure 7 (right), AEPO requires only about half the number of tool calls to achieve superior performance compared to vanilla and clipping-optimized RL algorithms. Additionally, compared to the agentic RL algorithm ARPO, AEPO consistently reduces the number of tool calls. We attribute this enhanced efficiency to the entropy pre-monitoring phase, which balances the allocation of rollout exploration resources based on the information gain from the problem and tool usage. This ensures that AEPO not only broadens the rollout exploration space but also achieves efficient web agent training.

(2) *Entropy Stability Analysis.* To better quantify the effectiveness of entropy-balanced policy optimization during policy updates, we present the RL training curves for 10 reasoning tasks. Figure 8 illustrates the dynamic visualization of entropy loss and validation set accuracy across 10 reasoning tasks throughout the training steps. We observe that using clipping-optimized RL often encounters entropy instability during training, leading to performance collapse. In contrast, AEPO demonstrates a more stable entropy curve compared to other reinforcement learning algorithms. Interestingly, sharp fluctuations in entropy loss do not improve training stability and effectiveness. Instead, maintaining a consistently high and stable entropy dynamic is generally advantageous for ongoing performance enhancement. This observation supports our initial motivation, as AEPO employs entropy-balanced policy optimizations to foster more reasonable and stable entropy dynamics.

6 Related Work

6.1 Reinforcement Learning for Web Agent.

The emergence of agent reinforcement learning (RL)[122] has set the stage for the development of general-purpose web agents, a pursuit shared by both academia and industry. Initial efforts[6, 24, 40, 53, 81] established a foundation by enabling models to autonomously interact with search engines or code interpreters using rule-based RL. Building on this groundwork, subsequent innovations have emerged: Tool-star [13] incorporates multi-tool usage within agentic RL, while other studies [7, 12, 36, 82, 88, 98, 99, 110, 128] enhance efficiency and stability through redesigned reward functions. MemAgent introduces memory mechanisms during the RL phase to better manage contextual information [117]. Additionally, recent research [27, 39, 49, 112] explores comprehensive asynchronous training frameworks for web agents. Building on these advancements, Tongyi Deep

Research [21, 47, 72, 83, 90, 106, 109, 121] aims to fully leverage the post-training paradigm. This includes data synthesis, RL algorithm optimization, and report generation, thereby broadening the scope of web agent training. To minimize resource consumption during training, another line of research seeks to simulate search engines using the generative capabilities of large models for self-alignment [20, 86].

Recently, agentic RL methods [14, 26, 34, 55] have focused on optimizing foundational RL algorithms for web agents, employing tree-structured rollouts for autonomous branch sampling under high entropy. While these methods have advanced web agent training, they often overlook the challenges posed by high-entropy tokens. Several single-turn RL studies [9, 58, 85, 126] have emphasized that stable entropy training is crucial for enhancing model performance. However, this aspect remains largely unexplored in multi-turn agentic RL. In this paper, we introduce AEPO to achieve entropy-balanced web agent RL training.

6.2 Agentic Reinforcement Learning.

Reinforcement learning (RL) plays a crucial role in helping large language model (LLM) agents adapt to dynamic and open environments [59, 60, 78]. Foundational studies such as DQN [62] and AlphaZero [79] have shown that self-play-based RL can endow agents with skills ranging from natural language understanding to strategic gameplay [63]. Building on these foundations, value-based RL methods have been applied to improve embodied intelligence in hardware control and complex gaming tasks [1, 65, 75, 89, 102, 120]. Recent advancements, like RAGEN [104, 129], incorporate reasoning states and environmental interactions into turn-level responses using trajectory-level RL. To enhance tool-integrated reasoning, several studies [6, 24, 24, 40, 40, 51, 53, 80, 81, 86] utilize rule-based RL to enable LLMs to autonomously invoke external tools (e.g., search engines, Python compilers) to improve reasoning accuracy. Further research, including ToolRL [68], Tool-Star [13], and OTC [99], explores the integration of multiple tools and enhances tool-use efficiency. Efforts by Kimi Deepresearcher⁴ and Web-sailor [48] focus on optimizing RL algorithms to better handle deepsearch’s long context scenarios. With the surge in reasoning capabilities of Multimodal large language models (MLLMs), several works have effectively broadened the scope of this field by combining agentic RL in the multimodal domain with external tools [2, 69–71, 73, 100, 111].

Although many studies enhance tool invocation through reward shaping and rollout mechanisms, trajectory-level RL alone often struggles to effectively capture the multi-turn, long-horizon characteristics of LLM-based agent behavior. This challenge has led to the development of ARPO, which aims to learn step-level tool-use behavior patterns.

7 Conclusion

In this paper, we introduce Agentic Entropy-Balanced Policy Optimization (AEPO), an agentic RL algorithm that effectively balances entropy during both rollout and policy update phases. Initially, we quantify two inherent entropy-driven challenges in preliminary experiments. AEPO comprises two core components: (1) a dynamic entropy-balanced rollout mechanism that adaptively allocates the sampling budget between global and branch sampling through entropy pre-monitoring, while imposing a branch penalty on consecutive high-entropy tool-call steps to prevent oversampling; (2) Entropy-Balanced Policy Optimization, which incorporates a stop-gradient operation in the high-entropy clipping term to preserve and rescale gradients on high-entropy tokens, alongside entropy-aware advantage estimation to focus learning on high-uncertainty tokens. Experiments across 14 benchmarks demonstrate that AEPO consistently outperforms seven mainstream agentic RL algorithms. Quantitative analyses confirm its scalability and stability, offering valuable insights for training general web agents.

