HINDSIGHT-ANCHORED POLICY OPTIMIZATION: TURNING FAILURE INTO FEEDBACK IN SPARSE REWARD SETTINGS

Yuning Wu*, **Ke Wang***, **Devin Chen, Kai Wei**
Amazon
{yuningwu, kewangv, devichen, kaiwe}@amazon.com

ABSTRACT

Reinforcement Learning with Verifiable Rewards (RLVR) has emerged as a promising paradigm for post-training reasoning models. However, group-based methods such as Group Relative Policy Optimization (GRPO) face a critical dilemma in sparse-reward settings: pure Reinforcement Learning (RL) suffers from advantage collapse and high-variance gradient estimation, while mixed-policy optimization introduces persistent distributional bias. To resolve this dilemma, we introduce Hindsight-Anchored Policy Optimization (HAPO). HAPO employs the Synthetic Success Injection (SSI) operator, a hindsight mechanism that selectively anchors optimization to teacher demonstrations during failure. This injection is governed by a Thompson sampling-inspired gating mechanism, creating an autonomous, self-paced curriculum. Theoretically, we demonstrate that HAPO achieves *asymptotic consistency*: by naturally annealing the teacher signal as the policy improves, HAPO recovers the unbiased on-policy gradient. This ensures off-policy guidance acts as a temporary scaffold rather than a persistent ceiling, enabling the model to surpass the limitations of static teacher forcing.

1 INTRODUCTION

Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2025) provides a critical mechanism for enhancing the reasoning capabilities of large language models. While standard Reinforcement Learning (RL) (Sutton & Barto, 2018) allows models to explore diverse solution paths and collect environmental feedback, its effectiveness is limited by the base model's initialization and suffers from inefficient exploration in sparse-reward environments (Yue et al., 2025; Zeng et al., 2025). Conversely, Supervised Fine-Tuning (SFT) (Ouyang et al., 2022; Wei et al., 2022) efficiently distills expert knowledge for rapid adaptation, but it is prone to overfitting and catastrophic forgetting. The prevailing "SFT-then-RL" recipe (Yoshihara et al., 2025) combines these approaches sequentially, but encounters inherent *distribution drift*: SFT constrains the model to a narrow imitation-based manifold that sometimes conflicts with RL's exploration requirements. As the model explores, its policy distribution often drifts away from expert behaviors, leading to suboptimal updates and the forgetting of verified reasoning patterns.

To circumvent these challenges, recent work has focused on integrating RL and SFT within a unified training framework (Zhang et al., 2025; Lv et al., 2026; Yan et al., 2025; Fu et al., 2025; Liu et al., 2025a; Ma et al., 2025; Su et al., 2025; Huang et al., 2025). In these works, the model policy is trained to maximize a composite objective function containing both RL and SFT objectives using predefined masking strategies at various granularities (token, sample, or group level), where selected RL-generated content is replaced with teacher demonstrations. However, these methods treat all samples equally and use static replacement strategies that ignore the dynamic training context. Additionally, the distribution shift between self-exploration trajectories and teacher demonstrations leads to suboptimal learning dynamics. This raises a key question: *How can we adaptively determine when to leverage SFT guidance versus RL exploration while mitigating distribution shift?*

