Your Group-Relative Advantage Is Biased

Fengkai Yang[1,3,4] **Zherui Chen**[2] **Xiaohan Wang**[4] **Xiaodong Lu**[1,4] **Jiajun Chai**[4]
Guojun Yin[4] **Wei Lin**[4] **Shuai Ma**[1] **Fuzhen Zhuang**[1*] **Deqing Wang**[1]
Yaodong Yang[3] **Jianxin Li**[1] **Yikun Ban**[1†]

[1]Beihang University [2]University of California, Berkeley [3]Peking University [4]Meituan
yangfengkai@stu.pku.edu.cn yikunban@buaa.edu.cn

Abstract

Reinforcement Learning from Verifier Rewards (RLVR) has emerged as a widely used approach for post-training large language models on reasoning tasks, with group-based methods such as GRPO and its variants gaining broad adoption. These methods rely on group-relative advantage estimation to avoid learned critics, yet its theoretical properties remain poorly understood. In this work, we characterize conditional estimation bias in group-relative advantages among non-degenerate groups. For mean-centered advantages under binary rewards, our analysis shows negative conditional mean estimation errors for hard prompts and positive errors for easy prompts. At the response level, correct answers to hard prompts are underweighted, while incorrect answers to easy prompts receive insufficient penalties. To address this issue, we propose History-Aware Adaptive Difficulty Weighting (**HA-DW**), an adaptive reweighting scheme that adjusts advantage estimates based on an evolving difficulty anchor and training dynamics. Both theoretical analysis and experiments on five mathematical reasoning benchmarks demonstrate that HA-DW consistently improves performance when integrated into GRPO and its variants. Our results suggest that correcting biased advantage estimation is critical for robust and efficient RLVR training. The code is available at https://github.com/yangfengkai/HA-DW

1 Introduction

After the success of DeepSeek-R1 [5], *Reinforcement Learning with Verifiable Rewards* (RLVR) has rapidly emerged as a simple yet powerful paradigm for training reasoning-oriented Large Language Models (LLMs). The key driver behind this leap forward is *Group Relative Policy Optimization* (GRPO) [31]. GRPO leverages *within-group* average reward as a baseline estimator to compute advantages with a small number of generated rollouts for each sampled prompt, thereby avoiding the need for an additional critic model. The group-relative policy optimization paradigm it introduces has gained significant attention due to its simplicity and practical effectiveness. Numerous variants have been proposed to boost training stability and performance, expanding applicability to broader domains such as agent and multimodal training. Representative approaches include GSPO [54], DAPO [48], Dr.GRPO [23], GiGPO [8], and StepGRPO [50]. However, despite its empirical success, the *intra-group* advantage estimation still lacks detailed theoretical characterization which is critical to the performance in group-relative paradigm [45, 36].

In this paper, we characterize the conditional mean estimation error of group-relative advantages within non-degenerate groups. For mean-centered advantages under binary rewards, this error is negative for *hard prompts* and positive for *easy prompts*. A separate reward-conditioned analysis

*Fuzhen Zhuang is affiliated with the School of Artificial Intelligence, Beihang University.
†Corresponding Author. Yikun Ban is affiliated with the State Key Laboratory of Complex & Critical Software Environment, Beihang University. If you have any questions, feel free to contact yikunb@buaa.edu.cn or yangfengkai@stu.pku.edu.cn.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

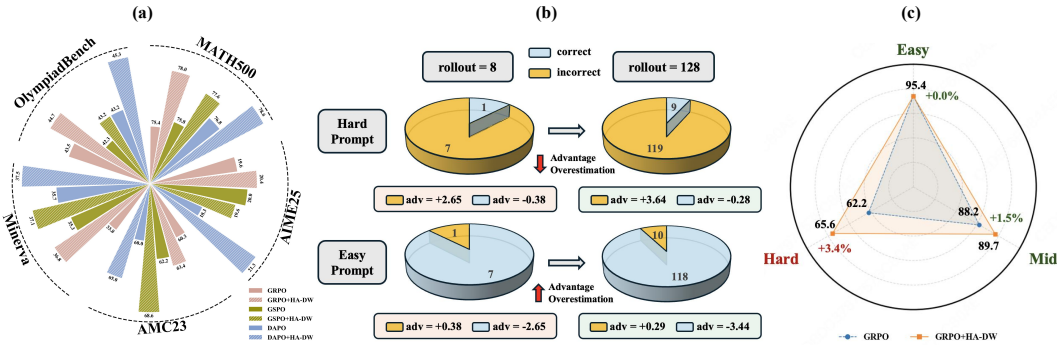

Figure 1: (a) Comparison of the performance of RL algorithms with and without HA-DW on Qwen3-4B-Base across five mathematical reasoning benchmarks. (b) Significant biased advantage estimation on the MATH dataset under 8 and 128 rollouts. (c) Performance gain by GRPO+HA-DW on MATH500 stratified by difficulty levels.

shows that correct responses to hard prompts receive underestimated positive advantages, while incorrect responses to easy prompts receive insufficiently negative advantages, as presented in Section 2.2. Such systematic bias can cause the policy to under-learn from hard questions while over-exploiting easy ones, ultimately hurting both training stability and generalization. As illustrated by the representative example in Figure 1(b), group-relative estimation can introduce substantial bias in advantage estimation for group-based RL algorithms. Our empirical results further corroborate this phenomenon, with consistent evidence reported in Appendix G.1.

Motivated by these findings, we propose a novel policy optimization algorithm *History-Aware Adaptive Difficulty Weighting (HA-DW)* that leverages an evolving difficulty anchor that integrates long-term reward trends and historical training information and adaptively reweights advantage estimates based on the group-relative advantage estimator to mitigate the bias induced by group-based advantage estimation. As a plug-and-play module, it amplifies the signal of hard prompts while reducing overexploitation of easy ones and enables a more principled balance between exploration and exploitation in RL training. This emphasis on learning near the model's capability boundary aligns with the epistemic exploration perspective, which highlights exploration as a mechanism for sustained capability improvement [1]. The overall framework is depicted in Figure 2. Our goal is not to model all RLVR settings, but to expose a previously overlooked statistical bias in group-relative algorithms and demonstrate that even lightweight corrections can yield consistent gains. Our main contributions are summarized as follows:

[Discovery]. We characterize difficulty-dependent conditional mean estimation errors in group-relative advantages and provide reward-conditioned results showing underweighted correct responses to hard prompts and insufficiently penalized incorrect responses to easy prompts.

[Algorithm]. We propose HA-DW that dynamically adjusts advantage weights, theoretically effectively compensate for the bias of group-relative advantage estimation.

[Performance]. As illustrated in Figure 1(a), we validate our approach through extensive experiments on mathematical and code generation benchmarks with robust improvements. Notably, it achieves performance equivalent to increasing sampling under a limited computational budget.

2 Why Your Advantage Estimation is Biased?

In this section, we theoretically analyze the biased estimation in group-relative algorithms. For clarity of exposition, the derivation in this section mainly focuses on the mean-centered group-relative advantage without standard-deviation normalization. Firstly, we provide the prerequisite definitions.

2.1 Definitions

At training step t, we sample a prompt $x_t \sim D$. Given x_t, a group-relative RL algorithm samples G responses $\{y_{t,i}\}_{i=1}^{G}$ independently from the current policy $\pi_{\theta_t}(\cdot \mid x_t)$. Each response $y_{t,i}$ receives a corresponding scalar reward $r_{t,i} \in \{0, 1\}$, forming the reward set $\{r_{t,i}\}_{i=1}^{G}$, where $r(\cdot)$ is the reward

function and we denote $r(y_{t,i})$ by $r_{t,i}$ for brevity. The **group-relative advantage** $\hat{A}_{t,i}$ is computed as:

$$\hat{A}_{t,i} = r_{t,i} - \hat{p}_t, \quad \hat{p}_t = \frac{1}{G}\sum_{i=1}^{G} r_{t,i}, \tag{1}$$

where $\hat{p}_t$ is the group baseline. Let $\pi_{\theta_{\text{old}}}$ denote the reference (behavior) policy. Then, the group-relative policy optimization objective is defined as:

$$J_{\text{group}}(\theta) = \frac{1}{G}\sum_{i=1}^{G} \psi\Big(\frac{\pi_\theta(y_{t,i} \mid x_t)}{\pi_{\theta_{\text{old}}}(y_{t,i} \mid x_t)}\Big)\, \phi\big(\hat{A}_{t,i}\big), \tag{2}$$

Here, $\psi(\cdot)$ denotes a function applied to the importance sampling ratio (e.g., identity, clipping, or logarithmic transformation), and $\phi(\cdot)$ is a transformation applied to the advantage term. These two functions are introduced to unify a broad family of group-relative policy optimization variants under a single formulation.

Definition 1 (Expected Reward)**.** *Assume the reward function $r(\cdot)$ is binary, i.e., $r(\cdot) \in \{0, 1\}$. Given a prompt $x_t \sim D$ and a policy π_{θ_t}, let $y_t \sim \pi_{\theta_t}(\cdot \mid x_t)$ be a sampled response. The expected reward of policy π_{θ_t} on prompt x_t is defined as:*

$$p_t = \mathbb{E}_{y_t \sim \pi_{\theta_t}(\cdot|x_t)}\big[r(y_t)\big] = \mathbb{P}(r(y_t) = 1 \mid x_t, \pi_{\theta_t}). \tag{3}$$

In the RLVR setting, p_t represents the expected reward under policy π_{θ_t} given x_t, while $\hat{p}_t$ can be regarded as an empirical estimator of p_t obtained from a finite group of sampled responses. This motivates the following definition.

Definition 2 (Expected Advantage)**.** *Given a prompt $x_t \sim D$, let $y_{t,i} \sim \pi_{\theta_t}(\cdot \mid x_t)$ be a sampled response with corresponding reward $r_{t,i}$. The expected advantage is defined as:*

$$A_{t,i} = r_{t,i} - p_t. \tag{4}$$

Thus, in the RLVR setting, $A_{t,i}$ represents the *expected* advantage of response $y_{t,i}$ under policy π_{θ_t} given x_t, while $\hat{A}_{t,i}$ can be regarded as an empirical estimator of $A_{t,i}$ obtained from a finite group of sampled responses. Most group-relative RL algorithms rely on $\hat{A}_{t,i}$ for policy updates, differing primarily in how $\hat{A}_{t,i}$ is processed or transformed within their respective optimization objectives.

2.2 Fundamental Discovery

Next, we present a formal formulation of the problem. Given a prompt $x_t \sim D$, let p_t denote the expected reward of policy π_{θ_t} on x_t. We then sample G responses independently from $\pi_{\theta_t}(\cdot \mid x_t)$. In RLVR, rewards are often binary, especially in mathematical and formal reasoning tasks where verifiers return pass/fail signals. Under this widely adopted setting, it is natural to model the reward associated with each response as a Bernoulli random variable:

$$r_{t,i} \sim \text{Bernoulli}(p_t), \quad \forall i \in [G]. \tag{5}$$

Let $R = \sum_{i=1}^{G} r_{t,i}$ denote the total reward within the group. The empirical group baseline is given by $\hat{p}_t = R/G$.

Definition 3 (Prompt Difficulty)**.** *Given a prompt x_t, a policy π_{θ_t}, and a classification threshold $\tau_{\text{cls}} \in [0, 1/2)$, we define:*

- *x_t is a **hard prompt** if $p_t < 1/2 - \tau_{\text{cls}}$;*
- *x_t is a **moderate prompt** if $1/2 - \tau_{\text{cls}} \le p_t \le 1/2 + \tau_{\text{cls}}$;*
- *x_t is an **easy prompt** if $p_t > 1/2 + \tau_{\text{cls}}$.*

The actual difficulty gap is $d_t := |p_t - 1/2|$. The threshold τ_{cls} determines class membership, whereas d_t measures the prompt's distance from $p_t = 1/2$. Unless otherwise stated, the subsequent analysis uses $\tau_{\text{cls}} = 0$.

In group-based policy optimization, the group-relative advantage estimator satisfies $\hat{A}_{t,i} = 0$ for all $i \in [G]$ when either $R = 0$ or $R = G$, resulting in zero gradients and hence no parameter updates. In practice, such degenerate groups do not contribute to learning and are either explicitly discarded or implicitly ignored by GRPO-style algorithms. Accordingly, our analysis focuses on the effective update regime, namely groups for which at least one response receives a non-zero advantage. This corresponds to the non-degenerate event:

$$\mathcal{S} \coloneqq \{1 \le R \le G-1\}. \tag{6}$$

Conditioning on $\mathcal{S}$ isolates non-degenerate groups for the analysis of advantage estimation error. The resulting statements concern this conditional distribution; they do not assert that the unconditionally sampled group baseline is biased. Next, we present the main results.