⁴<https://moonshotai.github.io/Kimi-Researcher/>

References

- [1] Hao Bai, Yifei Zhou, Jiayi Pan, Mert Cemri, Alane Suhr, Sergey Levine, and Aviral Kumar. 2024. DigiRL: Training In-The-Wild Device-Control Agents with Autonomous Reinforcement Learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amin Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/1704ddd0bb89f159dfe609b32c889995-Abstract-Conference.html
- [2] Shuai Bai, Keping Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923* (2025).
- [3] Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, Chengjun Xiao, Chengyu Du, Chi Zhang, Chu Qiao, Chunhai Zhang, Chunhui Du, Congchao Guo, Da Chen, Deming Ding, Dianjun Sun, Dong Li, Enwei Jiao, Haigang Zhou, Haimo Zhang, Han Ding, Haohai Sun, Haoyu Feng, Huaiqiang Cai, Haichao Zhu, Jian Sun, Jiaqi Zhuang, Jieren Cai, Jiayuan Song, Jin Zhu, Jingyang Li, Jinhao Tian, Jinli Liu, Junhao Xu, Junjie Yan, Junteng Liu, Junxian He, Kaiyi Feng, Ke Yang, Kecheng Xiao, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Li, Lin Zheng, Lingde Du, Lingyu Yang, Lunbin Zeng, Minghui Yu, Mingliang Tao, Mingyuan Chi, Mozhi Zhang, Mujeje Lin, Nan Hu, Nongyu Di, Peng Gao, Pengfei Li, Pengyu Zhao, Qibing Ren, Qidi Xu, Qile Li, Qin Wang, Rong Tian, Ruitao Leng, Shaoxiang Chen, Shaoyu Chen, Shengmin Shi, Shitong Weng, Shuchang Guan, Shuqi Yu, Sichen Li, Songguan Zhu, Tengfei Li, Tianchi Cai, Tianrun Liang, Weiyu Cheng, Weize Kong, Wenkai Li, Xiancai Chen, Xiangjun Song, Xiaoxiao Luo, Xiao Su, Xiaobiao Li, Xiaodong Han, Xinzhu Hou, Xuan Lu, Xun Zou, Xuyang Shen, Yan Gong, Yan Ma, Yang Wang, Yiqi Shi, Yiran Zhong, and Yonghong Duan. 2025. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. *CoRR abs/2506.13585* (2025). <https://doi.org/10.48550/ARXIV.2506.13585>
- [4] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *CoRR abs/2402.03216* (2024). <https://doi.org/10.48550/ARXIV.2402.03216>
- [5] Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, et al. 2025. xbench: Tracking Agents Productivity Scaling with Profession-Aligned Real-World Evaluations. *arXiv preprint arXiv:2506.13651* (2025).
- [6] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan, Wen Zhang, Huajun Chen, Fan Yang, Zennan Zhou, and Weipeng Chen. 2025. ReSearch: Learning to Reason with Search for LLMs via Reinforcement Learning. *arXiv:cs.AI/2503.19470* <https://arxiv.org/abs/2503.19470>
- [7] Yifei Chen, Guanting Dong, and Zhicheng Dou. 2025. Toward Effective Tool-Integrated Reasoning via Self-Evolved Preference Learning. *arXiv preprint arXiv:2509.23285* (2025).
- [8] Yifei Chen, Guanting Dong, Yutao Zhu, and Zhicheng Dou. 2025. Revisiting RAG Ensemble: A Theoretical and Mechanistic Analysis of Multi-RAG System Collaboration. *arXiv preprint arXiv:2508.13828* (2025).
- [9] Daixuan Cheng, Shaohan Huang, Xuekai Zhu, Bo Dai, Wayne Xin Zhao, Zhenliang Zhang, and Furu Wei. 2025. Reasoning with Exploration: An Entropy Perspective. *CoRR abs/2506.14758* (2025). <https://doi.org/10.48550/ARXIV.2506.14758>
- [10] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. 2025. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. *CoRR abs/2501.17161* (2025). <https://doi.org/10.48550/ARXIV.2501.17161>
- [11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).
- [12] Yong Deng, Guoqing Wang, Zhenzhe Ying, Xiaofeng Wu, Jinzhen Lin, Wenwen Xiong, Yuqin Dai, Shuo Yang, Zhanwei Zhang, Qiwen Wang, Yang Qin, Yuan Wang, Quanxing Zha, Sunhao Dai, and Changhua Meng. 2025. Atom-Searcher: Enhancing Agentic Deep Research via Fine-Grained Atomic Thought Reward. *CoRR abs/2508.12800* (2025). <https://doi.org/10.48550/ARXIV.2508.12800>
- [13] Guanting Dong, Yifei Chen, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Yutao Zhu, Hangyu Mao, Guorui Zhou, Zhicheng Dou, and Ji-Rong Wen. 2025. ToolStar: Empowering LLM-Brained Multi-Tool Reasoner via Reinforcement Learning. *CoRR abs/2505.16410* (2025). <https://doi.org/10.48550/ARXIV.2505.16410>