*Equal contribution

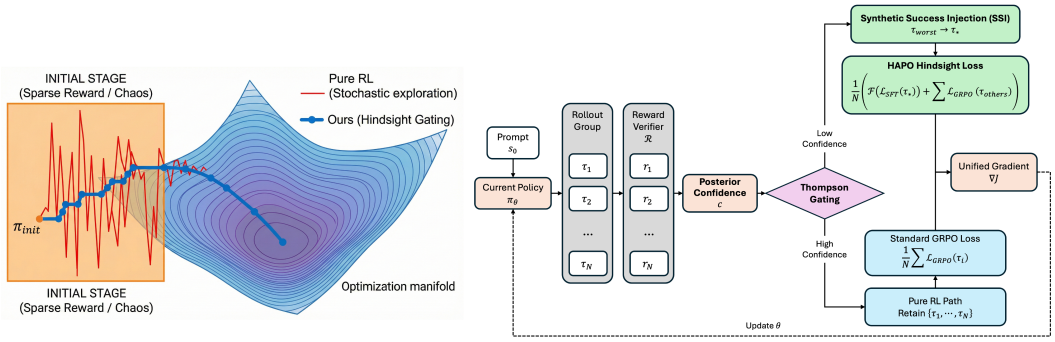

Figure 1: Hindsight-Anchored Policy Optimization (HAPO) system architecture

In this paper, we propose **Hindsight-Anchored Policy Optimization (HAPO)** to address the challenge of adaptive RL-SFT integration. Inspired by hindsight experience replay (Andrychowicz et al., 2018), HAPO introduces a dynamic gating mechanism that monitors policy competence via Thompson sampling. Unlike static mixed-policy approaches such as LUFFY (Yan et al., 2025)) and SRFT (Fu et al., 2025) that rely on fixed masking strategies, HAPO responds to distribution drift by selectively anchoring optimization to teacher demonstrations only during low-confidence failure modes, while prioritizing pure RL exploration when the confidence is high. This adaptive anchoring effectively mitigates catastrophic forgetting without compromising the model's ability to generalize beyond the teacher distribution.

Our preliminary evaluations on mathematical reasoning benchmarks indicate that HAPO achieves competitive performance compared to static mixed-policy methods, matching LUFFY's performance on AIME2024 while substantially outperforming it on MATH-500 (+2.4).

Our Contributions We present HAPO, a theoretically grounded framework for robust policy adaptation for resolving the conflict between exploration and imitation. We introduce the Synthetic Success Injection (SSI) operator, a dynamic mechanism that actively offers hindsight correction by anchoring gradient calculations to verified teacher demonstrations during failure modes, particularly in sparse reward scenarios. To govern this intervention, we propose a self-paced reward gating curriculum inspired by Thompson sampling, which dynamically aligns the teacher's influence with the model's evolving competence. Theoretically, we prove that this mechanism ensures *asymptotic consistency*: as the policy improves, the intervention probability naturally vanishes, recovering the unbiased on-policy gradient and effectively eliminating the persistent distributional bias inherent in static mixed-policy approaches.

2 Related Work

Reinforcement Learning for Reasoning The post-training of Large Language Models (LLMs) has recently pivoted toward Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2025). Algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Group Relative Policy Optimization (GRPO) (Shao et al., 2024) have demonstrated that sophisticated behaviors, including self-correction and multi-step Chain-of-Thought (CoT) reasoning, can emerge from simple rule-based feedback. However, recent studies (Yue et al., 2025) found that on-policy RL is fundamentally bounded by the model's initial "cognitive boundaries". In sparse reward settings, these methods frequently encounter a "cold start" problem where the model fails to discover any successful answers (Yu et al., 2025), leading to a lack of guiding signals. HAPO directly addresses this by introducing the Synthetic Success Injection (SSI) operator to anchor optimization specifically during these failure modes.

Challenges in Exploration and Imitation Balancing exploration and imitation in policy optimization remains a fundamental challenge. The sequential "SFT-then-RL" recipe often induces catastrophic forgetting due to the distribution drift between off-policy data and on-policy exploration (Zhang et al., 2025). While mixed-policy methods like LUFFY (Yan et al., 2025) and

CHORD (Zhang et al., 2025) attempt to mitigate the issue via static policy shaping or token-wise weighting, they frequently introduce persistent distributional bias. HAPO distinguishes itself by using the SSI as a dynamic anchor rather than a static constraint, providing hindsight correction that responds to drift without constantly tethering the optimal policy to the teacher's manifold.

Hybrid Post-Training Strategies Recent hybrid strategies like HPT (Lv et al., 2026) and ReLIFT (Ma et al., 2025) switch between SFT and RL based on heuristic performance measurements. In contrast, HAPO employs a Thompson sampling-inspired gating mechanism to establish a principled, self-paced curriculum. Unlike SRFT (Fu et al., 2025), which relies on local sample mixing, HAPO's probabilistic gate ensures that the intervention probability naturally decays to zero as the model's competence improves. This property guarantees *asymptotic consistency*, allowing the framework to eventually recover the unbiased on-policy gradient and surpass the potential limitations of the teacher.

3 Preliminaries and Problem Formulation

In this section, we establish the theoretical foundations underlying our approach by reviewing the relevant mathematical concepts from reinforcement learning to Thompson sampling, and formally define the optimization problem that HAPO aims to solve.