Theorem 1 (Conditional Mean Advantage Error)**.** *Fix a prompt x_t and the current policy π_{θ_t}. Suppose $G \ge 2$ and $r_{t,1}, \dots, r_{t,G} \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_t)$ with $p_t \in (0,1)$. Condition on $\mathcal{S} = \{1 \le R \le G-1\}$. For every $i \in [G]$, the conditional mean estimation error satisfies*

$$\begin{cases} \mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}] < 0, & p_t < 1/2, \\ \mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}] = 0, & p_t = 1/2, \\ \mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}] > 0, & p_t > 1/2. \end{cases}$$

Theorem 1 characterizes the conditional mean estimation error within non-degenerate groups; its closed-form expression and proof are provided in Appendix F.1. Unconditionally, $\mathbb{E}[\hat{p}_t] = p_t$ and $\mathbb{E}[\hat{A}_{t,i} - A_{t,i}] = 0$. Conditioning on $\mathcal{S}$ shifts the expected group baseline upward when $p_t < 1/2$ and downward when $p_t > 1/2$, producing negative and positive conditional mean errors, respectively. The zero conditional mean of the centered advantage does not imply a zero conditional mean estimation error. The latter vanishes only at $p_t = 1/2$; this does not imply that the error also vanishes after conditioning on the response reward.

Corollary 1 (Reward-Conditioned Advantages)**.** *Under the assumptions of Theorem 1, if $p_t < 1/2$, then*

$$\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i} = 1, \mathcal{S}] < 1 - p_t.$$

If $p_t > 1/2$, then

$$\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i} = 0, \mathcal{S}] > -p_t.$$

Thus, in conditional expectation, a correct response to a hard prompt receives an underestimated positive advantage, while an incorrect response to an easy prompt receives an insufficiently negative advantage. The comparison targets $1 - p_t$ and $-p_t$ are fixed once the response reward is specified. The closed-form expressions and proof are provided in Appendix F.2.

Dependence on the difficulty gap. For $G \ge 2$ and $d \in [0, 1/2)$, define

$$B_G(d) := \frac{(\frac{1}{2} - d)(\frac{1}{2} + d)^G - (\frac{1}{2} + d)(\frac{1}{2} - d)^G}{1 - (\frac{1}{2} + d)^G - (\frac{1}{2} - d)^G}. \tag{7}$$

By symmetry, the absolute conditional mean error satisfies

$$\left| \mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}] \right| = B_G(d_t), \qquad B_G(0) = 0.$$

For each fixed $G \ge 2$, as $d \to 0$,

$$B_G(d) = \frac{G-1}{2^{G-1} - 1}\, d + O_G(d^3). \tag{8}$$

The leading coefficient is positive and decreases with G. Hence the conditional mean bias grows locally with the actual distance from $p_t = 1/2$, while increasing the group size reduces this local sensitivity. Setting $\tau_{\text{cls}} = 0$ changes the classification boundary; the conditional mean error vanishes when $d_t = 0$, equivalently $p_t = 1/2$. These statements concern the mean-centered error conditioned on $\mathcal{S}$, before standard-deviation normalization. The derivation is provided in Appendix F.3.

However, the expectation-level result in Theorem 1 alone is insufficient to characterize the *probability* of overestimation or underestimation of $\hat{A}_{t,i}$. We provide the following probabilistic result.

Theorem 2. *Under the condition of Theorem 1, suppose x_t is a **hard** prompt ($p_t < 0.5$). Then, for any $\epsilon \in (0, \mathbb{E}[\hat{p}_t \mid \mathcal{S}] - p_t)$, we have:*

$$\mathbb{P}\left(A_{t,i} - \hat{A}_{t,i} > \epsilon \mid \mathcal{S}\right) = \frac{\sum_{k=\lfloor G(p_t+\epsilon)\rfloor+1}^{G-1} \binom{G}{k} p_t^k (1-p_t)^{G-k}}{1-(1-p_t)^G - p_t^G}.$$

*Simlarily, suppose x_t is an **easy** prompt ($p_t > 0.5$). Then, for any $\epsilon \in (0, p_t - \mathbb{E}[\hat{p}_t \mid \mathcal{S}])$, we have:*

$$\mathbb{P}\left(\hat{A}_{t,i} - A_{t,i} > \epsilon \mid \mathcal{S}\right) = \frac{\sum_{k=1}^{\lceil G(p_t-\epsilon)\rceil-1} \binom{G}{k} p_t^k (1-p_t)^{G-k}}{1-(1-p_t)^G - p_t^G}.$$

Theorem 2 provides a distribution-level characterization of how likely group-relative advantage estimation is to *underestimate* or *overestimate* the true advantage, depending on prompt difficulty. In contrast to expectation-level results, this theorem quantifies the exact probability mass of large estimation errors under finite group sizes. The conditional concentration bound and mean-bias rates are provided in Appendix F.5.

It is well known that generating multiple rollouts per prompt is computationally expensive in practice. Consequently, existing RLVR methods typically sample only a small number of responses (e.g., $G = 8$) for each prompt x_t to estimate $\hat{p}_t$ [49, 21, 33] . Motivated by this practical constraint, we derive the following corollaries based on Theorem 2, which explicitly characterize the estimation behavior under small group sizes.

Corollary 2. *Under the condition of Theorem 2, suppose the group size satisfies $2 \le G \le 8$, and assume that p_t is uniformly distributed over $[0, 1]$. Then, for any $i \in [G]$, the following inequalities hold:*

$$\begin{aligned}
&\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < 0.5\Big) > 0.63, && \mathbb{P}\Big(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > 0.5\Big) > 0.63,\\
&\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < 0.25\Big) > 0.78, && \mathbb{P}\Big(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > 0.75\Big) > 0.78,\\
&\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < 0.125\Big) = 1.00, && \mathbb{P}\Big(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > 0.875\Big) = 1.00.
\end{aligned}$$

Corollary 2 shows that, with high probability, the group-relative advantage estimator $\hat{A}_{t,i}$ *underestimates* the true advantage $A_{t,i}$ for hard prompts and *overestimates* $A_{t,i}$ for easy prompts, under the practical set of G. These probability statements concern increasingly extreme ranges of p_t. The dependence of the conditional mean bias on the actual difficulty gap d_t is characterized by Equations (7) and (8).

Corollary 3. *Under the condition of Corollary 2, suppose $G \ge 6$. The following inequalities hold:*

$$\mathbb{P}\left(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < \frac{2}{G}\right) > 0.78, \quad \mathbb{P}\left(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > \frac{G-2}{G}\right) > 0.78. \tag{9}$$

Corollary 4. *Under the condition of Theorem 2, suppose $G \ge 2$. Then, for any $i \in [G]$, the following inequalities hold surely:*

$$\hat{A}_{t,i} < A_{t,i} \text{ as } p_t < \tfrac{1}{G}, \qquad \hat{A}_{t,i} > A_{t,i} \text{ as } p_t > \tfrac{G-1}{G}. \tag{10}$$

Corollary 4 shows that the group-based advantage estimator $\hat{A}_{t,i}$ necessarily *underestimates* the true advantage $A_{t,i}$ for extremely difficult prompts ($p_t < 1/G$), and *overestimates* $A_{t,i}$ for extremely easy prompts ($p_t > (G-1)/G$). Detailed derivation process is presented in Appendix F.

2.3 Auxiliary Results

Theorem 3 (Conditional Bias under Bounded Continuous Rewards)**.** *At training step t, let $G \ge 2$ and draw i.i.d. bounded rewards $r_{t,1}, \ldots, r_{t,G} \sim \mathcal{D}(p_t)$ with $r_{t,i} \in [0, 1]$ and $\mathbb{E}[r_{t,i}] = p_t$. For*

a threshold $\sigma \in [0,1]$, consider the update event $S_\sigma = \{\exists\, i \neq j : |r_{t,i} - r_{t,j}| > \sigma\}$. Assume $\mathbb{P}(S_\sigma) > 0$. Then, conditioned on performing an update, the advantage estimation bias satisfies:

$$\mathbb{E}\Big[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma\Big] = p_t - \mathbb{E}[\hat{p}_t \mid S_\sigma]\,, \qquad \forall i \in \{1,\ldots,G\}. \tag{11}$$

Consequently, whenever the update event S_σ changes the conditional expectation of the group mean, i.e.,

$$\mathbb{E}\Big[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma\Big] \neq 0, \tag{12}$$

the group-relative advantage estimator is conditionally biased, even under bounded continuous rewards. Moreover, the bias direction is deterministic in the following extreme regimes:

$$\begin{cases} \mathbb{E}\Big[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma\Big] < 0, & \text{if } p_t \leq \frac{\sigma}{G}, \\ \mathbb{E}\Big[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma\Big] > 0, & \text{if } p_t \geq 1 - \frac{\sigma}{G}. \end{cases}$$

Thus, the conditional bias is not an artifact of binary rewards; it persists under bounded continuous rewards whenever the update-selection event distorts the conditional group mean.

While the binary reward assumption covers many standard RLVR setups, we recognize that real-world reward signals can be more general. To address this, we extend our analysis to **continuous bounded reward distributions**. Theorem 3 suggests that the core bias phenomenon is not an artifact of the Bernoulli reward assumption but is prevalent across a broader class of bounded reward models. The more detailed analysis is presented in Appendix F.9.

Theorem 4 (Bias under Standard-Deviation Normalization)**.** *Under the retention condition $\mathcal{S} = \{1 \leq R \leq G-1\}$, for a correct response $r_{t,i} = 1$, the standard-deviation-normalized group-relative advantage satisfies:*

$$\mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \;\middle|\; r_{t,i} = 1, \mathcal{S}\right] \neq \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}. \tag{13}$$

Moreover, the bias direction is deterministic in the extreme regimes:

$$\begin{cases} \mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \mid r_{t,i} = 1, \mathcal{S}\right] < \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}, & \text{if } p_t < \frac{1}{G}, \\ \mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \mid r_{t,i} = 1, \mathcal{S}\right] > \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}, & \text{if } p_t > \frac{G-1}{G}. \end{cases}$$

Thus, standard-deviation normalization rescales the group-relative advantage but does not eliminate its intrinsic conditioning bias.

We further examine whether the bias can be removed by the standard-deviation normalization commonly used in group-relative policy optimization. Since this normalization divides the empirical advantage by the within-group reward standard deviation, one might expect it to correct the scale distortion caused by varying prompt difficulty. However, the normalization is itself computed from the same retained group samples and is therefore affected by the conditioning event. As a result, the estimator remains biased even after normalization. Normalization mainly stabilizes the scale of policy updates, whereas HA-DW targets the difficulty-dependent distortion induced by conditional group selection. The above theorem shows that this bias persists for correct responses and that its direction is deterministic in the extreme difficulty regimes. And the detailed derivation with standard-deviation normalization is provided in Appendix F.8.

3 Proposed Solution

Since the group-based advantage estimator is biased, we propose an algorithm to adjust the advantage estimation accordingly. The proposed approach consists of two key components: **Evolving Difficulty Anchor** and **History Aware Adaptive Difficulty Weighting**.

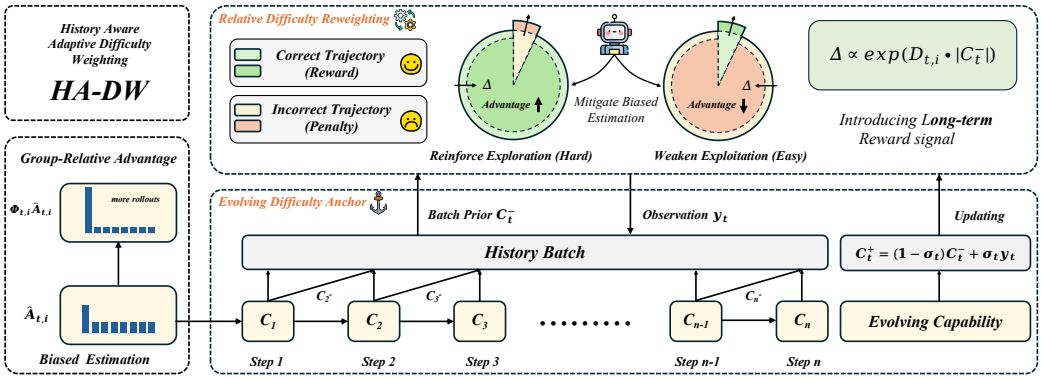

Figure 2: HA-DW consists of two collaborative phases. In the first phase, an evolving difficulty anchor incorporates cross-batch historical information by propagating the model's prior through a history buffer, capturing long-term reward trends. In the second phase, prompt weights are adaptively adjusted based on their estimated difficulty under the model's evolving state, compensating for biased advantage estimates.