- [14] Guanting Dong, Hangyu Mao, Kai Ma, Licheng Bao, Yifei Chen, Zhongyuan Wang, Zhongxia Chen, Jiazhen Du, Huiyang Wang, Fuzheng Zhang, Guorui Zhou, Yutao Zhu, Ji-Rong Wen, and Zhicheng Dou. 2025. Agentic Reinforced Policy Optimization. *CoRR abs/2507.19849* (2025). <https://doi.org/10.48550/ARXIV.2507.19849>
- [15] Guanting Dong, Xiaoshuai Song, Yutao Zhu, Runqi Qiao, Zhicheng Dou, and Ji-Rong Wen. 2025. Toward Verifiable Instruction-Following Alignment for Retrieval-Augmented Generation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, Toby Walsh, Julie Shah, and Zico Kolter (Eds.). AAAI Press, 23796–23804. <https://doi.org/10.1609/AAAI25.23.34551>
- [16] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2024. Progressive Multimodal Reasoning via Active Retrieval. *arXiv preprint arXiv:2412.14835* (2024).
- [17] Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Zhicheng Dou, and Ji-Rong Wen. 2024. Understand what LLM Needs: Dual Preference Alignment for Retrieval-Augmented Generation. *CoRR abs/2406.18676* (2024). <https://doi.org/10.48550/ARXIV.2406.18676>
- [18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*. AAAI Press, 226–231.
- [20] Yuchen Fan, Kaiyan Zhang, Heng Zhou, Yuxin Zuo, Yanxi Chen, Yu Fu, Xinwei Long, Xuekai Zhu, Che Jiang, Yuchen Zhang, Li Kang, Gang Chen, Cheng Huang, Zhizhou He, Bingning Wang, Lei Bai, Ning Ding, and Bowen Zhou. 2025. SSRL: Self-Search Reinforcement Learning. *CoRR abs/2508.10874* (2025). <https://doi.org/10.48550/ARXIV.2508.10874>
- [21] Runnan Fang, Shihao Cai, Baixuan Li, Jialong Wu, Guangyu Li, Wenbiao Yan, Xinyu Wang, Xiaobin Wang, Liangcai Su, Chen Zhang, et al. 2025. Towards General Agentic Intelligence via Environment Scaling. *arXiv preprint arXiv:2509.13311* (2025).
- [22] Tianqing Fang, Hongming Zhang, Zhisong Zhang, Kaixin Ma, Wenhao Yu, Haitao Mi, and Dong Yu. 2025. WebEvolver: Enhancing Web Agent Self-Improvement with Coevolving World Model. *CoRR abs/2504.21024* (2025). <https://doi.org/10.48550/ARXIV.2504.21024>
- [23] Tianqing Fang, Zhisong Zhang, Xiaoyang Wang, Rui Wang, Can Qin, Yuxuan Wan, Jun-Yu Ma, Ce Zhang, Jiaqi Chen, Xiyun Li, Hongming Zhang, Haitao Mi, and Dong Yu. 2025. Cognitive Kernel-Pro: A Framework for Deep Research Agents and Agent Foundation Models Training. *CoRR abs/2508.00414* (2025). <https://doi.org/10.48550/ARXIV.2508.00414>
- [24] Jiazhai Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjuan Zhong. 2025. ReTool: Reinforcement Learning for Strategic Tool Use in LLMs. *arXiv:cs.CL/2504.11536* <https://arxiv.org/abs/2504.11536>
- [25] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. Group-in-Group Policy Optimization for LLM Agent Training. *CoRR abs/2505.10978* (2025). <https://doi.org/10.48550/ARXIV.2505.10978>
- [26] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. 2025. Group-in-Group Policy Optimization for LLM Agent Training. *arXiv preprint arXiv:2505.10978* (2025).
- [27] Jiaxuan Gao, Wei Fu, Minyang Xie, Shusheng Xu, Chuiy He, Zhiyu Mei, Banghua Zhu, and Yu Wu. 2025. Beyond Ten Turns: Unlocking Long-Horizon Agentic Search with Large-Scale Asynchronous RL. *arXiv:cs.CL/2508.07976* <https://arxiv.org/abs/2508.07976>
- [28] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. ToRA: A Tool-Integrated Reasoning Agent for Mathematical Problem Solving. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. <https://openreview.net/forum?id=Ep0TtjVoap>
- [29] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).
- [30] Yuhuan Guo, Cong Guo, Aiwen Sun, Hongliang He, Xinyu Yang, Yue Lu, Yingji Zhang, Xuntao Guo, Dong Zhang, Jianzhuang Liu, Jiang Duan, Yijia Xiao, Liangjian Wen, Hai-Ming Xu, and Yong Dai. 2025. Web-CogReasoner: Towards Knowledge-Induced Cognitive Reasoning for Web Agents. *CoRR abs/2508.01858* (2025). <https://doi.org/10.48550/ARXIV.2508.01858>
- [31] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *CoRR abs/2002.08909* (2020). <https://arxiv.org/abs/2002.08909>
- [32] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information*

- Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual, Joaquin Vanschoren and Sai-Kit Yeung (Eds.). <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html>*
- [33] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A Multi-hop QA Dataset for Comprehensive Evaluation of Reasoning Steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8–13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 6609–6625. <https://doi.org/10.18653/V1/2020.COLING-MAIN.580>
- [34] Zhenyu Hou, Ziniu Hu, Yujiang Li, Rui Lu, Jie Tang, and Yuxiao Dong. 2025. TreeRL: Reinforcement Learning with On-Policy Tree Search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 – August 1, 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, 12355–12369. <https://aclanthology.org/2025.acl-long.604/>
- [35] Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262* (2025).
- [36] Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. 2025. Reinforcement Learning with Rubric Anchors. *CoRR abs/2508.12790* (2025). [https://doi.org/10.48550/ARXIV.2508.12790 arXiv:2508.12790](https://doi.org/10.48550/ARXIV.2508.12790)
- [37] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [38] Yuxiang Ji, Ziyu Ma, Yong Wang, Guanhua Chen, Xiangxiang Chu, and Liaoan Wu. 2025. Tree Search for LLM Agent Reinforcement Learning. *arXiv:cs.LG/2509.21240* <https://arxiv.org/abs/2509.21240>
- [39] Dongfu Jiang, Yi Lu, ZuoFeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, et al. 2025. Ver!Tool: Towards Holistic Agentic Reinforcement Learning with Tool Use. *arXiv preprint arXiv:2509.01055* (2025).
- [40] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-R1: Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning. *CoRR abs/2503.09516* (2025). <https://doi.org/10.48550/ARXIV.2503.09516 arXiv:2503.09516>
- [41] Jiajie Jin, Xiaoxi Li, Guanting Dong, Yutao Zhang, Yutao Zhu, Yang Zhao, Hongjin Qian, and Zhicheng Dou. 2025. Decoupled Planning and Execution: A Hierarchical Reasoning Framework for Deep Search. *arXiv:cs.AI/2507.02652* <https://arxiv.org/abs/2507.02652>
- [42] Jiajie Jin, Yutao Zhu, Xinyi Yang, Chenghao Zhang, and Zhicheng Dou. 2024. FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research. *CoRR abs/2405.13576* (2024). <https://doi.org/10.48550/ARXIV.2405.13576 arXiv:2405.13576>
- [43] Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation. *arXiv:cs.CL/2409.12941* <https://arxiv.org/abs/2409.12941>
- [44] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [45] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). <https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html>
- [46] Chengpeng Li, Guanting Dong, Mingfeng Xue, Ru Peng, Xiang Wang, and Dayiheng Liu. 2024. DataMath: Decomposition of Thought with Code Assistance and Self-correction for Mathematical Reasoning. *CoRR abs/2407.04078* (2024). <https://doi.org/10.48550/ARXIV.2407.04078 arXiv:2407.04078>
- [47] Kuan Li, Zhongwang Zhang, Huifeng Yin, Rui Ye, Yida Zhao, Liwen Zhang, Litou Ou, Dingchu Zhang, Xixi Wu, Jialong Wu, et al. 2025. WebSailor-V2: Bridging the Chasm to Proprietary Agents via Synthetic Data and Scalable Reinforcement Learning. *arXiv preprint arXiv:2509.13305* (2025).
- [48] Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litou Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, Weizhou Shen, Junkai Zhang, Dingchu Zhang, Xixi Wu, Yong Jiang, Ming Yan, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. WebSailor: Navigating Super-human Reasoning for Web Agent. *arXiv:cs.CL/2507.02592* <https://arxiv.org/abs/2507.02592>
- [49] Weizhen Li, Jianbo Lin, Zhusong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qixiang Wang, Hongxuan Lu, Tianrui Qin, Chenghao Zhu, Yi Yao, Shuying Fan, Xiaowan Li, Tiannan Wang, Pai Liu, King Zhu, He Zhu, Dingfeng Shi, Piaohong Wang, Yeyi Guan, Xiangru Tang, Minghao Liu, Yuchen Eleanor Jiang, Jian Yang, Jiaheng Liu, G Zhang, and Wangchunshu Zhou. 2025. Chain-of-Agents: End-to-End Agent Foundation Models via Multi-Agent Distillation and Agentic RL. *CoRR abs/2508.13167* (2025). <https://doi.org/10.48550/ARXIV.2508.13167 arXiv:2508.13167>
- [50] Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. *CoRR abs/2501.05366* (2025). <https://doi.org/10.48550/ARXIV.2501.05366 arXiv:2501.05366>
- [51] Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. WebThinker: Empowering Large Reasoning Models with Deep Research Capability. *arXiv preprint arXiv:2504.21776* (2025).
- [52] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yongkang Wu, Zhonghua Li, Qi Ye, and Zhicheng Dou. 2024. RetroLLM: Empowering Large Language Models to Retrieve Fine-grained Evidence within Generation. *CoRR abs/2412.11919* (2024). <https://doi.org/10.48550/ARXIV.2412.11919 arXiv:2412.11919>
- [53] Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. ToRL: Scaling Tool-Integrated RL. *CoRR abs/2503.23383* (2025). <https://doi.org/10.48550/ARXIV.2503.23383 arXiv:2503.23383>
- [54] Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. 2025. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445* (2025).
- [55] Yizhi Li, Qingshui Gu, Zhoufutu Wen, Ziniu Li, Tianshun Xing, Shuyue Guo, Tianyu Zheng, Xin Zhou, Xingwei Qu, Wangchunshu Zhou, et al. 2025. Treepo: Bridging the gap of policy optimization and efficacy and inference efficiency with heuristic tree-based modeling. *arXiv preprint arXiv:2508.17445* (2025).
- [56] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's Verify Step by Step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net. <https://openreview.net/forum?id=v8L0pN6EOi>
- [57] Jia Liu, ChangYi He, YingQiao Lin, MingMin Yang, FeiYang Shen, and ShaoGuo Liu. 2025. Ettrl: Balancing exploration and exploitation in llm test-time reinforcement learning via entropy mechanism. *arXiv preprint arXiv:2508.11356* (2025).
- [58] Zihe Liu, JiaShun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, et al. 2025. Part I: Tricks or traps? A deep dive into RL for LLM reasoning. *arXiv preprint arXiv:2508.08221* (2025).
- [59] Xing Han Lu, Amirhossein Kazemnejad, Nicholas Meade, Arkil Patel, Dongchan Shin, Alejandra Zambrano, Karolina Stanczak, Peter Shaw, Christopher J. Pal, and Siva Reddy. 2025. AgentRewardBench: Evaluating Automatic Evaluations of Web Agent Trajectories. *CoRR abs/2504.08942* (2025). <https://doi.org/10.48550/ARXIV.2504.08942 arXiv:2504.08942>
- [60] Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for General AI Assistants. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net. <https://openreview.net/forum?id=fibxvahvs3>
- [61] MiniMax. 2025. MiniMax-M1: Scaling Test-Time Compute Efficiently with Lightning Attention. *arXiv:cs.CL/2506.13585* <https://arxiv.org/abs/2506.13585>
- [62] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shaze Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nat.* 518, 7540 (2015), 529–533. <https://doi.org/10.1038/NATURE14236>
- [63] Karthik Narasimhan, Tejas D. Kulkarni, and Regina Barzilay. 2015. Language Understanding for Text-based Games using Deep Reinforcement Learning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17–21, 2015*, Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Mardon (Eds.). The Association for Computational Linguistics, 1–11. <https://doi.org/10.18653/V1/D15-1001>
- [64] OpenAI. 2024. Learning to Reason with LLMs. <https://openai.com/index-learning-to-reason-with-langs>
- [65] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. 2019. Advantage-Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. *CoRR abs/1910.00177* (2019). [arXiv:1910.00177 https://arxiv.org/abs/1910.00177](https://arxiv.org/abs/1910.00177)
- [66] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025.