3.1 Markov Decision Process

A Markov Decision Process (MDP) (Sutton & Barto, 2018) is defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the sets of state and action spaces, $\mathcal{P}$ is the transition probability operator, $\mathcal{R}$ is the reward operator, and γ is the discount factor. For LLMs, we reformulate MDP as follows (Murphy, 2025): each state $s_t \in \mathcal{S}$ contains the current context (prompt plus generated tokens), each action $a_t \in \mathcal{A}$ is the next generated token, the state transition probability p_t defined by $\mathcal{P}$ is deterministic, the reward operator $\mathcal{R}$ treats all time steps equally without any temporal decay and the discount factor $\gamma = 1$. For each episode, it consist of state s_t and action a_t in time horizon T steps, denoted as a trajectory $\tau = \{s_0, a_1, \cdots s_T, a_T\}$. The objective is to learn a policy π_θ that maximizes expected return $\mathcal{J}(\theta)$, mathematically:

$$\arg\max_{\theta} \mathcal{J}(\theta) = \mathbb{E}_{\tau \sim \pi_\theta | s_0} \left[\mathcal{R}(\tau) \right] \tag{1}$$

3.2 Group Relative Policy Optimization

The natural approach to maximize the objective in Eq. (1) is Proximal Policy Optimization (PPO) (Schulman et al., 2017). However, PPO requires both actor and critic networks, creating computational and memory bottleneck for training large language models. To address these limitations, Group Relative Policy Optimization (GRPO) (Shao et al., 2024) was proposed as efficient alternatives that eliminate the critic network by using relative performance of grouped trajectories to estimate advantages.

Given a curated dataset $\mathcal{D} = \{(s_0^i, \tau_*^i) : i \in \{1, \ldots, M\}\}$, where s_0^i is the prompt (initial state) and τ_i^* is the teacher trajectory. For each prompt s_0^i, N trajectories are sampled using the old policy $\pi_{\theta_{\text{old}}}$, forming a group of samples denoted as $\mathcal{G}^i = \{\tau_j^i : j \in \{1, \ldots, N\}\}$. The GRPO computes the advantage of each trajectory by normalizing rewards within the group $\mathcal{G}^i$:

$$A_j^i = \frac{\mathcal{R}(\tau_j^i) - \text{mean}(\{\mathcal{R}(\tau_k^i) : \tau_k^i \in \mathcal{G}^i\})}{\text{std}(\{\mathcal{R}(\tau_k^i) : \tau_k^i \in \mathcal{G}^i\})} \tag{2}$$

Considering the clipped surrogate objective from PPO, the GRPO objectives aggregates over all groups:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \frac{1}{\sum_{i=1}^{M} \sum_{j=1}^{N} |\tau_j^i|} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{t=1}^{|\tau_j^i|} \text{CLIP}(r_{j,t}^i(\theta), A_j^i, \epsilon) \tag{3}$$

where $r^i_{j,t}(\theta) = \frac{\pi_\theta(\tau^i_{j,t}|s^i_0,\tau^i_{j,<t})}{\pi_{\theta_{\text{old}}}(\tau^i_{j,t}|s^i_0,\tau^i_{j,<t})}$ is the importance sampling ratio and $\text{CLIP}(r, A, \epsilon) = \min[r \cdot A, \text{clip}(r; 1-\epsilon, 1+\epsilon) \cdot A]$ is an operator to ensures the updated policy remains within a trust region of the old policy. Following recent studies (Yu et al., 2025; Liu et al., 2025b), we exclude the KL penalty as it has minimal impact on performance.

3.3 Thompson Sampling

Thompson sampling (Sutton & Barto, 2018) is a Bayesian approach to the exploration-exploitation tradeoff that selects actions by sampling from the posterior distribution of each action's expected reward. In LLMs, we define the prompt quality parameter $\alpha_{s^i_0} \in [0, 1]$ as the true expected reward under the current policy π_θ. Formally, $\alpha_{s^i_0} = \mathbb{E}_{\tau\sim\pi_\theta|s^i_0}[\mathcal{R}(\tau)]$, which is intractable before trajectory sampling. Since the underlying distribution of $\alpha_{s^i_0}$ is unknown, we model this uncertainty using a Beta distribution $\alpha_{s^i_0} \sim \text{Beta}(1, 1)$. We define the reward operator $\mathcal{R}$ as:

$$\mathcal{R}(\tau) = \begin{cases} 1 & \text{if } \tau \text{ outputs the correct final answer} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

For each prompt s^i_0, the corresponding group of trajectories $\mathcal{G}^i$ can be viewed as Bernoulli trials, where each trajectory either succeeds (reward = 1) or fails (reward = 0) with probability $\alpha_{s^i_0}$. The total number of successes $S_i = \sum_{j=1}^{N} \mathcal{R}(\tau^i_j)$ follows a Binomial distribution $S_i \sim \text{Binomial}(N, \alpha_{s^i_0})$. This allows us to apply Beta-Binomial conjugacy for the posterior distribution (Bishop, 2007):

$$\alpha_{s^i_0} \mid \mathcal{G}^i \sim \text{Beta}(1 + S_i, 1 + N - S_i) \tag{5}$$

The Bayesian confidence score for a given initial state is then defined as the posterior mean:

$$c_i = \frac{1 + S_i}{2 + N} \tag{6}$$

which naturally balances observed performance with prior uncertainty and converges to the empirical success rate as more data is collected.