3.1 Evolving Difficulty Anchor

To track the evolving model state across batches, we propose the cross-batch difficulty anchor framework that integrates long-term reward trends and historical information. Let B_t denote the total number of responses in batch t. Model updates are guided by observations of the current batch's prompt accuracy b_t together with historical information, defined as:

$$b_t = \frac{K_t}{B_t}, \qquad K_t = \sum_{i=1}^{B_t} r_{t,i}. \tag{14}$$

We treat the model's solving capability C_t as a latent belief state. At training step t, the observation b_t is used to update the prior belief C_t^- to the posterior belief C_t^+ via a Kalman-style update [2, 51]:

$$C_t^+ = (1 - \sigma_t)\, C_t^- + \sigma_t\, b_t, \quad \eta_t \in [0, 1]. \tag{15}$$

The forgetting factor σ_t controls the influence of historical information and is dynamically modulated by model stability. Specifically, we compute the average belief over the previous m batches as:

$$\bar{C}_t = \frac{1}{m} \sum_{j=1}^{m} C_{t-j}, \tag{16}$$

and define the σ_t as corresponding standard deviation: $\sigma_t = \sqrt{\frac{1}{m}\sum_{j=1}^{m} \left(C_{t-j} - \bar{C}_t\right)^2}$. Between consecutive steps, the posterior belief C_t^+ serves as the prior belief for the next batch: $C_t^+ \to C_{t+1}^-$. Overall, C_t enables the model to aggregate information across historical batches via belief updates and to condition its training strategy on this evolving belief.

3.2 History Aware Adaptive Difficulty Weighting (HA-DW)

To rectify the inherent bias in group-based advantage estimation, we introduce HA-DW, which dynamically adjusts advantage weights based on the model's evolving state while incorporating long-term reward signals. Coupled with the evolving difficulty anchor, we define the history-based prompt difficulty as:

$$\mathrm{diff}_t^{\mathrm{his}} = \hat{p}_t - C_t^-, \tag{17}$$

where $\mathrm{diff}_t^{\mathrm{his}}$ captures both the magnitude and direction of a prompt's difficulty relative to the current model belief. To determine the *direction* of adjustment, we use the evolving difficulty anchor as a reference and define:

$$D_{t,i} = -\operatorname{sgn}\left(\hat{A}_{t,i}\right) \cdot \operatorname{sgn}\left(\mathrm{diff}_t^{\mathrm{his}}\right), \tag{18}$$

where $\mathrm{sgn}(\cdot)$ denotes the sign function. Next, we quantify the *magnitude* of adjustment using the absolute history-based difficulty: $M_t = \left|\mathrm{diff}_t^{\mathrm{his}}\right|$. Here, M_t measures the extent to which the prompt deviates from the model's current capability. We then define the history-aware reweighting factor as:

$$\Phi_{t,i} = \lambda_{\mathrm{scale}} \cdot \exp(D_{t,i} \cdot M_t)\,, \tag{19}$$

where λ_{scale} is a scaling constant, and the exponential form ensures smooth and multiplicative adjustment of advantage weights. The resulting HA-DW objective is:

$$L_{\text{HA-DW}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\psi\Big(\frac{\pi_\theta(y_{t,i} \mid x_t)}{\pi_{\theta_{\mathrm{old}}}(y_{t,i} \mid x_t)}\Big) \cdot \phi\big(\hat{A}_{t,i}\big) \cdot \Phi_{t,i}, \tag{20}$$

where $\psi(\cdot)$ and $\phi(\cdot)$ follow specific definitions in group-relative RL algorithms.

Intuitively, $\Phi_{t,i}$ adjusts the estimated advantage for each prompt, thereby correcting systematic bias identified in our analysis. HA-DW mitigates distortion within retained non-degenerate groups, rather than recovering gradients from degenerate groups. And it can be seamlessly integrated as a plug-and-play module into group-relative methods, improving reasoning performance under fixed rollouts while effectively mitigating biased advantage estimation. Detailed instantiations for GRPO and related algorithms are provided in the Appendix D

4 Theoretical Analysis

In this section, we provide a theoretical analysis of the effectiveness of the proposed adjustment strategy. We begin by analyzing how reweighting the empirical baseline $\hat{p}_t$ affects the expected bias.

Lemma 1 (Baseline Rectification)**.** *Given a prompt $x_t \sim D$ and the policy π_{θ_t}, let $\tilde{p}_t = c \cdot \hat{p}_t$ be the rectified group baseline. Assume $p_t \in [\eta_{\mathrm{margin}},\, 1-\eta_{\mathrm{margin}}]$ for some $\eta_{\mathrm{margin}} \in (0, 1/2]$. Given any $\delta \in (0,1)$, we can define that:*

$$\epsilon_\delta := \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big(1-(1-\eta_{\mathrm{margin}})^G - \eta_{\mathrm{margin}}^G\big)}\right)}. \tag{21}$$

Let

$$I_t := \big[\hat{p}_t - \epsilon_\delta,\ \hat{p}_t + \epsilon_\delta\big] \cap [\eta_{\mathrm{margin}}, 1-\eta_{\mathrm{margin}}], \quad A(p_t) := 1-(1-p_t)^G - p_t^G.$$

Fix any $\epsilon > 0$, we define:

$$c_{\mathrm{low}} := \sup_{p_t \in I_t} \frac{(p_t-\epsilon)\,A(p_t)}{p_t(1-p_t^{G-1})}, \qquad c_{\mathrm{high}} := \inf_{p_t \in I_t} \frac{(p_t+\epsilon)\,A(p_t)}{p_t(1-p_t^{G-1})}. \tag{22}$$

Then, with probability at least $1-\delta$ conditional on $\mathcal{S}$, for any choice $c \in (c_{\mathrm{low}},\ c_{\mathrm{high}})$, we can derive that:

$$\mathbb{E}[\tilde{p}_t \mid \mathcal{S}] \in (p_t - \epsilon,\ p_t + \epsilon). \tag{23}$$

Specifically, we consider adjusting the empirical group baseline using a reweighting factor c. From the perspective of the expected estimation bias, Lemma 1 that an appropriate choice of c can effectively reduce estimation bias. Detailed derivations are provided in Appendix F.7. Next, we now proceed to present the main theoretical result.

Theorem 5. *Under the condition of Lemma 1, suppose there exists a scaling factor λ_{scale} in Equation* (19) *such that:*

$$\lambda_{\mathrm{scale}} \in \left(\frac{1+\frac{(1-c_{\mathrm{high}})\hat{p}_t}{1-\hat{p}_t}}{\exp(D_{t,i}M_t)},\ \frac{1+\frac{(1-c_{\mathrm{low}})\hat{p}_t}{1-\hat{p}_t}}{\exp(D_{t,i}M_t)}\right) \cup \left(\frac{c_{\mathrm{low}}}{\exp(D_{t,i}M_t)},\ \frac{c_{\mathrm{high}}}{\exp(D_{t,i}M_t)}\right).$$

Then, HA-DW algorithm provably mitigates the bias of group-relative advantage:

$$\Big|\mathbb{E}\big[\hat{A}_{t,i}\cdot\Phi_{t,i} \mid \mathcal{S}\big] - A_{t,i}\Big| < \Big|\mathbb{E}\big[\hat{A}_{t,i} \mid \mathcal{S}\big] - A_{t,i}\Big|.$$

Theorem 5 shows that, with an appropriate choice of the scaling parameter λ_{scale}, the HA-DW adjustment yields advantage estimates that are closer to the true advantage $A_{t,i}$ in expectation. This theoretical result provides principled guidance for selecting λ_{scale} in practice.

Algorithm	Math						Code		
	MATH500	AIME25	AMC23	Minerva	OlympiadBench	Avg.	HumanEval	MBPP	Avg.
Qwen-3-4B-Base									
GRPO	75.4	19.6	60.3	33.8	43.5	46.5	70.1	58.2	64.2
↪ + HA-DW	78.0	20.4	63.4	36.8	44.7	**48.7**	73.2	60.4	**66.8**
GSPO	75.8	20.0	62.2	35.3	42.3	47.1	69.5	58.6	64.1
↪ + HA-DW	77.6	19.6	68.6	37.1	43.2	**49.2**	72.0	59.4	**65.7**
DAPO	76.8	18.3	60.0	35.7	43.2	46.8	71.3	60.0	65.6
↪ + HA-DW	78.6	21.3	65.0	37.5	45.3	**49.5**	73.8	62.6	**68.2**
Qwen-3-8B-Base									
GRPO	78.8	20.4	64.2	38.2	46.4	49.6	78.7	65.6	72.2
↪ + HA-DW	80.0	22.9	72.8	39.7	47.1	**52.5**	81.1	68.0	**74.6**
GSPO	78.6	21.7	67.0	37.9	45.9	50.2	78.0	64.8	71.4
↪ + HA-DW	80.2	22.1	66.5	41.9	47.6	**51.7**	80.5	67.4	**74.0**
DAPO	79.2	20.4	67.5	39.3	47.2	50.7	79.9	66.0	72.9
↪ + HA-DW	82.8	23.3	70.0	40.8	50.0	**53.4**	84.1	68.4	**76.3**
LLaMA-3.2-3B-Instruct									
GRPO	51.4	2.7	31.7	22.8	19.9	25.7	48.8	43.2	46.0
↪ + HA-DW	53.2	3.3	35.0	23.9	20.1	**27.1**	51.2	45.6	**48.4**
GSPO	48.6	1.9	30.9	23.2	19.8	24.9	49.4	43.6	46.5
↪ + HA-DW	50.4	2.3	32.7	22.4	21.0	**25.8**	50.6	46.0	**48.3**
DAPO	52.4	2.5	35.0	22.4	20.2	26.5	50.0	44.6	47.3
↪ + HA-DW	53.2	3.1	37.5	24.6	22.3	**28.1**	53.0	47.2	**50.1**

Table 1: Overall results across models (Qwen, LLaMA) and different group-relative RL algorithms (GRPO, GSPO, DAPO). We report the performance of different base RL algorithms, and the corresponding accuracy when applied HA-DW for each model scale and family.

Threshold	MATH500	AIME25	AMC23	Minerva	OlympiadBench	AVG
Base	75.4	19.6	60.3	33.8	43.5	46.5
0.4 (fixed)	77.0	18.5	63.1	**37.5**	44.3	48.1
0.5 (fixed)	76.6	20.0	62.7	35.7	44.0	47.8
0.6 (fixed)	76.8	**21.3**	61.1	36.4	44.3	48.0
C_t	**78.0**	20.4	**63.4**	36.8	**44.7**	**48.7**

Table 2: Ablation on the effectiveness of dynamic threshold for RL training using Qwen3-4B-Base. C_t denotes the dynamic threshold.

5 Experiments

Setups. We conduct our experiments on Qwen3-4B-Base, Qwen3-8B-Base [38] and LLaMA-3.2-3B-Instruct on five common-used mathematical reasoning benchmarks and two code generation benchmarks. We apply our proposed method on top of several representative group-relative reinforcement learning algorithms: GRPO, GSPO, and DAPO. We compare the performance of group-relative algorithms applying HA-DW to original ones, verifying the effectiveness and scalability of our method. We conduct RL training within the VeRL framework [34] on a single node with 8 $\times$ NVIDIA A100 GPUs. More implementation details in Appendix E.

5.1 Main Results

Our main results are presented in Table 1. Notably, group-based RL algorithms (GRPO, GSPO and DAPO) equipped with HA-DW outperforms original methods across seven benchmarks. We observed clear and consistent improvements across benchmarks on different models of different scales and family. Overall, the results underscore that HA-DW compensates for advantage estimation bias via dynamic reweighting to fully leverage these overshadowed critical prompts, thereby unlocking the potential performance gains in RL.

To validate our method's effectiveness in extending model capabilities, we divided the MATH500 dataset into three difficulty levels: Easy (Level 1), Mid (Levels 2-3), and Hard (Levels 4-5). We

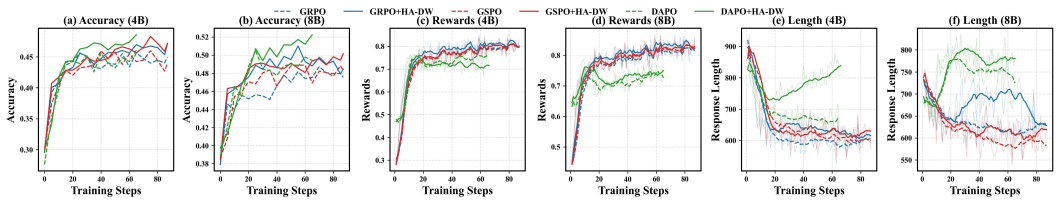

Figure 3: Comparison of training dynamics under different training strategies. Average accuracy across five benchmarks, training reward and response length of Qwen3-4B-Base and Qwen3-8B-Base on different training methods.

evaluated Qwen3-4B-Base trained with GRPO and GRPO+HA-DW, as shown in Figure 1(c). The performance on Easy and Mid levels was comparable for both methods, but GRPO+HA-DW outperformed GRPO by 3.4% on Hard prompts. This improvement is due to our history-based dynamic reweighting strategy, which enhances exploration on hard prompts while reducing unnecessary exploitation on easy ones. Simultaneously, it substantiates the existence of bias estimation indirectly.