- Humanity's last exam. *arXiv preprint arXiv:2501.14249* (2025).
- [67] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and Narrowing the Compositional Gap in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6–10, 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, 5687–5711. <https://doi.org/10.18653/V1/2023.FINDINGS-EMNLP.378>
- [68] Cheng Qian, Emre Can Acikgoz, Qi He, Hongru Wang, Xiusi Chen, Dilek Hakkani-Tür, Gokhan Tur, and Heng Ji. 2025. Toolrl: Reward is all tool learning needs. *arXiv preprint arXiv:2504.13958* (2025).
- [69] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhu Wu, Minhu Wu, Jiapeng Wang, Yifan Zhang, Zhuoma GongQue, Chong Sun, Yida Xu, Yadong Xue, et al. 2025. V-oracle: Making progressive reasoning in deciphering oracle bones for you and me. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2024–2050.
- [70] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhu Wu, Chong Sun, Xiaoshuai Song, Zhuoma Gongque, Shanglin Lei, Zhi Wei, Miaoqian Zhang, R unfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Dia o, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-Math: Does Your Large Multimodal Model Achieve Human-like Mathematical Reasoning? *CoRR abs/2407.01284* (2024). <https://doi.org/10.48550/ARXIV.2407.01284> arXiv:2407.01284
- [71] Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. 2025. We-Math 2.0: A Versatile MathBook System for Incentivizing Visual Mathematical Reasoning. *CoRR abs/2508.10433* (2025). <https://doi.org/10.48550/ARXIV.2508.10433> arXiv:2508.10433
- [72] Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, et al. 2025. WebResearcher: Unleashing unbounded reasoning capability in Long-Horizon Agents. *arXiv preprint arXiv:2509.13309* (2025).
- [73] Yujia Qin, Yining Ye, Junjie Fang, Haoming Wang, Shihao Liang, Shizuo Tian, Junda Zhang, Jiahao Li, Yunxin Li, Shijue Huang, et al. 2025. UI-TARS: Pioneering Automated GUI Interaction with Native Agents. *arXiv preprint arXiv:2501.12326* (2025).
- [74] 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 Tian, 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, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv:cs.CL/2412.15115* <https://arxiv.org/abs/2412.15115>
- [75] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR abs/1707.06347* (2017). <https://doi.org/10.48550/ARXIV.2402.03300> arXiv:1707.06347
- [76] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *CoRR abs/2402.03300* (2024). <https://doi.org/10.48550/ARXIV.2402.03300> arXiv:2402.03300
- [77] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2024. HybridFlow: A Flexible and Efficient RLHF Framework. *arXiv preprint arXiv:2409.19256* (2024).
- [78] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Jonatan Bisk, Adam Trischler, and Matthew Hausknecht. 2020. Alfworld: Aligning text and embodied environments for interactive learning. *arXiv preprint arXiv:2010.03768* (2020).
- [79] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy P. Lillicrap, Karen Simonyan, and Demis Hassabis. 2017. Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm. *CoRR abs/1712.01815* (2017). <https://doi.org/10.48550/ARXIV.2503.05592> arXiv:1712.01815
- [80] Joykirat Singh, Raghab Magazine, Yash Pandya, and Akshay Nambi. 2025. Agen-tic Reasoning and Tool Integration for LLMs via Reinforcement Learning. *arXiv preprint arXiv:2505.01441* (2025).
- [81] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement Learning. *CoRR abs/2503.05592* (2025). <https://doi.org/10.48550/ARXIV.2503.05592> arXiv:2503.05592
- [82] Huatong Song, Jinhao Jiang, Wenqing Tian, Zhipeng Chen, Yuhuan Wu, Jiahao Zhao, Yingqian Min, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-Searcher++: Incentivizing the Dynamic Knowledge Acquisition of LLMs via Reinforcement Learning. *CoRR abs/2505.17005* (2025). <https://doi.org/10.48550/ARXIV.2505.17005> arXiv:2505.17005
- [83] Liangcai Su, Zhen Zhang, Guangyu Li, Zhuo Chen, Chenxi Wang, Maojia Song, Xinyu Wang, Kuan Li, Jialong Wu, Xuanzhong Chen, et al. 2025. Scaling agents via continual pre-training. *arXiv preprint arXiv:2509.13310* (2025).
- [84] Zhenpeng Su, Leiyu Pan, Xue Bai, Dening Liu, Guanting Dong, Jiaming Huang, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025. Klear-Reasoner: Advancing Reasoning Capability via Gradient-Preserving Clipping Policy Optimization. *CoRR abs/2508.07629* (2025). <https://doi.org/10.48550/ARXIV.2508.07629> arXiv:2508.07629
- [85] Zhenpeng Su, Leiyu Pan, Minxuan Lv, Yuntao Li, Wenping Hu, Fuzheng Zhang, Kun Gai, and Guorui Zhou. 2025. CE-GPO: Controlling Entropy via Gradient-Preserving Clipping Policy Optimization in Reinforcement Learning. *arXiv preprint arXiv:2509.20712* (2025).
- [86] Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Yan Zhang, Fei Huang, and Jingren Zhou. 2025. ZeroSearch: Incentivize the Search Capability of LLMs without Searching. *arXiv:cs.CL/2505.04588* <https://arxiv.org/abs/2505.04588>
- [87] Jiejun Tan, Zhicheng Dou, Wen Wang, Mang Wang, Weipeng Chen, and Ji-Rong Wen. 2024. HtmlRAG: HTML is Better Than Plain Text for Modeling Retrieved Knowledge in RAG Systems. *CoRR abs/2411.02959* (2024). <https://doi.org/10.48550/ARXIV.2411.02959> arXiv:2411.02959
- [88] Jiejun Tan, Zhicheng Dou, Yan Yu, Jiehan Cheng, Qiang Ju, Jian Xie, and Ji-Rong Wen. 2025. HierSearch: A Hierarchical Enterprise Deep Search Framework Integrating Local and Web Searches. *arXiv preprint arXiv:2508.08088* (2025).
- [89] Weihao Tan, Wentao Zhang, Shanqi Liu, Longtao Zheng, Xinrun Wang, and Bo An. 2024. True Knowledge Comes from Practice: Aligning Large Language Models with Embodied Environments via Reinforcement Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024*. OpenReview.net. <https://openreview.net/forum?id=hILVmJ4Uvu>
- [90] Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. WebShaper: Agentically Data Synthesizing via Information-Seeking Formalization. *CoRR abs/2507.15061* (2025). <https://doi.org/10.48550/ARXIV.2507.15061> arXiv:2507.15061
- [91] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534* (2025).
- [92] Kimi Team, Angang Du, Bofei Bo, Bofei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599* (2025).
- [93] Meituan LongCat Team, Bei Li, Bingye Lei, Bo Wang, Bolin Rong, Chao Wang, Chao Zhang, Chen Gao, Chen Zhang, Cheng Sun, et al. 2025. LongCat-Flash Technical Report. *arXiv preprint arXiv:2509.01322* (2025).
- [94] Qwen Team. 2024. QwQ: Reflect Deeply on the Boundaries of the Unknown. <https://qwenlm.github.io/blog/qwq-32b-preview/>
- [95] Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown. *Hugging Face* (2024).
- [96] Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
- [97] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. *Transactions of the Association for Computational Linguistics* 10 (2022), 539–554.
- [98] Hongru Wang, Cheng Qian, Manling Li, Jiahao Qiu, Boyang Xue, Mengdi Wang, Heng Ji, and Kam-Fai Wong. 2025. Toward a Theory of Agents as Tool-Use Decision-Makers. *CoRR abs/2506.00886* (2025). <https://doi.org/10.48550/ARXIV.2506.00886> arXiv:2506.00886
- [99] Hongru Wang, Cheng Qian, Wanjun Zhong, Xiusi Chen, Jiahao Qiu, Shijue Huang, Bowen Jin, Mengdi Wang, Kam-Fai Wong, and Heng Ji. 2025. OTC: Optimal Tool Calls via Reinforcement Learning. *arXiv preprint arXiv:2504.14870* (2025).
- [100] Haoming Wang, Haoyang Zou, Huatong Song, Jiazhan Feng, Junjie Fang, Junting Lu, Longxiang Liu, Qinyu Luo, Shihao Liang, Shijue Huang, et al. 2025. Ui-tars-2 technical report: Advancing gui agent with multi-turn reinforcement learning. *arXiv preprint arXiv:2509.02544* (2025).
- [101] 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. 2025. Beyond the 80/20 Rule: High-Entropy Minority Tokens Drive Effective Reinforcement Learning for LLM Reasoning. *CoRR abs/2506.01939* (2025). <https://doi.org/10.48550/ARXIV.2506.01939> arXiv:2506.01939
- [102] Taiyi Wang, Zhihao Wu, Jianheng Liu, Jianye Hao, Jun Wang, and Kun Shao. 2024. DistRL: An Asynchronous Distributed Reinforcement Learning Framework for On-Device Control Agents. *CoRR abs/2410.14803* (2024). <https://doi.org/10.48550/ARXIV.2410.14803> arXiv:2410.14803
- [103] Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Lucas Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, Weizhu Chen, Shuohang Wang, Simon Shaolei Du, and Yelong Shen. 2025. Reinforcement Learning for Reasoning in Large Language Models with One Training Example. *arXiv preprint arXiv:2504.20571* (2025).
- [104] Zihuan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, Eli Gottlieb, Monica Lam, Yiping Lu, Kyunghyun Cho, Jiajun Wu, Li Fei-Fei, Lijuan Wang,