4 Hindsight-Anchored Policy Optimization

In this section, we detail the design of our HAPO algorithm, including the Synthetic Success Injection (SSI) operator and Thompson sampling-inspired gating mechanism. We then formally define the HAPO objective function and provide convergence analysis and theoretical justification.

4.1 The Synthetic Success Injection (SSI) Operator

When a group $\mathcal{G}^i$ exhibits low confidence, the model's policy requires additional guidance to improve learning. To address this scenario, we define the Synthetic Success Injection (SSI) operator $\mathcal{T}$ that operates at the group level. Within a low-confidence group $\mathcal{G}^i$, the poorest-performing trajectory $j^* = \arg\min_j \mathcal{R}(\tau^i_j)$ is identified and replaced by a high-confidence teacher sample τ^i_* derived from a verified solution, mathematically:

$$\mathcal{T}(\mathcal{G}^i) = \{\tau^i_1, ..., \tau^i_{j^*-1}, \tau^i_*, \tau^i_{j^*+1}, ..., \tau^i_N\} \tag{7}$$

This operator injects high-confidence guidance into groups where the model struggles, enabling more effective learning by anchoring the policy updates with expert demonstrations.

4.2 THOMPSON SAMPLING INSPIRED SELF-PACED REWARD GATING

In a group $\mathcal{G}^i$, applying the operator $\mathcal{T}$ is not always necessary. When most trajectories succeed (e.g., $N-1$ out of N samples receive reward 1), the current policy π_θ already handles the prompt s_0^i confidently. To determine when operator $\mathcal{T}$ is needed, we introduce a Bayesian confidence score inspired by Thompson sampling in Eq. (6). This score, computed as the posterior mean of trajectory success rates, provides a principled measure that determines whether the operator $\mathcal{T}$ should be applied. Algorithm 1 details this procedure.

Algorithm 1 Thompson Sampling-Inspired Gating

Require: Group of trajectories $\{\mathcal{G}^i : i \in \{1, ..., M\}\}$, threshold $\gamma \in (0, 1)$
1: **for** $i = 1$ to M **do**
2: Compute rewards $\mathcal{R}(\tau_j^i)$ and Bayesian confidence score $c_i = \frac{1+S_i}{2+N}$
3: **if** $c_i < \gamma$ **then**
4: $\mathcal{G}^i = \mathcal{T}(\mathcal{G}^i)$ ▷ Low confidence → Replace worst with teacher sample
5: **end if**
6: **end for**
7: **return** $\{\mathcal{G}^i : i \in \{1, ..., M\}\}, \{c_i : i \in \{1, ..., M\}\}$

In practice, the threshold γ can be a constant, sigmoid, or step function to dynamically adjust gating decisions based on training progress. When the Bayesian confidence score c_i is low, the gate opens and we apply operator $\mathcal{T}$ to provide teacher samples τ_*^i for supervised learning. When confidence is high, the gate remains closed and we continue with pure RL. This adaptive mechanism provides hindsight guidance when the model struggles while maintaining exploration when it performs well.

4.3 HAPO OBJECTIVE FUNCTION

After the Thompson sampling-inspired gating, each group $\mathcal{G}^i$ contains both original trajectories $\{\tau_j^i : j \in \{1, \ldots, N\} \backslash \{j^*\}\}$ and teacher trajectories $\{\tau_*^i\}$. The advantage A_j^i for each sample within a group is computed using the same method as in Eq. (2). Considering two trajectory types, the HAPO objective is proposed based on Eq. (1), where original trajectories represent online generation and follow the GRPO policy gradient objective, while teacher trajectories are offline references that require supervised fine-tuning objective, mathematically:

$$\mathcal{J}_{\text{HAPO}}(\theta) = \frac{1}{\sum_{i=1}^{M} \sum_{\tau_j^i \in \mathcal{G}^i} |\tau_j^i|} \sum_{i=1}^{M} \sum_{\tau_j^i \in \mathcal{G}^i} \mathcal{L}(\theta; \tau_j^i) \tag{8}$$