Training Dynamics. Figure 3 demonstrates the temporal dynamics of average accuracy across seven benchmarks, training rewards throughout the training process, and response lengths of Qwen3-4B-Base and Qwen3-8B-Base. RL algorithms applied HA-DW converge to a higher performance plateau in accuracy and acquired higher reward compared to the original RL algorithms, suggesting that the application of HA-DW boosts the exploration of hard prompts and weakens the exploitation of easy ones by mitigating the biased advantage estimation. All methods were trained until reaching stable performance plateaus; Figure 3 shows only a truncated portion of the full trajectories, while the displayed range already contains the best checkpoint of each method. The reward curves have also clearly plateaued, indicating that HA-DW's gains persist near convergence rather than merely reflecting faster early-stage optimization. In addition, our method encourages longer reasoning, greatly improving its reasoning abilities [17, 5]. HA-DW is capable of incentivizing the model to produce more sophisticated reasoning chain of thoughts to tackle more challenging tasks.

Ablation Study on C_t. We evaluate the effectiveness of the dynamic threshold C_t by comparing it with a fixed thresholds across five mathematical benchmarks, as shown in Table 2. Experiments on Qwen3-4B-Base with GRPO-based training show that dynamic adjustment achieves the best performance. Removing C_t degrades performance, while a fixed threshold still improves over the baseline by partially mitigating biased estimation. By incorporating cross-batch information, C_t captures long-term reward signals and further enhances RL performance.

Supplementary Experiments (Appendix G) Due to space limitations, we include the following additional experiments in Appendix G: (1) empirical verification of advantage estimation bias, (2) an ablation study on the group size G (Appendix G.5), (3) an ablation study on the scaling parameter λ_{scale}, (4) experiments extended to instruction following and tool use. (5) effect on easy prompts, (6) an ablation study on rollout token budget (7) robustness to reward noise, and (8) significance analysis.

6 Related Work

GRPO and GRPO Variants. Following the success of Deepseek-R1 [5], GRPO has attracted widespread attention. To achieve better performance, numerous GRPO-based variants have been proposed. However, these variants adopt static prompt difficulty and suffer from insufficient exploration of model's capability. More related work are placed in Appendix C.

7 Conclusion

Our work characterizes difficulty-dependent conditional estimation bias in group-relative advantages within non-degenerate groups. To address this issue, we propose *HA-DW*, which dynamically adjusts advantage weights based on the model's evolving state. Extensive experiments demonstrate that *HA-DW* effectively improves reasoning performance by mitigating biased advantage estimation.

Acknowledgments and Disclosure of Funding

This research was supported by China NSFC (No. 62506024). The authors declare no competing interests.

References

[1] Yikun Ban, Fengkai Yang, Fangzheng Chen, Yibo Wang, Zhijun Chen, Zhongyi Li, Zixuan Huang, Xiaoyuan Zhang, Gongxun Li, Zehao Chen, Huaiyang Wang, Xiaodong Lu, Yaocheng Yang, Pengcheng Wei, Wan Tian, Zherui Chen, Zhixia Zhang, Hongyan Xie, Lingyu Lv, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Tianxiang Ai, Ruijie Wang, Haoyi Zhou, Chaochao Lu, Wanxiang Che, Fuzhen Zhuang, Ning Ding, Qianqian Xu, Deqing Wang, Yaodong Yang, and Jianxin Li. Epistemic exploration toward artificial general intelligence. Zenodo, 2026. URL https://doi.org/10.5281/zenodo.20522905.

[2] Stefano Battilotti, Alessandro Borri, Filippo Cacace, Massimiliano D'Angelo, and Alfredo Germani. A consensus kalman filter on L2 spaces. *Autom.*, 183:112530, 2026. doi: 10.1016/J.AUTOMATICA.2025.112530. URL https://doi.org/10.1016/j.automatica.2025.112530.

[3] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press, 2013. ISBN 978-0-19-953525-5. doi: 10.1093/ACPROF:OSO/9780199535255.001.0001. URL https://doi.org/10.1093/acprof:oso/9780199535255.001.0001.

[4] Weizhe Chen, Sven Koenig, and Bistra Dilkina. LSPO: length-aware dynamic sampling for policy optimization in LLM reasoning. *CoRR*, abs/2510.01459, 2025. doi: 10.48550/ARXIV.2510.01459. URL https://doi.org/10.48550/arXiv.2510.01459.

[5] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948.

[6] Yuyang Ding, Chi Zhang, Juntao Li, Haibin Lin, Xin Liu, and Min Zhang. FAPO: flawed-aware policy optimization for efficient and reliable reasoning. *CoRR*, abs/2510.22543, 2025. doi: 10.48550/ARXIV.2510.22543. URL https://doi.org/10.48550/arXiv.2510.22543.

[7] Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1406–1415. PMLR, 2018. URL http://proceedings.mlr.press/v80/espeholt18a.html.

[8] Lang Feng, Zhenghai Xue, Tingcong Liu, and Bo An. Group-in-group policy optimization for llm agent training. *arXiv preprint arXiv:2505.10978*, 2025.

[9] Pietro Di Gianantonio and Abbas Edalat. A domain-theoretic framework for conditional probability and bayesian updating in programming. *CoRR*, abs/2502.00949, 2025. doi: 10.48550/ARXIV.2502.00949. URL https://doi.org/10.48550/arXiv.2502.00949.

[10] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*. Springer Series in Statistics. Springer, 2009. ISBN 9780387848570. doi: 10.1007/978-0-387-84858-7. URL https://doi.org/10.1007/978-0-387-84858-7.

[11] 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. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850.

Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.211. URL https://doi.org/10.18653/v1/2024.acl-long.211.

[12] Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss Cook, and Jingrui He. Llm-forest: Ensemble learning of llms with graph-augmented prompts for data imputation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6921–6936, 2025.

[13] 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. In Joaquin Vanschoren and Sai-Kit Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

[14] Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, et al. Adaptive batch-wise sample scheduling for direct preference optimization. *Advances in Neural Information Processing Systems*, 38:38490–38522, 2026.

[15] Zixuan Huang, Xin Xia, Yuxi Ren, Jianbin Zheng, Xuanda Wang, Zhixia Zhang, Hongyan Xie, Songshi Liang, Zehao Chen, Xuefeng Xiao, et al. Does your reasoning model implicitly know when to stop thinking? *arXiv preprint arXiv:2602.08354*, 2026.

[16] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 652–661. JMLR.org, 2016. URL http://proceedings.mlr.press/v48/jiang16.html.

[17] Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. The impact of reasoning step length on large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 1830–1842. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.108. URL https://doi.org/10.18653/v1/2024.findings-acl.108.

[18] Zhongyi Li, Wan Tian, Jingyu Chen, Kangyao Huang, Huiming Zhang, Hui Yang, Tao Ren, Jinyang Jiang, Yijie Peng, Yikun Ban, et al. Adaptive robust estimator for multi-agent reinforcement learning. *arXiv preprint arXiv:2603.21574*, 2026.

[19] Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026.

[20] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

[21] Bingshuai Liu, Ante Wang, Zijun Min, Liang Yao, Haibo Zhang, Yang Liu, Anxiang Zeng, and Jinsong Su. Spec-rl: Accelerating on-policy reinforcement learning with speculative decoding, 2025. Rollouts generated using vLLM (rollout N=8).

[22] Dengcan Liu, Fengkai Yang, Xiaohan Wang, Shurui Yan, Jiajun Chai, Jiahao Li, Yikun Ban, Zhendong Mao, Wei Lin, and Guojun Yin. Cdrrm: Contrast-driven rubric generation for reliable and interpretable reward modeling. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 3152–3163, 2026.

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

[24] Xiaodong Lu, Xiaohan Wang, Jiajun Chai, Guojun Yin, Wei Lin, Zhijun Chen, Yu Luo, Fuzhen Zhuang, Yikun Ban, and Deqing Wang. Contextual rollout bandits for reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2602.08499*, 2026.

[25] Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 1046–1054, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/c3992e9a68c5ae12bd18488bc579b30d-Abstract.html.

[26] Kevin P. Murphy. *Machine learning - a probabilistic perspective*. Adaptive computation and machine learning series. MIT Press, 2012. ISBN 0262018020.

[27] Daniil Plyusov, Alexey Gorbatovski, Boris Shaposhnikov, Viacheslav Sinii, Alexey Malakhov, and Daniil Gavrilov. F-grpo: Don't let your policy learn the obvious and forget the rare. *arXiv preprint arXiv:2602.06717*, 2026.

[28] Zikun Qu, Min Zhang, Mingze Kong, Xiang Li, Zhiwei Shang, Zhiyong Wang, Yikun Ban, Shuang Qiu, Yao Shu, and Zhongxiang Dai. T-pop: Test-time personalization with online preference feedback. *arXiv preprint arXiv:2509.24696*, 2025.

[29] Zikun Qu, Min Zhang, Mingze Kong, Zhiwei Shang, Zhengyu Chen, Yikun Ban, Shuang Qiu, and Zhongxiang Dai. Spot: Sparse probing and outcome calibration for on-policy distillation. *arXiv preprint arXiv:2608.04419*, 2026.

[30] Robert J Serfling. Some elementary results on poisson approximation in a sequence of bernoulli trials. *Siam review*, 20(3):567–579, 1978.

[31] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

[32] Guobin Shen, Chenxiao Zhao, Xiang Cheng, Lei Huang, and Xing Yu. Vespo: Variational sequence-level soft policy optimization for stable off-policy llm training. *arXiv preprint arXiv:2602.10693*, 2026.

[33] Yuanzhe Shen, Zisu Huang, Zhengkang Guo, Yide Liu, Guanxu Chen, Ruicheng Yin, Xiaoqing Zheng, and Xuanjing Huang. Intentionreasoner: Facilitating adaptive LLM safeguards through intent reasoning and selective query refinement. *CoRR*, abs/2508.20151, 2025. doi: 10.48550/ARXIV.2508.20151. URL https://doi.org/10.48550/arXiv.2508.20151.

[34] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

[35] Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. KTAE: A model-free algorithm to key-tokens advantage estimation in mathematical reasoning. *CoRR*, abs/2505.16826, 2025. doi: 10.48550/ARXIV.2505.16826. URL https://doi.org/10.48550/arXiv.2505.16826.

[36] Zelin Tan, Hejia Geng, Mulei Zhang, Xiaohang Yu, Guancheng Wan, Yifan Zhou, Qiang He, Xiangyuan Xue, Heng Zhou, Yutao Fan, Zhongzhi Li, Zaibin Zhang, Guibin Zhang, Chen Zhang, Zhenfei Yin, and Lei Bai. Scaling behaviors of LLM reinforcement learning post-training: An empirical study in mathematical reasoning. *CoRR*, abs/2509.25300, 2025. doi: 10.48550/ARXIV.2509.25300. URL https://doi.org/10.48550/arXiv.2509.25300.

[37] Zichang Tan, Ajian Liu, Jun Wan, Hao Li, Zhen Lei, Guodong Guo, and Stan Z. Li. Cross-batch hard example mining with pseudo large batch for ID vs. spot face recognition. *IEEE Trans. Image Process.*, 31:3224–3235, 2022. doi: 10.1109/TIP.2021.3137005. URL https://doi.org/10.1109/TIP.2021.3137005.

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

[39] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=rkxoh24FPH.

[40] Huaiyang Wang, Xiaojie Li, Xiaohan Wang, Zhixia Zhang, Xiaodong Lu, Zixuan Huang, Jiajun Chai, Guojun Yin, Deqing Wang, Haoyi Zhou, et al. Policy improvement reinforcement learning. *arXiv preprint arXiv:2604.00860*, 2026.

[41] Jinpeng Wang, Jieming Zhu, and Xiuqiang He. Cross-batch negative sampling for training two-tower recommenders. In Fernando Diaz, Chirag Shah, Torsten Suel, Pablo Castells, Rosie Jones, and Tetsuya Sakai, editors, *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 1632–1636. ACM, 2021. doi: 10.1145/3404835.3463032. URL https://doi.org/10.1145/3404835.3463032.

[42] Xun Wang, Haozhi Zhang, Weilin Huang, and Matthew R. Scott. Cross-batch memory for embedding learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 6387–6396. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00642. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_Cross-Batch_Memory_for_Embedding_Learning_CVPR_2020_paper.html.

[43] Xinran Wu, Kun Yue, Huashuai Liu, and Liang Duan. Learning conditional probability distributions for robust probabilistic inference in bayesian network. In Meeyoung Cha, Chanyoung Park, Noseong Park, Carl Yang, Senjuti Basu Roy, Jessie Li, Jaap Kamps, Kijung Shin, Bryan Hooi, and Lifang He, editors, *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025, Seoul, Republic of Korea, November 10-14, 2025*, pages 3438–3447. ACM, 2025. doi: 10.1145/3746252.3761154. URL https://doi.org/10.1145/3746252.3761154.