- Yejin Choi, and Manling Li. 2025. RAGEN: Understanding Self-Evolution in LLM Agents via Multi-Turn Reinforcement Learning. *arXiv:cs.LG/2504.20073* <https://arxiv.org/abs/2504.20073>
- [105] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. WebDancer: Towards Autonomous Information Seeking Agency. *arXiv:cs.CL/2505.22648* <https://arxiv.org/abs/2505.22648>
- [106] Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. WebDancer: Towards Autonomous Information Seeking Agency. *CoRR abs/2505.22648* (2025). <https://doi.org/10.48550/ARXIV.2505.22648> arXiv:2505.22648
- [107] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. WebWalker: Benchmarking LLMs in Web Traversal. *CoRR abs/2501.07572* (2025). <https://doi.org/10.48550/ARXIV.2501.07572> arXiv:2501.07572
- [108] Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Deyu Zhou, Pengjun Xie, and Fei Huang. 2025. WebWalker: Benchmarking LLMs in Web Traversal. *arXiv:cs.CL/2501.07572* <https://arxiv.org/abs/2501.07572>
- [109] Xixi Wu, Kuan Li, Yida Zhao, Liwen Zhang, Litu Ou, Huifeng Yin, Zhongwang Zhang, Yong Jiang, Pengjun Xie, Fei Huang, Minhao Cheng, Shuai Wang, Hong Cheng, and Jingren Zhou. 2025. ReSum: Unlocking Long-Horizon Search Intelligence via Context Summarization. *arXiv preprint arXiv:2509.13313* (2025).
- [110] Yang Xiao, Mohan Jiang, Jie Sun, Keyu Li, Jifan Lin, Yumin Zhuang, Ji Zeng, Shijie Xie, Qishuo Hua, Xuefeng Li, et al. 2025. LIMI: Less is More for Agency. *arXiv preprint arXiv:2509.17567* (2025).
- [111] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. 2024. LLaVA-01: Let Vision Language Models Reason Step-by-Step. *arXiv preprint arXiv:2411.10440* (2024).
- [112] Zhenghai Xue, Longtao Zheng, Qian Liu, Yingru Li, Xiaosen Zheng, Zejun Ma, and Bo An. 2025. SimpleTIR: End-to-End Reinforcement Learning for Multi-Turn Tool-Integrated Reasoning. *arXiv preprint arXiv:2509.02479* (2025).
- [113] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binuyan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan 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, Junyong 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. 2025. Qwen3 Technical Report. *CoRR abs/2505.09388* (2025). <https://doi.org/10.48550/ARXIV.2505.09388> arXiv:2505.09388
- [114] Zhicheng Yang, Zhijiang Guo, Yinya Huang, Xiaodan Liang, Yiwei Wang, and Jing Tang. 2025. TreeRPO: Tree Relative Policy Optimization. *arXiv preprint arXiv:2506.05183* (2025).
- [115] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629* (2022).
- [116] Dian Yu, Baolin Peng, Ye Tian, Linfeng Song, Haitao Mi, and Dong Yu. 2024. SlaM: Self-Improving Code-Assisted Mathematical Reasoning of Large Language Models. *CoRR abs/2408.15565* (2024). <https://doi.org/10.48550/ARXIV.2408.15565> arXiv:2408.15565
- [117] Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiyng Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2025. MemAgent: Reshaping Long-Context LLM with Multi-Conv RL-based Memory Agent. *CoRR abs/2507.02259* (2025). <https://doi.org/10.48550/ARXIV.2507.02259> arXiv:2507.02259
- [118] Qiyng 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. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. *CoRR abs/2503.14476* (2025). <https://doi.org/10.48550/ARXIV.2503.14476> arXiv:2503.14476
- [119] Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, Yean Cheng, Yifan An, Yilin Niu, Yuanhao Wen, Yushi Bai, Zhengxiao Du, Zihan Wang, Zilin Zhu, Bohan Zhang, Bosi Wen, Bowen Wu, Bowen Xu, Can Huang, Casey Zhao, Changpeng Cai, Chao Yu, Chen Li, Chendi Ge, Chenghui Huang, Chenhui Zhang, Chenxi Xu, Chenzheng Zhu, Chuang Li, Congfeng Yin, Daoyan Lin, Dayong Yang, Dazhi Jiang, Ding Ai, Erle Zhu, Fei Wang, Gengzheng Pan, Guo Wang, Hailong Sun, Haitao Li, Haiyang Li, Haiyi Hu, Hanyu Zhang, Hao Peng, Hao Tai, Haoke Zhang, Haoran Wang, Haoyu Yang, He Liu, He Zhao,
- Hongwei Liu, Hongxi Yan, Huan Liu, Hui long Chen, Ji Li, Jiajng Zhao, Jiamin Ren, Jian Jiao, Jian Zhao, Jianyang Yan, Jiaqi Wang, Jiayi Gui, Jiayue Zhao, Jie Liu, Jijie Li, Jing Li, Jing Lu, Jingsen Wang, Jingwei Yuan, Jingxuan Li, Jingzhao Du, Jinhua Du, Jinxin Liu, Junkai Zhi, Junli Gao, Ke Wang, Lekang Yang, Liang Xu, Lin Fan, Lindong Wu, Lintao Ding, Lu Wang, Man Zhang, Minghao Li, Minghuan Xu, Mingming Zhao, and Mingshu Zhai. 2025. GLM-4.5: Agentic, Reasoning, and Coding (ARC) Foundation Models. *CoRR abs/2508.06471* (2025). <https://doi.org/10.48550/ARXIV.2508.06471> arXiv:2508.06471
- [120] Simon Zhai, Hao Bai, Zipeng Lin, Jiayi Pan, Peter Tong, Yifei Zhou, Alane Suhr, Saining Xie, Yann LeCun, Yi Ma, and Sergey Levine. 2024. Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/c848b7d3adc08fc0bf1df3101ba6728-Abstract-Conference.html
- [121] Dingchu Zhang, Yida Zhao, Jialong Wu, Baixuan Li, Wenbiao Yin, Liwen Zhang, Yong Jiang, Yufeng Li, Kewei Tu, Pengjun Xie, and Fei Huang. 2025. EvolveSearch: An Iterative Self-Evolving Search Agent. *CoRR abs/2505.22501* (2025). <https://doi.org/10.48550/ARXIV.2505.22501> arXiv:2505.22501
- [122] Guibin Zhang, Hejia Geng, Xiaohang Yu, Zhenfei Yin, Zaibin Zhang, Zelin Tan, Heng Zhou, Zhongzhi Li, Xiangyuan Xue, Yijiang Li, et al. 2025. The Landscape of Agentic Reinforcement Learning for LLMs: A Survey. *arXiv preprint arXiv:2509.02547* (2025).
- [123] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR abs/2309.01219* (2023). <https://doi.org/10.48550/ARXIV.2309.01219> arXiv:2309.01219
- [124] 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. 2025. Group Sequence Policy Optimization. *CoRR abs/2507.18071* (2025). <https://doi.org/10.48550/ARXIV.2507.18071> arXiv:2507.18071
- [125] Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, et al. 2025. First Return, Entropy-Eliciting Explore. *arXiv preprint arXiv:2507.07017* (2025).
- [126] Tianyu Zheng, Tianshun Xing, Qingshui Gu, Taoran Liang, Xingwei Qu, Xin Zhou, Yizhi Li, Zhoufutu Wen, Chenghua Lin, Wenhao Huang, et al. 2025. First return, entropy-eliciting explore. *arXiv preprint arXiv:2507.07017* (2025).
- [127] Huichi Zhou, Yihang Chen, Siyuan Guo, Xue Yan, Kin Hei Lee, Zihan Wang, Ka Yiu Lee, Guchun Zhang, Kun Shao, Linyi Yang, et al. 2025. Agentfly: Fine-tuning llm agents without fine-tuning llms. *arXiv preprint arXiv:2508.16153* (2025).
- [128] Yuanchen Zhou, Shuo Jiang, Jie Zhu, Junhui Li, Lifan Guo, Feng Chen, and Chi Zhang. 2025. Fin-PRM: A Domain-Specialized Process Reward Model for Financial Reasoning in Large Language Models. *CoRR abs/2508.15202* (2025). <https://doi.org/10.48550/ARXIV.2508.15202> arXiv:2508.15202
- [129] Yifei Zhou, Andrea Zanette, Jiayi Pan, Sergey Levine, and Aviral Kumar. 2024. Archer: Training language model agents via hierarchical multi-turn rl. *arXiv preprint arXiv:2402.19446* (2024).
- [130] Kun Zhu, Xiaocheng Feng, Xiyuan Du, Yuxuan Gu, Weijiang Yu, Haotian Wang, Qianglong Chen, Zheng Chu, Jingchang Chen, and Bing Qin. 2024. An information bottleneck perspective for effective noise filtering on retrieval-augmented generation. *arXiv preprint arXiv:2406.01549* (2024).