$$\mathcal{L}(\theta; \tau_j^i) = \begin{cases} \sum_{t=1}^{|\tau_*^i|} \mathcal{F}(\pi_\theta(\tau_{j,t}^* | s_0^i, \tau_{j,<t}^*), c_i) & \text{if } \tau_j^i = \tau_*^i \text{ (hindsight anchored)} \\ \sum_{t=1}^{|\tau_j^i|} \text{CLIP}(r_{j,t}^i(\theta), A_j^i, \epsilon) & \text{otherwise} \end{cases} \tag{9}$$

where $\mathcal{F}$ is the policy shaping operator that reshaping probability distribution of actions (tokens) τ_t^* based on Bayesian confidence score c_i .

4.4 THEORETICAL ANALYSIS

In this section, we analyze the convergence properties of HAPO. We demonstrate that our method not only converges to a stationary point but also achieves asymptotic consistency with the pure RL objective, theoretically outperforming static mixed-policy strategies which suffer from persistent asymptotic bias.

4.4.1 CONVERGENCE TO STATIONARY POINT

Let $\hat{g}(\theta)$ denote the stochastic gradient estimator of the HAPO objective $\mathcal{J}_{\text{HAPO}}(\theta)$. Based on the gating mechanism in Algorithm 1, this estimator switches between a hindsight-anchored gradient $\hat{g}_{\text{teach}}$ (when $c_i < \gamma$) and a pure policy gradient $\hat{g}_{\text{RL}}$ (when $c_i \geq \gamma$).

Theorem 4.1 (Convergence)**.** *Assume the policy π_θ is differentiable, the reward function is bounded, and the gradients of both the shaping operator $\mathcal{F}$ and the CLIP loss satisfy $\|\nabla\mathcal{L}\| \leq G$. With a decaying learning rate $\eta_t = \mathcal{O}(1/\sqrt{t})$, the HAPO algorithm converges to a stationary point of the implicit dynamic objective.*

Sketch. The gradient estimator $\hat{g}(\theta)$ is a bounded stochastic variable. Specifically, both the teacher-forced gradient derived from $\mathcal{F}$ (conceptually similar to cross-entropy) and the GRPO gradient (bounded importance weights via clipping) have bounded norms. The variance of the HAPO estimator is bounded by:

$$\mathbb{V}[\hat{g}(\theta)] \leq \max(\mathbb{V}[\hat{g}_{\text{teach}}], \mathbb{V}[\hat{g}_{\text{RL}}]) \leq \sigma^2 < \infty \tag{10}$$

Standard non-convex optimization theory for SGD states that if the gradient estimator has bounded second moments, the algorithm converges such that $\lim_{T\to\infty} \mathbb{E}[\|\nabla\mathcal{J}(\theta_T)\|^2] \to 0$, provided the descent direction is valid. In the Hindsight phase ($c_i < \gamma$), the teacher term τ_*^i provides a high-bias but consistent descent direction, pulling the policy into a region of non-zero rewards. Once confidence improves such that $c_i \geq \gamma$, the algorithm transitions to the pure RL phase, which is unbiased w.r.t. the true reward objective. □

4.4.2 Asymptotic Consistency vs. Mixed-Policy Methods

A key advantage of HAPO over static mixed-policy approaches is the elimination of asymptotic bias.

Theorem 4.2 (Asymptotic Purity)**.** *Let π^* be an optimal policy such that for any prompt s_0^i, the expected success rate $\mu^* > \gamma$. As $\pi_{\theta_t} \to \pi^*$, the probability of applying the biased teacher replacement $\mathcal{T}(\mathcal{G}^i)$ vanishes.*

Proof. Let $S_i = \sum_{j=1}^N \mathcal{R}(\tau_j^i)$ be the number of correct responses in a group. S_i follows a Binomial distribution $B(N, \mu(\theta))$. The Bayesian confidence score defined in Eq. (6) is monotonic in S_i. The gating condition $c_i < \gamma$ is equivalent to $S_i < k_\gamma$, where $k_\gamma = \gamma(2+N) - 1$. As the policy improves such that its success rate $\mu(\theta)$ satisfies $\mu(\theta) > \gamma$, the probability of the low-confidence event decays exponentially via Hoeffding's inequality:

$$P(c_i < \gamma) = P(S_i < k_\gamma) \leq \exp\left(-2N(\mu(\theta) - \gamma)^2\right) \tag{11}$$