[44] Xuan Xie, Xuan Wang, and Wenjie Wang. Dagrpo: Rectifying gradient conflict in reasoning via distinctiveness-aware group relative policy optimization. *arXiv preprint arXiv:2512.06337*, 2025.

[45] Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework for reinforce-style LLM training. *CoRR*, abs/2510.04996, 2025. doi: 10.48550/ARXIV.2510.04996. URL https://doi.org/10.48550/arXiv.2510.04996.

[46] Huei-Fang Yang, Kevin Lin, and Chu-Song Chen. Cross-batch reference learning for deep classification and retrieval. In Alan Hanjalic, Cees Snoek, Marcel Worring, Dick C. A. Bulterman, Benoit Huet, Aisling Kelliher, Yiannis Kompatsiaris, and Jin Li, editors, *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pages 1237–1246. ACM, 2016. doi: 10.1145/2964284.2964324. URL https://doi.org/10.1145/2964284.2964324.

[47] Zhuliang Yao, Yue Cao, Shuxin Zheng, Gao Huang, and Stephen Lin. Cross-iteration batch normalization. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12331–12340. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.01215. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yao_Cross-Iteration_Batch_Normalization_CVPR_2021_paper.html.

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

[49] Enci Zhang, Xingang Yan, Wei Lin, Tianxiang Zhang, and Qianchun Lu. Learning like humans: Advancing LLM reasoning capabilities via adaptive difficulty curriculum learning and expert-guided self-reformulation. *CoRR*, abs/2505.08364, 2025. doi: 10.48550/ARXIV.2505.08364. URL https://doi.org/10.48550/arXiv.2505.08364.

[50] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[51] Qinghua Zhang. Stability analysis of the kalman filter under practical conditions. *Autom.*, 183:112670, 2026. doi: 10.1016/J.AUTOMATICA.2025.112670. URL https://doi.org/10.1016/j.automatica.2025.112670.

[52] Zhixia Zhang, Zixuan Huang, Gongxun Li, Huaiyang Wang, Chengyi Yuan, Xin Xia, Deqing Wang, Fuzhen Zhuang, Shuai Ma, Ning Ding, et al. Heterogeneous agent collaborative reinforcement learning. *arXiv preprint arXiv:2603.02604*, 2026.

[53] Yuzhong Zhao, Yue Liu, Junpeng Liu, Jingye Chen, Xun Wu, Yaru Hao, Tengchao Lv, Shaohan Huang, Lei Cui, Qixiang Ye, Fang Wan, and Furu Wei. Geometric-mean policy optimization. *CoRR*, abs/2507.20673, 2025. doi: 10.48550/ARXIV.2507.20673. URL https://doi.org/10.48550/arXiv.2507.20673.

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

[55] Jiaru Zou, Yikun Ban, Zihao Li, Yunzhe Qi, Ruizhong Qiu, Ling Yang, and Jingrui He. Transformer copilot: Learning from the mistake log in LLM fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=MRvxlTlkNQ.

Appendix

A Limitations

This work reveals an intrinsic limitation of group-relative RL—namely, biased advantage estimation under non-degenerate sampling—and proposes HA-DW to effectively mitigate this issue. Our study primarily focuses on the issue of group-wise estimation bias, restricting the application of HA-DW to group-relative methods. Nevertheless, estimation bias is pervasive, and future work will focus on extending this concept to a broader scope. HA-DW is theoretically applicable to tasks beyond standard RLVR. However, the analysis and experiments in this work are mainly conducted under the RLVR setting for reasoning tasks, such as mathematical reasoning and code generation. Its potential applications in broader domains remain to be further explored.

B Future Work

Future work will explore extending HA-DW beyond mathematical reasoning to broader RLVR settings, including agentic tasks, long-horizon decision making, and environments with non-binary or noisy reward signals. As group-relative optimization is increasingly adopted in emerging RL paradigms, HA-DW may serve as a general difficulty-aware reweighting module that can be integrated with these methods to improve training efficiency and optimization stability. Another promising direction is to develop more robust difficulty estimation under imperfect verifiers and to investigate mechanisms that can additionally recover useful training signals from degenerate all-correct or all-incorrect groups.

C More Related Work

Group-based RLVR. Artificial intelligence has achieved significant advances in recent years[55, 12, 28, 19, 29, 14, 22]. Recent studies have proposed numerous improvements to group-based reinforcement learning algorithms[15, 24, 18, 40]. DaGRPO [44] tackles GRPO's instability and poor sample efficiency (caused by low distinctiveness in on-policy rollouts) by introducing sequence-level gradient rectification to filter low-distinctiveness pairs and off-policy anchor augmentation to restore learning signals on hard prompts. To address the advantage reversion and advantage mirror issues of fixed advantage formulations in GRPO that fail to adapt to samples with varying trajectory certainty. LSPO [4] adopts length-aware dynamic sampling to retain shortest/longest responses, addressing the ineffectiveness of RLVR training for LLM reasoning. GMPO [53] uses the geometric mean of token-level rewards (replacing GRPO's arithmetic mean) to resolve unstable policy updates from outlier importance sampling ratios. FAPO [6] uses a generative reward model (GenRM) to detect flawed-positive rollouts and a parameter-free reward penalty, addressing unreliable reasoning patterns and performance limitations caused by such rollouts in RLVR. F-GRPO [27] extends GRPO with focal-style reweighting to down-weight easy prompts, reduce overfitting to common solutions, and preserve rare correct ones. VESPO [32] stabilizes off-policy LLM training through variational sequence-level soft importance reweighting, avoiding length-normalization bias while remaining robust to stale and asynchronous rollouts. KTAE [35] uses statistical analysis to quantify tokens' association with correct rollouts and combines it with rollout-level advantages, solving the coarse granularity issue of GRPO that ignores token-specific contributions. Dr.GRPO [23] further revisits GRPO-style reinforcement learning by removing the response-length normalization term and adopting a simpler group-relative objective, showing that such normalization choices can substantially affect training stability and final reasoning performance. [52] investigates advantage estimation in heterogeneous multi-agent collaborative reinforcement learning to facilitate effective knowledge transfer and collaborative policy optimization across agents.

Leverage of Cross-batch Signals. Cross-batch signals have found widespread application across numerous domains. SamS [14] leverages a cross-batch scheduler to identify high-quality samples for direct preference optimization. XBM [42] improves embedding learning by leveraging memory from previous batches to enhance the consistency and quality of embeddings. CBNS [41] introduces a method to improve negative sampling in embedding learning by utilizing negative samples from different batches, enhancing the model's ability to learn more robust and generalized representations. CIBN [47] extends traditional batch normalization across iterations, rather than within a single batch,

to improve model convergence and generalization. CBRL [46] utilizes reference samples from different batches during training to improve the learning of deep classification and retrieval models. CBHEM-PLB [37] combines cross-batch hard example mining with a pseudo large batch strategy to improve face recognition models.

Biased Estimation. Considerable research effort has been directed towards addressing the critical challenge of biased estimation. The Bias–Variance Tradeoff theory [10, 26] suggests that as a model's complexity increases, its bias decreases but its variance increases, and vice versa. It emphasizes that there is a balance between bias and variance that affects the overall error in model predictions, and finding the optimal model complexity is crucial to minimize both bias and variance. Retrace [25] addresses the challenge of bias estimation in off-policy reinforcement learning. It proposes a retracing technique to mitigate the bias caused by off-policy data, which can lead to inaccurate value estimates. V-trace [7, 3] introduces a method for improving off-policy reinforcement learning by applying importance-weighted corrections to the value function updates in actor-critic algorithms, mitigating bias in off-policy data. DR-OVR [16, 39] combines importance sampling and regression to correct for bias in off-policy value estimation, making it more stable and accurate.

D Detailed Instantiations for GRPO and Related Algorithms

In this section, we present detailed instantiations of three group-relative reinforcement learning algorithms: GRPO, GSPO, and DAPO. And t and τ denote training step and token index in this part.

GRPO streamlines PPO by discarding the value network without compromising stability. Instead of fitting a baseline, it derives the advantage using group-relative normalization. This group-normalized advantage is then assigned uniformly to all tokens in the response, formulating the clipped surrogate loss:

$$J_{\text{GRPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_{t,i}|}\sum_{\tau=1}^{|o_{t,i}|}\min\left(r_{t,i,\tau}(\theta)\hat{A}_{t,i,\tau},\ \text{clip}\left(r_{t,i,\tau}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{t,i,\tau}\right), \tag{24}$$

where ϵ is the clipping hyperparameter and $r_{t,i,\tau}$ is is the importance sampling ratio comparing the new and old policy:

$$r_{t,i,\tau}(\theta) = \frac{\pi_\theta(y_{t,i,\tau} \mid x_t, y_{t,i,<\tau})}{\pi_{\theta_{\text{old}}}(y_{t,i,\tau} \mid x_t, y_{t,i,<\tau})}. \tag{25}$$

And GRPO defines the group advantage by subtracting the average reward of the group and normalizing by its standard deviation:

$$\hat{A}_{t,i,\tau} = \frac{R(x_t, o_{t,\tau}) - \text{mean}\left(\{R(x_t, o_{t,j})\}_{j=1}^{G}\right)}{\text{std}\left(\{R(x_t, o_{t,j})\}_{j=1}^{G}\right)}, \tag{26}$$

where $R(x, o)$ denotes the reward function.

The objective function of GRPO applied with HA-DW can be denoted as:

$$\begin{aligned} &J_{\text{GRPO+HA-DW}}(\theta) \\ &= \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_{t,i}|}\sum_{\tau=1}^{|o_{t,i}|}\min\left(r_{t,i,\tau}(\theta)\hat{A}_{t,i,\tau}\cdot\Phi_{t,i},\ \text{clip}\left(r_{t,i,\tau}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_{t,i,\tau}\cdot\Phi_{t,i}\right), \end{aligned} \tag{27}$$

where $\Phi_{t,i}$ is the history-aware reweighting factor defined earlier.

GSPO optimizes policy learning by defining importance ratios at the sequence level, eliminating the need for a critic model. Rather than relying on a separate value network, it computes advantages through normalized relative rewards of group responses. This sequence-level advantage is directly used for policy updates without token-level processing, yielding the following objective function:

$$J_{\text{GSPO}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\min\left(r_{t,i}(\theta)\hat{A}_{t,i},\ \text{clip}\Big(r_{t,i}(\theta),\ 1-\epsilon,\ 1+\epsilon\Big)\hat{A}_{t,i}\right), \tag{28}$$

where the sequence-level importance sampling ratio $r_i(\theta)$ can be denoted as:

$$r_{t,i}(\theta) = \frac{\pi_\theta(y_{t,i}|x_t)}{\pi_{\theta_{\text{old}}}(y_{t,i}|x_t)} = \frac{\prod_{\tau=1}^{|y_{t,i}|} \pi_\theta(y_{t,i,\tau}|x_t, y_{t,i,<\tau})}{\prod_{t=1}^{|y_{t,i}|} \pi_{\theta_{\text{old}}}(y_{t,i,\tau}|x_t, y_{t,i,<\tau})}, \tag{29}$$

where the advantage for GSPO can be denoted as:

$$\hat{A}_{t,i} = \frac{R(x_t, o_{t,i}) - \text{mean}\left(\{R(x_t, o_{t,j})\}_{j=1}^G\right)}{\text{std}\left(\{R(x_t, o_{t,j})\}_{j=1}^G\right)} \tag{30}$$

And the objective function of GSPO+HA-DW is:

$$\begin{aligned} &J_{\text{GSPO+HA-DW}}(\theta) \\ &= \frac{1}{G}\sum_{i=1}^{G} \min\left(r_{t,i}(\theta)\hat{A}_{t,i}\cdot\Phi_{t,i},\ \text{clip}\Big(r_{t,i}(\theta),\ 1-\epsilon,\ 1+\epsilon\Big)\hat{A}_{t,i}\cdot\Phi_{t,i}\right). \end{aligned} \tag{31}$$

DAPO's key feature is operating at the token level instead of treating full responses as single units, ensuring each token in sampled output o_i contributes proportionally to gradient updates. This fine-grained optimization boosts training stability and delivers more informative feedback for LLMs. The objective function is defined as:

$$J_{\text{DAPO}}(\theta) = \frac{1}{\sum_{i=1}^{G}|o_{t,i}|}\sum_{i=1}^{G}\sum_{\tau=1}^{|o_{t,i}|} \min\left(r_{t,i,\tau}(\theta)\hat{A}_{t,i,\tau}, \text{clip}\big(r_{t,i,\tau}(\theta), 1-\epsilon, 1+\epsilon'\big)\hat{A}_{t,i,\tau}\right). \tag{32}$$

DAPO introduces two key mechanisms: decoupled clipping and dynamic sampling, to address the limitations of traditional group-based methods. Decoupled clipping refines the trust region for more stable updates, while dynamic sampling mitigates estimation bias by adaptively reweighting samples based on their distribution.