Appendix

A Proof of the Gradient of AEPO

In this section, we will comprehensively detail the theoretical derivation of AEPO's forward propagation formulas and how they lead to the backward propagation formulas. Specifically:

We begin with the loss function:

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \min \left(\delta \tilde{A}^{(t)}, \text{clip} \left(\delta, 1 - \epsilon_l, \frac{1 + \epsilon_h}{\text{sg}(\delta)} \delta \right) \tilde{A}^{(t)} \right) \right], \quad (12)$$

where $\delta = r_t^{(j)}(\theta)$ represents the importance ratio, and $\text{sg}(\cdot)$ is the stop-gradient operator. Given that $\tilde{A}^{(t)}$ is a constant and $\nabla_\theta \delta = \delta \phi_\theta(a_{j,t}, s_{j,t})$, the gradient of $f(\delta)$ can be expressed as:

$$\nabla_\theta f(\delta) = \tilde{A}^{(t)} s(\delta) \delta \phi_\theta(a_{j,t}, s_{j,t}), \quad (13)$$

where $s(\delta)$ depends on the range of δ . Therefore, we consider three scenarios:

(1) If $\tilde{A}^{(t)} > 0$ and $\delta > 1 + \epsilon_h$: The upper clipping boundary is active, so $\partial f / \partial \delta = (1 + \epsilon_h) / \text{sg}(\delta)$, effectively simplifying to $(1 + \epsilon_h)$.

(2) If $\tilde{A}^{(t)} < 0$ and $\delta < 1 - \epsilon_l$: The lower clipping boundary dominates, leading to $\partial f / \partial \delta = 0$, causing the gradient to vanish.

The region is unclipped, resulting in $\partial f / \partial \delta = \delta$.

By combining all cases, we derive:

$$\mathcal{F}_{j,t}(\theta) = \begin{cases} 1 + \epsilon_h, & \tilde{A}^{(t)} > 0, \delta > 1 + \epsilon_h, \\ 0, & \tilde{A}^{(t)} < 0, \delta < 1 - \epsilon_l, \\ \delta, & \text{otherwise.} \end{cases} \quad (14)$$

Thus, the gradient update is given by:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \mathcal{F}_{j,t}(\theta) \cdot \phi_\theta(a_{j,t}, s_{j,t}) \cdot \tilde{A}^{(t)} \right]. \quad (15)$$

B Discussion of the Gradient Forms in Clipping-optimized RL

In this section, we discuss the gradient differences between AEPO and clipping-optimized RL algorithms to gain insight into the differences in their policy update stages [85].

B.1 CISPO

$$\mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \delta \tilde{A}^{(t)} \log \pi_\theta(a_t^{(j)} | s_t^{(j)}) \right]. \quad (16)$$

By expanding the gradient of the loss function, we obtain:

$$\nabla_\theta \mathcal{L} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \mathcal{F}_{j,t}(\theta) \phi_\theta(a_{j,t}, s_{j,t}) \tilde{A}^{(t)} \right], \quad (17)$$

where

$$\mathcal{F}_{j,t}(\theta) = \begin{cases} 1 - \epsilon_l, & \tilde{A}^{(t)} < 0, \delta < 1 - \epsilon_l, \\ 1 + \epsilon_h, & \tilde{A}^{(t)} > 0, \delta > 1 + \epsilon_h, \\ 1 - \epsilon_l, & \tilde{A}^{(t)} > 0, \delta < 1 - \epsilon_l, \\ 1 + \epsilon_h, & \tilde{A}^{(t)} < 0, \delta > 1 + \epsilon_h, \\ \delta, & \text{otherwise.} \end{cases} \quad (18)$$

As shown in Eq. (14), AEPO modifies the CISPO objective by introducing an asymmetric clipping rule that deactivates gradient flow when both $\tilde{A}^{(t)} < 0$ and $\delta < 1 - \epsilon_l$. In CISPO, the gradient factor remains $F_t(\theta) = 1 - \epsilon_l$ for this

region, propagating a fixed penalty regardless of sample reliability. AEPO, instead, sets $F_t(\theta) = 0$, effectively filtering out low-confidence negative advantages. This simple but principled change prevents unstable gradient signals from low-likelihood rollouts and reduces the variance introduced by symmetric updates. Consequently, AEPO achieves smoother optimization dynamics and more stable convergence, especially under high-entropy exploration regimes where CISPO often exhibits oscillatory behavior. Figure 8 provides experimental evidence for this discussion.

B.2 GPPO

$$\mathcal{L}^{\text{GPPO}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \ell^{(t)} \right], \quad (19)$$

where

$$\ell^{(t)} = \begin{cases} \beta_1 \cdot \frac{1 - \epsilon_l}{\text{sg}(\delta)} \delta \tilde{A}^{(t)}, & \tilde{A}^{(t)} < 0, \delta < 1 - \epsilon_l, \\ \beta_2 \cdot \frac{1 + \epsilon_h}{\text{sg}(\delta)} \delta \tilde{A}^{(t)}, & \tilde{A}^{(t)} > 0, \delta > 1 + \epsilon_h, \\ \delta \tilde{A}^{(t)}, & \text{otherwise.} \end{cases} \quad (20)$$

By expanding its gradient, we have:

$$\nabla_\theta \mathcal{L}^{\text{GPPO}} = \mathbb{E}_{x \sim \mathcal{D}} \left[\frac{1}{\sum_{j=1}^G T_j} \sum_{j=1}^G \sum_{t=1}^{T_j} \mathcal{F}_{j,t}(\theta) \phi_\theta(a_{j,t}, s_{j,t}) \tilde{A}^{(t)} \right], \quad (21)$$

where

$$\mathcal{F}_{j,t}(\theta) = \begin{cases} \beta_1(1 - \epsilon_l), & \tilde{A}^{(t)} < 0, \delta < 1 - \epsilon_l, \\ \beta_2(1 + \epsilon_h), & \tilde{A}^{(t)} > 0, \delta > 1 + \epsilon_h, \\ \delta, & \text{otherwise.} \end{cases} \quad (22)$$

Compared with GPPO, which retains bounded (non-zero) gradients inside clipped regions via its β -scaled correction terms, AEPO enforces a stricter rule: **residual gradients in the region $\tilde{A}^{(t)} < 0, \delta < 1 - \epsilon_l$ are discarded (i.e., $F_t(\theta) = 0$)**. Empirically, we find that agentic RL training is particularly sensitive to GPPO-style pessimistic suppression. Since our goal is to exploit high-entropy tokens with positive rewards, excessive pessimism can hamper effective credit assignment for these tokens. AEPO therefore removes residual negative updates while allowing high-entropy, positively rewarded tokens to fully contribute to the gradient, improving both stability and credit propagation in long-horizon agentic tasks.