Consequently, $\lim_{t\to\infty} P(c_i < \gamma) \to 0$. The expected gradient becomes:

$$\lim_{t\to\infty} \mathbb{E}[\hat{g}_t(\theta)] = \mathbb{E}[\hat{g}_{\text{RL}}(\theta)] = \nabla\mathcal{J}_{\text{RL}}(\theta) \tag{12}$$

In contrast, static mixed-policy methods optimize a static mixture $\mathcal{J}_{\text{mix}} = \mathcal{J}_{\text{RL}} + \lambda\mathcal{J}_{\text{SFT}}$, leading to a stationary point where $\nabla\mathcal{J}_{\text{RL}} = -\lambda\nabla\mathcal{J}_{\text{SFT}} \neq 0$, resulting in persistent bias towards the teacher distribution. □

4.4.3 Bias-Variance Decomposition of Convergence Error

While both HAPO and static mixed-policy methods nominally follow an $\mathcal{O}(1/\sqrt{T})$ convergence rate characteristic of SGD, the composition of the *effective* error differs fundamentally. We analyze the error in terms of the optimality gap with respect to the true RL objective $\mathcal{J}_{\text{RL}}$.

For a static mixed-policy approach, the convergence is bounded by the variance of the mixed estimator and an approximation bias:

$$\mathbb{E}[\|\nabla\mathcal{J}_{\text{RL}}(\theta_T)\|] \lesssim \underbrace{\frac{\sigma_{\text{mix}}}{\sqrt{T}}}_{\text{Optimization Error}} + \underbrace{\lambda\|\nabla\mathcal{L}_{\text{SFT}}(\theta^*_{\text{RL}})\|}_{\text{Asymptotic Bias}} \tag{13}$$

The bias term arises because the optimization stabilizes at the stationary point of the *mixed* objective, not the true RL objective. If the teacher policy is suboptimal (i.e., $\nabla\mathcal{L}_{\text{SFT}} \neq 0$ at the RL optimum), the model remains tethered to the teacher's limitations.

In contrast, HAPO uses the low-variance teacher signal early to reduce gradient variance σ when reward signals are sparse, but eliminates the bias term asymptotically as the gating mechanism deactivates:

$$\mathbb{E}[\|\nabla \mathcal{J}_{\text{RL}}(\theta_T)\|] \lesssim \frac{\sigma_{\text{adaptive}}}{\sqrt{T}} + 0 \tag{14}$$

This implies that for high-precision reasoning tasks where the teacher data provides helpful initialization but may be suboptimal compared to the ground-truth reward, HAPO theoretically allows the model to surpass the teacher, achieving zero asymptotic bias.

5 EXPERIMENTS

In this section, we present implementation details and preliminary experimental evaluations demonstrating HAPO's competitive performance on mathematical reasoning tasks compared to baseline models.[1]

Model	AIME2024	MATH-500	Olympiad
Qwen2.5-Math-7B	16.7	65.2	30.0
GRPO	27.0	83.0	49.2
SFT	30.0	83.6	43.2
SFT-then-RL	30.0	84.8	48.6
SRFT	26.7	<u>85.2</u>	50.0
LUFFY	<u>36.7</u>	84.6	**51.8**
HAPO	**36.7**	**87.0**	<u>51.4</u>

Table 1: Main experiment results on mathematical reasoning benchmarks based on Qwen2.5-Math-7B. **Bold** and <u>underline</u> indicate the best and second-best results, respectively.

5.1 EXPERIMENTAL SETUP

Training Setup We conduct our experiments using OpenR1-Math-46k-8192 (Yan et al., 2025), a curated dataset of verified mathematical reasoning trajectories generated by DeepSeek-R1 (Face, 2025). Following established practices in mathematical reasoning (Yan et al., 2025; Huang et al., 2025; Fu et al., 2025; Lv et al., 2026), we use Qwen2.5-Math-7B (Yang et al., 2024) as our base model and GRPO (Shao et al., 2024) excluding the KL penalty term (Liu et al., 2025b) as our main RL algorithm. Our training configuration includes a batch size of 128, constant learning rate of 1×10^{-6}, and trajectory generation temperature of 1.0. For the operator $\mathcal{T}$, we experiment with groups of size 8 and employ the same policy shaping operator $\mathcal{F}$ as prior work (Yan et al., 2025). The confidence threshold is set to $\gamma = 0.8$, with all remaining hyperparameters following established baselines (Yan et al., 2025; Fu et al., 2025).