Applying HA-DW on Equation (32), and we have:

$$\begin{aligned} &J_{\text{DAPO+HA-DW}}(\theta) \\ &= \frac{1}{\sum_{i=1}^{G}|o_{t,i}|}\sum_{i=1}^{G}\sum_{\tau=1}^{|o_{t,i}|} \min\left(r_{t,i,\tau}(\theta)\hat{A}_{t,i,\tau}\cdot\Phi_{t,i}, \text{clip}\big(r_{t,i,\tau}(\theta), 1-\epsilon, 1+\epsilon'\big)\hat{A}_{t,i,\tau}\cdot\Phi_{t,i}\right). \end{aligned} \tag{33}$$

E Setup Details

Models & Datasets. We conduct our experiments on Qwen3-4B-Base, Qwen3-8B-Base [38] and LLaMA-3.2-3B-Instruct to assess the mathematical reasoning performance of different algorithms across models of varying scales and family. Our training dataset is sourced from MATH dataset [13, 20] which contains $7.5k$ questions for training. Our evaluation suite includes: MATH500 [13], AMC23, AIME25, Minerva, and OlympiadBench [11]. To mitigate high variance on small benchmark sets and obtain reliable results, we report avg@16 on AIME25 and AMC23.

Baseline. We apply our proposed method on top of several representative group-relative reinforcement learning algorithms: GRPO, GSPO, and DAPO. We compare the performance of group-relative algorithms applying HA-DW to original ones, verifying the effectiveness and scalability of our method.

Implementation Details. We conduct RL training within the VeRL framework [34] on a single node with 8 $\times$ NVIDIA A100 GPUs. All experiments use a maximum prompt batch size of 1,024 and a maximum response length of 4,096. More hyperparameter details are provides in appendix E.

Training Hyperparameters. The detailed hyperparameters used during our training process on 6 different methods of 3 models (Qwen3-4B-Base, Qwen3-8B-Base and LLaMA-3.2-3B-Instruct) used in our experiments are demonstrated in Table 13.

F Theoretical Proof

F.1 Proof of Theorem 1

Fix a prompt and policy under the assumptions of Theorem 1. We analyze the non-degenerate event

$$\mathcal{S} \coloneqq \{1 \le R \le G-1\}.$$

Under the retention condition $\mathcal{S}$, $\mathbb{E}\left[\hat{p}_t \mid S\right]$ denotes the conditional expectation of the empirical estimation $\hat{p}_t = R/G$. And it can be derived what the relationship is between it and the expected reward p_t:

$$\begin{aligned}
\mathbb{E}\left[\hat{p}_t \mid \mathcal{S}\right] &= \mathbb{E}\left[\frac{R}{G} \mid \mathcal{S}\right] \\
&= \frac{1}{G} \cdot \frac{\mathbb{E}\left[R \cdot \mathbf{1}_{\{\mathcal{S}\}}\right]}{\mathbb{P}(\mathcal{S})} \\
&= \frac{1}{G} \cdot \frac{\mathbb{E}[R] - \mathbb{E}\left[R \cdot \mathbf{1}_{\{R=G\}}\right]}{\mathbb{P}(\mathcal{S})} \\
&= \frac{1}{G} \cdot \frac{Gp_t - G\mathbb{P}(R=G)}{\mathbb{P}(\mathcal{S})} \\
&= \frac{p_t - p_t^G}{1-(1-p_t)^G - p_t^G},
\end{aligned} \tag{34}$$

where the indicator function $\mathbf{1}_{\{\mathcal{S}\}}$ takes the value 1 if the event $\mathcal{S}$ occurs and 0 otherwise. Through the conditional expectation of $\hat{p}_t$, we can obtain that its expected value is more than p_t when $p_t < \frac{1}{2}$ and the baseline tends to be overestimated. Conversely, when $p_t > \frac{1}{2}$, the expected value is smaller than p_t, leading to an underestimation.

For every sampled group, $\sum_{i=1}^{G} \hat{A}_{t,i} = 0$. The rewards are exchangeable and $\mathcal{S}$ is permutation-invariant, so $\mathbb{E}[\hat{A}_{t,i} \mid \mathcal{S}] = 0$. The samplewise identity

$$\hat{A}_{t,i} - A_{t,i} = (r_{t,i} - \hat{p}_t) - (r_{t,i} - p_t) = p_t - \hat{p}_t$$

then gives

$$\begin{aligned}
&\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}] \\
&\quad = p_t - \mathbb{E}[\hat{p}_t \mid \mathcal{S}] \\
&\quad = p_t - \frac{p_t - p_t^G}{1-(1-p_t)^G - p_t^G} \\
&\quad = \frac{p_t(1-p_t)\left[p_t^{G-1} - (1-p_t)^{G-1}\right]}{1-(1-p_t)^G - p_t^G}.
\end{aligned}$$

The denominator and $p_t(1-p_t)$ are positive because $G \ge 2$ and $p_t \in (0,1)$. The remaining factor is negative, zero, or positive according as $p_t < 1/2$, $p_t = 1/2$, or $p_t > 1/2$. This proves Theorem 1. Without conditioning on $\mathcal{S}$, $\mathbb{E}[\hat{p}_t] = p_t$, and therefore $\mathbb{E}[\hat{A}_{t,i} - A_{t,i}] = 0$.

Lemma 2. *Under the condition of Theorem 1, the bias induced by the group-relative advantage is formulated as:*

$$\mathbb{E}\left[A_{t,i} - \hat{A}_{t,i} \mid \mathcal{S}\right] = \frac{p_t(1-p_t)^G + p_t^{G+1} - p_t^G}{1-(1-p_t)^G - p_t^G}. \tag{35}$$

Proof.

$$\mathbb{E}\left[\hat{p}_t \mid \mathcal{S}\right] - p_t = \frac{p_t(1-p_t)^G + p_t^{G+1} - p_t^G}{1-(1-p_t)^G - p_t^G}. \tag{36}$$

The identity $A_{t,i} - \hat{A}_{t,i} = \hat{p}_t - p_t$ completes the proof. □

F.2 Proof of Corollary 1

Under the assumptions of Theorem 1, the reward-conditioned advantages satisfy

$$\begin{aligned}\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=1, \mathcal{S}] &= \frac{(G-1)(1-p_t)}{G(1-p_t^{G-1})}, \\ \mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=0, \mathcal{S}] &= -\frac{(G-1)p_t}{G\big(1-(1-p_t)^{G-1}\big)}.\end{aligned} \tag{37}$$

Proof. Let $n = G-1$, and let K be the number of correct responses among the other n samples. Conditional on either value of $r_{t,i}$, independence gives $K \sim \mathrm{Binomial}(n, p_t)$. If $r_{t,i}=1$, then $\mathcal{S}$ is equivalent to $K \le n-1$, and $\hat{A}_{t,i} = (n-K)/G$. Since $n-K=0$ when $K=n$,

$$\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=1, \mathcal{S}] = \frac{\mathbb{E}[n-K]}{G(1-p_t^n)} = \frac{n(1-p_t)}{G(1-p_t^n)}.$$

If $r_{t,i}=0$, then $\mathcal{S}$ is equivalent to $K \ge 1$, and $\hat{A}_{t,i} = -K/G$. The term $K=0$ contributes nothing to $\mathbb{E}[K]$, so

$$\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=0, \mathcal{S}] = -\frac{\mathbb{E}[K]}{G[1-(1-p_t)^n]} = -\frac{np_t}{G[1-(1-p_t)^n]}.$$

Subtracting the corresponding targets yields

$$\begin{aligned}&\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=1, \mathcal{S}] - (1-p_t) \\ &\quad = \frac{(1-p_t)(Gp_t^{G-1}-1)}{G(1-p_t^{G-1})}, \\ &\mathbb{E}[\hat{A}_{t,i} \mid r_{t,i}=0, \mathcal{S}] + p_t \\ &\quad = \frac{p_t[1-G(1-p_t)^{G-1}]}{G[1-(1-p_t)^{G-1}]}.\end{aligned}$$

Since $G \le 2^{G-1}$ for $G \ge 2$, the first expression is negative when $p_t < 1/2$, and the second is positive when $p_t > 1/2$. □

F.3 Derivation of the Difficulty-Gap Formulation

Proof. Substituting $p_t = 1/2 - d$ into the conditional mean error in Theorem 1 gives $-B_G(d)$. Exchanging p_t and $1-p_t$ reverses its sign, so the value at $p_t = 1/2 + d$ is $B_G(d)$. For $d \in [0, 1/2)$, the numerator of $B_G(d)$ factors as

$$(\tfrac{1}{2}-d)(\tfrac{1}{2}+d)\left[(\tfrac{1}{2}+d)^{G-1} - (\tfrac{1}{2}-d)^{G-1}\right] \ge 0.$$

Its denominator is positive. Thus the absolute conditional mean error equals $B_G(d_t)$ and $B_G(0) = 0$.

For fixed G, let $N_G(d)$ and $D_G(d)$ denote the numerator and denominator of $B_G(d)$. The numerator is odd, the denominator is even, and

$$N_G'(0) = (G-1)2^{1-G}, \qquad D_G(0) = 1 - 2^{1-G} > 0.$$

Hence

$$\begin{aligned}N_G(d) &= (G-1)2^{1-G}d + O_G(d^3), \\ D_G(d) &= 1 - 2^{1-G} + O_G(d^2).\end{aligned}$$

Dividing gives

$$B_G(d) = \frac{G-1}{2^{G-1}-1}d + O_G(d^3).$$

Finally, for $a_G = (G-1)/(2^{G-1}-1)$,

$$a_G - a_{G+1} = \frac{(G-2)2^{G-1}+1}{(2^{G-1}-1)(2^G-1)} > 0.$$

Therefore, the leading coefficient decreases with G. □

F.4 Proof of Theorem 2 and Corollary 2

For hard prompts, in Theorem 2, we have:

$$\mathbb{P}\left(\hat{p}_t - p_t > \epsilon \mid \mathcal{S}\right) = \frac{\sum_{k=\lfloor G(p_t+\epsilon)\rfloor+1}^{G-1} \binom{G}{k} p_t^k (1-p_t)^{G-k}}{1-(1-p_t)^G - p_t^G}. \tag{38}$$

The above equation is given by the following argument: The conditioning event $\mathcal{S}$ restricts the sample space by excluding the outcome $R \in \{0, G\}$ (hence under $\mathcal{S}$ we only keep $R \in \{1, \ldots, G-1\}$). Let:

$$m(p_t) \coloneqq \lfloor G(p_t + \epsilon) \rfloor + 1. \tag{39}$$

Therefore, within the event $\mathcal{S}$, the deviation event A becomes:

$$A \cap \mathcal{S} = \{R \geq m(p_t)\} \cap \{1 \leq R \leq G-1\} = \{m(p_t) \leq R \leq G-1\}. \tag{40}$$

By definition of conditional probability, the numerator is the (unconditional) probability mass of all outcomes that satisfy the deviation requirement $\hat{p} - p > \epsilon$ and simultaneously, and satisfy the restriction imposed by $\mathcal{S}$. Because R is binomial, for any integer k we have:

$$\mathbb{P}(R = k) = \binom{G}{k} p_t^k (1-p_t)^{G-k}. \tag{41}$$

Summing over all admissible counts $k \in \{m(p_t), m(p_t)+1, \ldots, G-1\}$ yields:

$$\mathbb{P}(A \cap \mathcal{S}) = \sum_{k=m(p_t)}^{G-1} \mathbb{P}(R = k) = \sum_{k=m(p_t)}^{G-1} \binom{G}{k} p_t^k (1-p_t)^{G-k}. \tag{42}$$

Thus, based on the formula of conditional probability [43, 9], we can derive the conclusion of Theorem 2.

According to Theorem 2, we can formulate:

$$f(G, p_t) \coloneqq \mathbb{P}(\hat{p}_t - p_t > \epsilon \mid \mathcal{S}). \tag{43}$$

Assume that p_t follows a uniform distribution. And we define:

$$\mathbb{P}(G, p_{t_1}, p_{t_2}) \coloneqq \frac{1}{p_{t_2} - p_{t_1}} \int_{p_{t_1}}^{p_{t_2}} f(G, p_t) dp_t \tag{44}$$

where p_{t_1} and p_{t_2} is the expected reward. And $\mathbb{P}(G)$ reflects the probability that, when G is fixed, the baseline $\hat{p}_t$ is overestimated of group-relative RL algorithms over a certain expected-reward interval. For hard prompts with $p_t \in (0, 0.25)$ under different group size G, when $G \in [2, 8]$, we have:

G	$\mathbb{P}(G, 0, 0.25)$
2	0.999997499987
4	0.999995948256
6	0.827761785622
8	**0.781129955681**

Table 3: $\mathbb{P}(G, 0, 0.25)$ as a function of $G \in [2, 8]$.