C Baselines

In this section, we provide a detailed overview of the baseline models involved in all experiments, as follows:

C.1 RL algorithms

(1) Classical RL Method

- **GRPO** [76] is a reinforcement learning algorithm for fine-tuning large language models via group-based policy optimization. It optimizes model behaviors by comparing responses within sampled groups and assigning relative rewards, enabling more stable and sample-efficient policy updates.

- **Reinforce++** [35] extends the classic policy-gradient algorithm by incorporating variance reduction and adaptive normalization techniques. It improves training stability and sample efficiency when fine-tuning language models with scalar rewards, while keeping the overall objective aligned with standard REINFORCE.

(2) Clipping-optimized RL Method

- **DAPo** [118] decouples the clipping operation from the policy update to achieve more stable optimization, and introduces a dynamic sampling strategy that adaptively selects training examples to maintain effective

Algorithm 2 Agentic Entropy-Balanced Policy Optimization

Require: Reasoning model π_θ , external tools T , total rollout size k , entropy sensitivity β , branch penalty slope γ , clipping bounds ϵ_l, ϵ_h , entropy-aware weight α

```

1: Input: Dataset  $D$ 
2: Initialize reference model:  $\pi_\theta^{\text{old}} \leftarrow \pi_\theta$ 
3: for  $i = 1$  to  $C$  do
4:   Sample mini-batch  $D_b \subset D$ 
5:   // Dynamic Entropy-Balanced Rollout
6:   for each query  $q \in D_b$  do
7:     Generate 1 complete trajectory  $r$  to obtain  $H_{\text{root}}$  and  $H_{\text{tool}}^{\text{avg}}$ 
8:     Global rollout size  $m \leftarrow k \cdot \sigma(\beta(H_{\text{root}} - H_{\text{tool}}^{\text{avg}}))$ 
9:     Branch rollout size  $b \leftarrow k - m$ 
10:    Initialize rollout pool  $\mathcal{P} \leftarrow \emptyset$ 
11:    Consecutive-high-entropy counter  $l \leftarrow 0$ 
12:    while  $|\mathcal{P}| < m$  do
13:      Sample trajectory  $r$ ; add to  $\mathcal{P}$ 
14:    end while
15:    while  $b > 0$  and  $\exists r_j \in \mathcal{P}$  not terminated do
16:      Select a trajectory  $r \in \mathcal{P}$  at tool-call step  $t$ 
17:       $\Delta H_t \leftarrow \text{Normalize}(H_t - H_{\text{initial}})$ 
18:      Consecutive penalty  $\hat{P}(l) \leftarrow \gamma \cdot l$ 
19:      Branch probability  $P_t \leftarrow (\alpha + \beta \Delta H_t)(1 - \hat{P}(l))$ 
20:      if  $P_t > \tau$  then
21:        Branch  $Z$  sub-trajectories;  $b \leftarrow b - Z$ 
22:      else
23:         $l \leftarrow l + 1$  if  $\Delta H_t > 0$ 
24:      end if
25:    end while
26:    if  $b > 0$  then
27:      Sample  $b$  additional trajectories and add to  $\mathcal{P}$ 
28:    end if
29:  end for
30:  // Entropy-Balanced Policy Optimization
31:  for step = 1 to  $S$  do
32:    Compute standard advantage  $\hat{A}_{\text{Acc}}$  and entropy advantage  $\hat{A}_{\Delta H}$  via Eq. (10)
33:    Entropy-aware advantage  $\hat{A} \leftarrow \hat{A}_{\text{Acc}} \cdot (1 + \hat{A}_{\Delta H})^\alpha$ 
34:    for each token  $t$  in trajectory  $j$  do
35:      Importance ratio  $\delta \leftarrow \pi_\theta / \pi_{\theta^{\text{old}}}$ 
36:      if  $\delta > 1 + \epsilon_h$  and  $\hat{A} > 0$  then
37:        Gradient scaler  $\mathcal{F}_{j,t} \leftarrow 1 + \epsilon_h$ 
38:      else if  $\delta < 1 - \epsilon_l$  and  $\hat{A} < 0$  then
39:        Gradient scaler  $\mathcal{F}_{j,t} \leftarrow 0$ 
40:      else
41:        Gradient scaler  $\mathcal{F}_{j,t} \leftarrow \delta$ 
42:      end if
43:    end for
44:    Update parameters via Eq. 15
45:  end for
46: end for
47: Output: Fine-tuned model  $\pi_\theta$ 

```

gradient signals. These techniques together improve training efficiency and prevent performance degradation in long-horizon reasoning tasks.

- **GPPO** [84] extends the PPO framework by decoupling the clipping operation between the forward and backward passes. During optimization, the policy ratio is clipped in the forward computation to ensure bounded updates, while the original, unclipped ratio is used in the backward path to preserve complete gradient information.
- **CISPO** [61] reformulates ratio clipping by applying the constraint to importance sampling weights instead of policy ratios. It bounds update magnitudes in expectation while preserving token-level gradient information through unclipped policy ratios.

(3) **Agentic RL Method**

- **GIGPO** [26] groups complete trajectories at episode level to compute macro-relative advantages, and also retroactively groups actions sharing anchor states across trajectories at step level to compute micro-relative advantages. Both levels are combined without using a critic, preserving the critic-free nature while enabling per-step credit signals.
- **ARPO** [14] is an RL method tailored for multi-turn LLM agents. It introduces an entropy-based adaptive rollout scheme that increases sampling in steps with high uncertainty, and incorporates an advantage attribution mechanism to assign credit across branching tool-use interactions.

C.2 Web Search Agent

- **RAG** [44] (Retrieval-Augmented Generation) combines information retrieval with generative modeling to enhance the accuracy, reliability, and timeliness of outputs. It retrieves relevant information from an external knowledge base before generating responses, addressing internal knowledge gaps and reducing hallucinations.
- **Search-o1** [50] is a framework designed to enhance reasoning by integrating agentic RAG mechanisms with a Reason-in-Documents module. It improves accuracy, coherence, and reliability in reasoning tasks, outperforming native reasoning and traditional RAG methods in complex scenarios.
- **WebThinker** [51] is an open-source framework developed by Renmin University of China, enabling LLMs to autonomously search, explore web pages, and generate research reports. It employs direct preference optimization and iterative synthesis tools to enhance tool utilization capabilities.
- **ReAct** [115] combines reasoning and action to tackle complex tasks effectively. It allows models to generate reasoning steps and use external tools, such as search engines and databases, during decision-making, optimizing results through iterative processes.

D The Overall Algorithm Workflow of AEPO

In this section, we delve into the overall workflow of the Agentic Entropy-Balanced Policy Optimization (AEPO) algorithm, as depicted in Algorithm Diagram 2. The AEPO algorithm integrates dynamic entropy-balanced rollouts with entropy-balanced policy optimization to enhance multi-turn tool-use capabilities in large language models.