Evaluation Setup For evaluation, we use temperature 0.6 and a maximum generation length of 8,192 tokens. We assess our approach on three mathematical reasoning benchmarks: AIME2024 (LI et al., 2024), MATH-500 (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024). Following standard evaluation protocols, we report avg@32 for AIME2024 due to its limited test samples, while using pass@1 for both MATH-500 and OlympiadBench.

Baseline Comparison We evaluate our approach against two categories of baselines. First, we consider pure RL approaches without teacher demonstrations, specifically GRPO (Shao et al., 2024). Second, we compare against methods that incorporate teacher demonstrations through various non-adaptive strategies that apply expert trajectories uniformly without considering group level prompt quality: (1) SFT, which directly trains the model to imitate expert trajectories; (2) SFT-then-RL, following the standard two-stage pipeline where SFT precedes RL; (3) SRFT (Fu et al., 2025), which replaces one trajectory per group with an expert trajectory using SFT token loss; and (4) LUFFY (Yan et al., 2025), which also replaces one trajectory per group with an expert trajectory but incorporates policy shaping for SFT token loss.

[1] HAPO works for both mathematical and general domain reasoning tasks. In this paper, we focus on training and evaluating mathematical reasoning datasets and report the corresponding settings.

5.2 Main Results

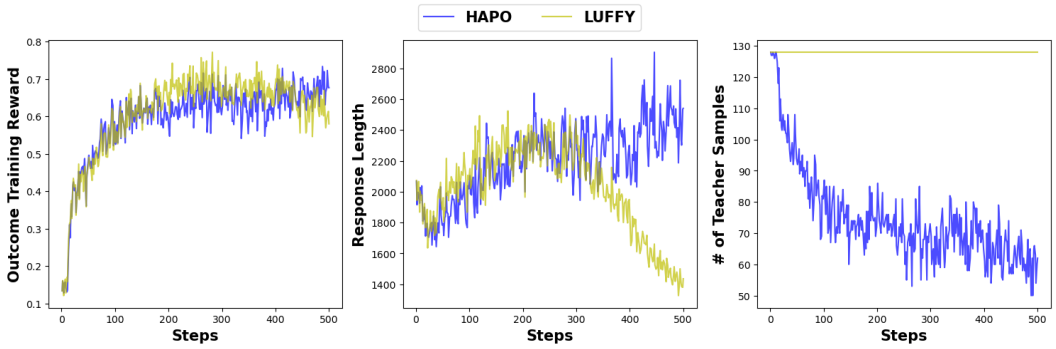

Figure 2: Training dynamics of HAPO compared with LUFFY. From left to right: average reward, generation length, and number of teacher samples during training. For fair comparison, both reward and generation length are computed by excluding trajectories guided by teacher demonstration.

Mathematical Reasoning Performance As demonstrated in Table 1, HAPO achieves strong performance across all benchmarks with scores of 36.7 (AIME2024), 87.0 (MATH-500), and 51.4 (Olympiad). Compared to pure RL methods, HAPO shows substantial improvements over GRPO with gains of **+9.7** (AIME2024), **+4.0** (MATH-500), and **+2.2** (Olympiad). When compared to LUFFY, HAPO achieves competitive performance on AIME2024 while substantially outperforming on MATH-500 with a **+2.4** improvement. These results confirm our central hypothesis that HAPO's adaptive integration of expert knowledge leads to more effective reasoning skill acquisition than both pure RL and static expert guidance approaches.

Training Dynamics Figure 2 illustrates the training dynamics comparison between HAPO and LUFFY, revealing several key differences in their learning behaviors: (1) Both methods achieve competitive reward performance with similar trajectories, indicating comparable optimization effectiveness. (2) The response length analysis shows divergent patterns: while both methods initially maintain longer outputs, LUFFY exhibits a notable decrease in generation length during middle to late-stage training, whereas HAPO sustains consistent response lengths throughout the entire training process. (3) The SFT sample utilization patterns differ markedly: HAPO demonstrates a significant reduction in SFT samples during the early training phase followed by continued fluctuations, suggesting adaptive adjustment to training dynamics. In contrast, LUFFY maintains stable SFT sample usage throughout training, indicating a more static approach to expert guidance integration.

6 Conclusions and Discussion

In this work, we introduced Hindsight-Anchored Policy Optimization (HAPO), an adaptive framework designed to resolve the distribution drift dilemma in RLVR. By coupling Synthetic Success Injection (SSI) operator with a Thompson sampling-inspired gating mechanism, HAPO creates a self-paced curriculum that dynamically anchors optimization to teacher demonstrations only during failure modes, theoretically ensuring asymptotic consistency with the unbiased on-policy gradient.