Similarly, we can calculate hard prompts with $p_t \in (0, 0.5)$ under different group size G.

We can conclude from Table 3, when $2 \leq G \leq 8$, $\mathbb{P}(G, 0, 0.25) > 0.78$. This reveals that for hard prompts whose $p_t \in (0, 0.25)$ when G is limited, its baseline $\hat{p}_t$ of group-relative RL algorithms is substantially likely to be overestimated. Similarly, due to the evident symmetry of the group-relative methods, for easy prompt with $p_t \in (0.75, 1)$, the baseline $\hat{p}_t$ is underestimated with the same probability distribution.

G	$\mathbb{P}(G, 0, 0.5)$
2	0.999994999975
4	0.776965795853
6	0.689721502158
8	**0.640944744224**

Table 4: $\mathbb{P}(G, 0, 0.5)$ as a function of $G \in [2, 8]$.

Based on the aforementioned conclusions, for group-based algorithms, when $G \in [2, 8]$, the probability of biased advantage estimation can be denoted as:

$$\begin{aligned} &\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < 0.25\Big) > 0.78, \\ &\mathbb{P}\Big(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > 0.75\Big) > 0.78. \end{aligned} \tag{45}$$

Similarlly, Table 4 can give:

$$\begin{aligned} &\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < 0.5\Big) > 0.63, \\ &\mathbb{P}\Big(\hat{A}_{t,i} > A_{t,i} \mid \mathcal{S},\ p_t > 0.5\Big) > 0.63. \end{aligned} \tag{46}$$

For $p_t > 0.875$ and $p_t < 0.125$, the results are the adaptation of Corollary 4.

F.5 Conditional Concentration and Bias Rates

Corollary 5 (Conditional Concentration and Bias Rates)**.** *Under Theorem 1's assumptions, for $\epsilon > 0$,*

$$\begin{aligned} &\mathbb{P}\Big(|\hat{A}_{t,i} - A_{t,i}| \geq \epsilon \,\Big|\, \mathcal{S}\Big) \\ &\qquad \leq \min\left\{1, \frac{2e^{-2G\epsilon^2}}{1-(1-p_t)^G - p_t^G}\right\}. \end{aligned}$$

For $\delta \in (0, 1)$, with conditional probability at least $1 - \delta$,

$$|\hat{A}_{t,i} - A_{t,i}| \leq \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big[1-(1-p_t)^G - p_t^G\big]}\right)}. \tag{47}$$

The conditional mean bias is derived below. For fixed $\eta_{\text{margin}} \in (0, 1/2]$, uniformly over $p_t \in [\eta_{\text{margin}}, 1 - \eta_{\text{margin}}]$,

$$\Big|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}]\Big| = O_{\eta_{\text{margin}}}\big((1-\eta_{\text{margin}})^G\big) \quad \text{as } G \to \infty.$$

If $p_t = O(G^{-1})$ or $1 - p_t = O(G^{-1})$, with $p_t \in (0, 1)$ varying with G,

$$\Big|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}]\Big| = O(G^{-1}).$$

For fixed δ, the deviation bound in Equation (47) is $O(G^{-1/2})$ whenever the retention probability $1 - (1 - p_t)^G - p_t^G$ is bounded away from zero. A smaller retention probability weakens this concentration bound. The conditional mean bias has the separate rates above and decays exponentially in G when p_t stays a fixed distance from both endpoints. The proof follows.

Proof. Using $\hat{A}_{t,i} - A_{t,i} = p_t - \hat{p}_t$, conditional probability and Hoeffding's inequality give

$$\begin{aligned} &\mathbb{P}\Big(|\hat{A}_{t,i} - A_{t,i}| \geq \epsilon \,\Big|\, \mathcal{S}\Big) \\ &\quad \leq \frac{\mathbb{P}(|\hat{p}_t - p_t| \geq \epsilon)}{1-(1-p_t)^G - p_t^G} \\ &\quad \leq \frac{2e^{-2G\epsilon^2}}{1-(1-p_t)^G - p_t^G}. \end{aligned}$$

Combining this with the upper bound 1 proves the concentration inequality. Setting

$$\epsilon = \sqrt{\frac{1}{2G}\log\Big(\frac{2}{\delta[1-(1-p_t)^G - p_t^G]}\Big)}$$

gives Equation (47).

The conditional mean bias is

$$\Big|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}]\Big| = \frac{\big|p_t(1-p_t)^G - (1-p_t)p_t^G\big|}{1-(1-p_t)^G - p_t^G}.$$

For $p_t \le 1/2$, the numerator of the absolute conditional mean error is $p_t(1-p_t)^G - (1-p_t)p_t^G \ge 0$, and

$$1-(1-p_t)^G - p_t^G = \mathbb{P}(\mathcal{S}) \ge \mathbb{P}(R=1) = Gp_t(1-p_t)^{G-1}.$$

Therefore,

$$\begin{aligned}\Big|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}]\Big| &\le \frac{p_t(1-p_t)^G}{Gp_t(1-p_t)^{G-1}} \\ &= \frac{1-p_t}{G} \le \frac{1}{G}.\end{aligned}$$

The case $p_t \ge 1/2$ follows by symmetry. This bound holds throughout $(0,1)$ and in particular establishes the stated extreme-regime rate.

If $p_t \in [\eta_{\text{margin}}, 1-\eta_{\text{margin}}]$, then, since $G \ge 2$,

$$\begin{aligned}1-(1-p_t)^G - p_t^G &\ge 1-(1-p_t)^2 - p_t^2 \\ &= 2p_t(1-p_t) \\ &\ge 2\eta_{\text{margin}}(1-\eta_{\text{margin}}),\end{aligned}$$

while

$$\begin{aligned}&\big|p_t(1-p_t)^G - (1-p_t)p_t^G\big| \\ &\quad\le p_t(1-p_t)^G + (1-p_t)p_t^G \\ &\quad\le (1-\eta_{\text{margin}})^G.\end{aligned}$$

Consequently,

$$\begin{aligned}\Big|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid \mathcal{S}]\Big| &\le \frac{(1-\eta_{\text{margin}})^G}{2\eta_{\text{margin}}(1-\eta_{\text{margin}})} \\ &= O_{\eta_{\text{margin}}}\big((1-\eta_{\text{margin}})^G\big),\end{aligned}$$

uniformly over the stated interval. □

F.6 Proof of Corollary 3 and Corollary 4

Let G be a large integer, for hard prompts, according to Theorem 2, we have:

$$\mathbb{P}\left(\hat{p}_t - p_t > \epsilon \mid \mathcal{S}\right) = \frac{\sum_{k=\lfloor G(p_t+\epsilon)\rfloor+1}^{G-1}\binom{G}{k}p_t^k(1-p_t)^{G-k}}{1-(1-p_t)^G - p_t^G}. \tag{48}$$

And we define:

$$f(p_t) \coloneqq \mathbb{P}(\hat{p}_t - p_t > \epsilon \mid \mathcal{S}). \tag{49}$$

We analyze the integral in the limit of large G using the Poisson approximation [30]. Let us perform the change of variable $x_t = Gp_t$. The limits of integration change from $[1/G, 2/G]$ to $[1,2]$, and $dp_t = dx_t/G$. We define the integral of interest:

$$\mathbb{P}(G_1, G_2) = \frac{G}{G_2 - G_1}\int_{G_1}^{G_2} f(x_t/G)\frac{dx_t}{G} = \int_1^2 f(x_t/G)\,dx_t. \tag{50}$$

First, we determine the summation lower bound $m(p_t)$. For $p_t \in [1/G, 2/G)$, we have $Gp_t \in [1,2)$. Consequently, $\lfloor Gp_t \rfloor = 1$, which implies:

$$m(p_t) = \lfloor Gp_t \rfloor + 1 = 2. \tag{51}$$

Next, we approximate the binomial terms. In the limit $G \to \infty$ with $Gp_t = x$ fixed, the binomial distribution converges to the Poisson distribution with parameter x_t. The denominator $Z(p_t)$ approximates to:

$$Z(p_t) = 1 - (x_t/G)^G - (1 - x_t/G)^G \xrightarrow{G\to\infty} 1 - e^{-x_t}. \tag{52}$$

The numerator is the probability that a Poisson random variable $K \sim \text{Pois}(x_t)$ takes a value $k \geq 2$ (ignoring the upper limit $G-1$ as the Poisson tail vanishes exponentially):

$$\begin{aligned} f(p_t) = \mathbb{P}(K \geq 2) &= \sum_{k=2}^{\infty} \frac{x_t^k e^{-x_t}}{k!} \\ &= 1 - \mathbb{P}(K=0) - \mathbb{P}(K=1) \\ &= 1 - e^{-x_t} - xe^{-x_t} \\ &= 1 - e^{-x_t}(1 + x_t). \end{aligned} \tag{53}$$

Substituting these approximations into $f(x_t/G)$, we obtain the limiting integrand $h(x_t)$:

$$h(x_t) = \frac{1 - e^{-x_t}(1+x_t)}{1 - e^{-x_t}} = 1 - \frac{x_t e^{-x_t}}{1 - e^{-x_t}} = 1 - \frac{x_t}{e^{x_t} - 1}. \tag{54}$$

Assume that p_t follows a uniform distribution. Calculating Equation (54) numerically, for sufficiently large G, we can show that $\mathbb{P}(0,2) = \frac{G}{2}\int_0^2 f(x_t)\,dx_t = G\frac{\mathbb{P}(0,1)+\int_1^2 h(x_t)\,dx_t}{2} \approx 0.7818$.

Next, we use numerical computation to show how large G should be, whose result can be found in the following table:

G	$\frac{G}{2}\int_0^{2/G} f(p_t)\,dp_t$
2	0.499997499987
3	0.749995833315
4	0.776965795853
5	0.780787089465
6	**0.781154327380**

Table 5: $\frac{G}{2}\int_0^{2/G} f(p_t)\,dp_t$ as a function of $G \in [2,6]$.

Thus, $G \geq 6$ is sufficiently large to have:

$$\mathbb{P}\Big(\hat{A}_{t,i} < A_{t,i} \mid \mathcal{S},\ p_t < \frac{2}{G}\Big) > 0.78\,. \tag{55}$$

Proof of Corollary 4. On S, we have $R \geq 1$, hence $\hat{p} = R/G \geq 1/G$. Since $p < 1/G$, it follows that $\hat{p} \geq 1/G > p$. On S, we have $R \leq G-1$, hence $\hat{p} = R/G \leq (G-1)/G$. Since $p > (G-1)/G$, it follows that $\hat{p} \leq (G-1)/G < p$. This leads to the Corollary.

F.7 Proof of Lemma 1 and Theorem 5

F.7.1 Proof of Lemma 1

Before proving Lemma 1, we begin with the following auxiliary result.

Lemma 3. *Define the non-degenerate event $\mathcal{S} := \{1 \leq S \leq G-1\}$, and $\epsilon \in (0, |p_t - \hat{p}_t|)$. If*

$$c \in \left(\frac{(p_t - \epsilon)\cdot\big(1 - (1-p_t)^G - p_t^G\big)}{p_t(1 - p_t^{G-1})},\ \frac{(p_t + \epsilon)\cdot\big(1 - (1-p_t)^G - p_t^G\big)}{p_t(1 - p_t^{G-1})} \right), \tag{56}$$

we have:

$$\mathbb{E}[\tilde{p}_t \mid \mathcal{S}] \in (p_t - \epsilon,\ p_t + \epsilon)\,. \tag{57}$$

Proof. We define the adjusted factor c to compensate for the bias in the advantage estimation which applied on empirical group baseline $\hat{p}_t$. The globally scaled estimator can be approximated as:

$$\tilde{p}_t(R) \coloneqq c\,\hat{p}_t = c\,\frac{R}{G}. \tag{58}$$

We can derive the conditional expectation of $\tilde{p}_t$ on non-degenerate event $\mathcal{S} = \{1 \le R \le G-1\}$:

$$\mathbb{E}[\tilde{p}_t \mid \mathcal{S}] = \mathbb{E}\left[c\frac{R}{G}\,\middle|\,\mathcal{S}\right] = \frac{c}{G}\mathbb{E}[R \mid \mathcal{S}] = \frac{c}{G}\frac{\mathbb{E}[R\cdot\mathbf{1}_{\{\mathcal{S}\}}]}{\mathbb{P}(\mathcal{S})}. \tag{59}$$

And we have:

$$\mathbb{E}[R\cdot\mathbf{1}_{\{S\}}] = \sum_{k=1}^{G-1} k\,\mathbb{P}(R=k) = \mathbb{E}[R] - G\mathbb{P}(R=G). \tag{60}$$

Because the only term excluded from $\sum_{k=0}^{G} k\mathbb{P}(R=k) = \mathbb{E}[R]$ is the $k=G$ term (the $k=0$ term is zero anyway). Using $\mathbb{E}[R] = Gp_t$ and $\mathbb{P}(R=G) = p_t^G$, we can obtain:

$$\mathbb{E}[R\cdot\mathbf{1}_{\{S\}}] = Gp_t - Gp_t^G = Gp_t(1-p_t^{G-1}). \tag{61}$$

Therefore:

$$\mathbb{E}[\tilde{p}_t \mid S] = \frac{c}{G}\cdot\frac{Gp_t(1-p_t^{G-1})}{1-(1-p_t)^G-p_t^G} = c\,p_t\,\frac{1-p_t^{G-1}}{1-(1-p_t)^G-p_t^G}, \tag{62}$$

which proves the stated conditional expectation formula.