Crucially, our analysis of training dynamics confirms the efficacy of HAPO's adaptive response strategy. Unlike LUFFY, which maintains static expert utilization and suffers from decreasing generation lengths, HAPO actively anneals its reliance on SFT samples as the policy improves and sustains consistent reasoning lengths throughout training. This behavior validates that HAPO successfully operates as a temporary scaffold rather than a persistent ceiling, mitigating the distributional bias inherent in fixed teacher forcing. Future work will explore scaling and evaluating HAPO on larger foundation models and general domain reasoning tasks.

REFERENCES

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay, 2018. URL https://arxiv.org/abs/1707.01495.

Christopher M. Bishop. *Pattern recognition and machine learning, 5th Edition*. Information science and statistics. Springer, 2007. ISBN 9780387310732. URL https://www.worldcat.org/oclc/71008143.

Hugging Face. Open r1: A fully open reproduction of deepseek-r1, January 2025. URL https://github.com/huggingface/open-r1.

Yuqian Fu, Tinghong Chen, Jiajun Chai, Xihuai Wang, Songjun Tu, Guojun Yin, Wei Lin, Qichao Zhang, Yuanheng Zhu, and Dongbin Zhao. Srft: A single-stage method with supervised and reinforcement fine-tuning for reasoning, 2025. URL https://arxiv.org/abs/2506.19767.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset, 2021. URL https://arxiv.org/abs/2103.03874.

Zeyu Huang, Tianhao Cheng, Zihan Qiu, Zili Wang, Yinghui Xu, Edoardo M. Ponti, and Ivan Titov. Blending supervised and reinforcement fine-tuning with prefix sampling, 2025. URL https://arxiv.org/abs/2507.01679.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Jia LI, Edward Beeching, Lewis Tunstall, Ben Lipkin, Roman Soletskyi, Shengyi Costa Huang, Kashif Rasul, Longhui Yu, Albert Jiang, Ziju Shen, Zihan Qin, Bin Dong, Li Zhou, Yann Fleureau, Guillaume Lample, and Stanislas Polu. Numinamath. [https://huggingface.co/AI-MO/NuminaMath-CoT](https://github.com/project-numina/aimo-progress-prize/blob/main/report/numina_dataset.pdf), 2024.

Mingyang Liu, Gabriele Farina, and Asuman Ozdaglar. Uft: Unifying supervised and reinforcement fine-tuning, 2025a. URL https://arxiv.org/abs/2505.16984.

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

Xingtai Lv, Yuxin Zuo, Youbang Sun, Hongyi Liu, Yuntian Wei, Zhekai Chen, Xuekai Zhu, Kaiyan Zhang, Bingning Wang, Ning Ding, and Bowen Zhou. Towards a unified view of large language model post-training, 2026. URL https://arxiv.org/abs/2509.04419.

Lu Ma, Hao Liang, Meiyi Qiang, Lexiang Tang, Xiaochen Ma, Zhen Hao Wong, Junbo Niu, Chengyu Shen, Runming He, Yanhao Li, Bin Cui, and Wentao Zhang. Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions, 2025. URL https://arxiv.org/abs/2506.07527.

Kevin Murphy. Reinforcement learning: An overview, 2025. URL https://arxiv.org/abs/2412.05265.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Mingyu Su, Jian Guan, Yuxian Gu, Minlie Huang, and Hongning Wang. Trust-region adaptive policy optimization, 2025. URL https://arxiv.org/abs/2512.17636.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning - an introduction, 2nd Edition*. MIT Press, 2018. URL http://www.incompleteideas.net/book/the-book-2nd.html.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL https://arxiv.org/abs/2109.01652.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance, 2025. URL https://arxiv.org/abs/2504.14945.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. URL https://arxiv.org/abs/2409.12122.

Hiroshi Yoshihara, Taiki Yamaguchi, and Yuichi Inoue. A practical two-stage recipe for mathematical llms: Maximizing accuracy with sft and efficiency with reinforcement learning, 2025. URL https://arxiv.org/abs/2507.08267.

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

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild, 2025. URL https://arxiv.org/abs/2503.18892.

Wenhao Zhang, Yuexiang Xie, Yuchang Sun, Yanxi Chen, Guoyin Wang, Yaliang Li, Bolin Ding, and Jingren Zhou. On-policy rl meets off-policy experts: Harmonizing supervised fine-tuning and reinforcement learning via dynamic weighting, 2025. URL https://arxiv.org/abs/2508.11408.