To mitigate the biased estimation, let:

$$\mathbb{E}[\tilde{p}_t \mid S] = p_t. \tag{63}$$

And we can solve:

$$c\,p_t\,\frac{1-p_t^{G-1}}{1-(1-p_t)^G-p_t^G} = p_t. \tag{64}$$

The analytical solution for this equation is:

$$c = \frac{1-(1-p_t)^G-p_t^G}{1-p_t^{G-1}}. \tag{65}$$

When the adjustment coefficient c falls within a specific range of values, we will have $|\tilde{p}_t - p_t| < |\hat{p}_t - p_t|$. We first let:

$$\epsilon = |\hat{p}_t - p_t|. \tag{66}$$

For $\mathbb{E}[\tilde{p}_t \mid S] = p_t + \epsilon$, solve:

$$c_+\,p_t\,\frac{1-p_t^{G-1}}{1-(1-p_t)^G-p_t^G} = p_t + \epsilon. \tag{67}$$

And we can derive:

$$c_+ = \frac{(p_t+\epsilon)\cdot\left(1-(1-p_t)^G-p_t^G\right)}{p_t(1-p_t^{G-1})} = \left(1+\frac{\epsilon}{p_t}\right)c. \tag{68}$$

For $\mathbb{E}[\tilde{p}_t \mid S] = p_t - \epsilon$, solve:

$$c_-\,p_t\,\frac{1-p_t^{G-1}}{1-(1-p_t)^G-p_t^G} = p_t - \epsilon. \tag{69}$$

Thus, we have:

$$c_- = \frac{(p_t-\epsilon)\cdot\left(1-(1-p_t)^G-p_t^G\right)}{p_t(1-p_t^{G-1})} = \left(1-\frac{\epsilon}{p_t}\right)c. \tag{70}$$

$$c \in \left(\frac{(p_t-\epsilon)\cdot\left(1-(1-p_t)^G-p_t^G\right)}{p_t(1-p_t^{G-1})},\ \frac{(p_t+\epsilon)\cdot\left(1-(1-p_t)^G-p_t^G\right)}{p_t(1-p_t^{G-1})}\right), \tag{71}$$

we have

$$\mathbb{E}[\tilde{p}_t \mid \mathcal{S}] \in (p_t-\epsilon,\ p_t+\epsilon)\,. \tag{72}$$

□

Lemma 4 (p_t-free concentration under $\mathcal{S}$)**.** *Define the non-degenerate event $\mathcal{S} := \{1 \le S \le G-1\}$. Assume $p_t \in [\eta_{\text{margin}},\, 1-\eta_{\text{margin}}]$ for some $\eta_{\text{margin}} \in (0, 1/2]$. Then for any $\zeta > 0$, we have:*

$$\mathbb{P}(|\hat{p}_t - p_t| < \zeta \,|\, \mathcal{S}) \ge \frac{1 - 2\exp(-2G\zeta^2) - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G}{1 - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G}. \tag{73}$$

Proof. Let $A := \{|\hat{p}_t - p_t| < \zeta\}$. By the definition of conditional probability:

$$\mathbb{P}(A \mid \mathcal{S}) = \frac{\mathbb{P}(A \cap \mathcal{S})}{\mathbb{P}(\mathcal{S})}. \tag{74}$$

We lower bound the numerator. Since $A \cap \mathcal{S} \supseteq A \setminus \mathcal{S}^-$, we have:

$$\mathbb{P}(A \cap \mathcal{S}) \ge \mathbb{P}(A) - \mathbb{P}(\mathcal{S}^-). \tag{75}$$

Next, note that $\mathcal{S}^- = \{S = 0\} \cup \{S = G\}$ and these two events are disjoint. Therefore:

$$\mathbb{P}(\mathcal{S}^-) = \mathbb{P}(S=0) + \mathbb{P}(S=G) = (1-p_t)^G + p_t^G. \tag{76}$$

Moreover, we can derive that:

$$\mathbb{P}(\mathcal{S}) = 1 - \mathbb{P}(\mathcal{S}^-) = 1 - (1-p_t)^G - p_t^G. \tag{77}$$

We now lower bound $\mathbb{P}(A)$ using Hoeffding's inequality. Since each $r_{t,i} \in [0,1]$ almost surely and $\{r_{t,i}\}_{i=1}^G$ are independent with $\mathbb{E}[r_{t,i}] = p_t$, Hoeffding's inequality yields:

$$\mathbb{P}(|\hat{p}_t - p_t| \ge \zeta) \le 2\exp(-2G\zeta^2), \tag{78}$$

equivalently:

$$\mathbb{P}(A) = \mathbb{P}(|\hat{p}_t - p_t| < \zeta) \ge 1 - 2\exp(-2G\zeta^2). \tag{79}$$

It remains to remove the dependence on p_t in $\mathbb{P}(\mathcal{S})$. Define $f(p) := p^G + (1-p)^G$. For $G \ge 1$, f is symmetric around $1/2$ and attains its maximum over $[\eta_{\text{margin}}, 1-\eta_{\text{margin}}]$ at the endpoints. Hence:

$$(1-p_t)^G + p_t^G = f(p_t) \le f(\eta_{\text{margin}}) = (1-\eta_{\text{margin}})^G + \eta_{\text{margin}}^G, \tag{80}$$

which implies:

$$\mathbb{P}(\mathcal{S}) = 1 - f(p_t) \ge 1 - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G. \tag{81}$$

Combining Equation (74) and (75) with Equation (79), (80), and (81), we can obtain that:

$$\begin{aligned} &\mathbb{P}(A \mid \mathcal{S}) && (82)\\ &\ge \frac{\mathbb{P}(A) - \mathbb{P}(\mathcal{S}^-)}{\mathbb{P}(\mathcal{S})} \\ &\ge \frac{\left[1 - 2\exp(-2G\zeta^2)\right] - \left[(1-\eta_{\text{margin}})^G + \eta_{\text{margin}}^G\right]}{1 - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G}, && (83) \end{aligned}$$

which completes the proof. □

Lemma 5 (Conditional p_t-free concentration under $\mathcal{S}$)**.** *Assume $p_t \in [\eta_{\text{margin}},\, 1-\eta_{\text{margin}}]$ for some $\eta_{\text{margin}} \in (0, 1/2]$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ conditional on $\mathcal{S}$, we have:*

$$|\hat{p}_t - p_t| < \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big(1 - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G\big)}\right)}. \tag{84}$$

Proof. Now choose γ such that the right-hand side of Equation (73) is at most δ, i.e.:

$$\frac{2\exp(-2G\gamma^2)}{1 - (1-\eta_{\text{margin}})^G - \eta_{\text{margin}}^G} \le \delta.$$

Solving for γ gives:

$$\gamma \geq \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big(1-(1-\eta_{\text{margin}})^G-\eta_{\text{margin}}^G\big)}\right)}.$$

Therefore, for

$$\gamma^\star := \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big(1-(1-\eta_{\text{margin}})^G-\eta_{\text{margin}}^G\big)}\right)},$$

we have $\mathbb{P}(|\hat{p}_t - p_t| \geq \gamma^\star \mid \mathcal{S}) \leq \delta$, equivalently, $\mathbb{P}(|\hat{p}_t - p_t| < \gamma^\star \mid \mathcal{S}) \geq 1-\delta$, which proves Equation (84). □

Finally, combining Lemma 3, 4, and 5 gives Lemma 1. We restated it here for completeness:

Lemma 6 (A p_t-free feasible range of c expressed via $\hat{p}_t$)**.** *Assume the conditions of Lemma 4 and define:*

$$\epsilon_\delta := \sqrt{\frac{1}{2G}\log\left(\frac{2}{\delta\big(1-(1-\eta_{\text{margin}})^G-\eta_{\text{margin}}^G\big)}\right)}. \tag{85}$$

Let:

$$I_t := \big[\hat{p}_t - \epsilon_\delta,\ \hat{p}_t + \epsilon_\delta\big] \cap [\eta_{\text{margin}}, 1-\eta_{\text{margin}}], \quad A(p) := 1-(1-p)^G - p^G. \tag{86}$$

Fix any $\epsilon > 0$, we define:

$$c_{\text{low}} := \sup_{p \in I_t} \frac{(p-\epsilon)\,A(p)}{p(1-p^{G-1})}, \tag{87}$$

and:

$$c_{\text{high}} := \inf_{p \in I_t} \frac{(p+\epsilon)\,A(p)}{p(1-p^{G-1})}. \tag{88}$$

Then, on the event $\{|\hat{p}_t - p_t| < \epsilon_\delta\}$ (which holds with probability at least $1-\delta$ conditional on $\mathcal{S}$), any choice

$$c \in (c_{\text{low}},\ c_{\text{high}}) \tag{89}$$

implies that the condition (71) *holds for the true p_t, and hence:*

$$\mathbb{E}[\tilde{p}_t \mid \mathcal{S}] \in (p_t - \epsilon,\ p_t + \epsilon).$$

F.7.2 Proof of Theorem 5

When applying adjustment on the advantage $\hat{A}_{t,i}$, we do not consider the standard deviation here, and assume that:

$$\Phi_{t,i}\hat{A}_{t,i} = r_{t,i} - \tilde{p}_t = r_{t,i} - c\hat{p}_t. \tag{90}$$

It is equivalent to:

$$\Phi_{t,i}r_{t,i} - \Phi_{t,i}\hat{p}_t = r_{t,i} - c\hat{p}_t. \tag{91}$$

And for correct responses with $r_{t,i} = 1$:

$$\Phi_{t,i} = \frac{1-c\hat{p}_t}{1-\hat{p}_t}. \tag{92}$$

While for incorrect responses with $r_{t,i} = 0$:

$$\Phi_{t,i} = c. \tag{93}$$

According to Equation (19):

$$\Phi_{t,i} = \lambda_{\text{scale}} \cdot \exp\left(D_{t,i} \cdot M_t\right), \tag{94}$$

the adjustment of $A_{t,i}$ can be categorized into four types. For responses in defined hard prompts with $r_{t,i} = 1$, the adjusted advantage can be denoted as:

$$\hat{A}^1_{t,i} = \lambda_{\text{scale}} \cdot \exp\left(M_t\right) \cdot \hat{A}_{t,i}. \tag{95}$$

For hard prompts, we have $c \in (0,1)$ and $\hat{p} \in (0,1)$. Based on Equation (71) and Lemma 6, to mitigate biased estimation, λ_{scale} satisfies:

$$\lambda^1_{\text{scale}} \in \left(\frac{1 + \frac{(1-c^{\text{hard}}_{\text{high}})\hat{p}_t}{1-\hat{p}_t}}{\exp(M_t)}, \ \frac{1 + \frac{(1-c^{\text{hard}}_{\text{low}})\hat{p}_t}{1-\hat{p}_t}}{\exp(M_t)} \right). \tag{96}$$

And for incorrect responses in hard prompts, we have:

$$\hat{A}^2_{t,i} = \frac{\lambda_{\text{scale}}}{\exp(M_t)} \cdot \hat{A}_{t,i}. \tag{97}$$

And we can set:

$$\lambda^2_{\text{scale}} \in \left(c^{\text{hard}}_{\text{low}} \cdot \exp(M_t), \ c^{\text{hard}}_{\text{high}} \cdot \exp(M_t) \right). \tag{98}$$

For easy prompts, we have $c > 1$ and $\hat{p} \in (0,1)$, thus for correct answers:

$$\lambda^3_{\text{scale}} \in \left(\left(1 + \frac{(1 - c^{\text{easy}}_{\text{high}})\hat{p}_t}{1-\hat{p}_t} \right) \cdot \exp(M_t), \ \left(1 + \frac{(1 - c^{\text{easy}}_{\text{low}})\hat{p}_t}{1-\hat{p}_t} \right) \cdot \exp(M_t) \right). \tag{99}$$

And for incorrect responses:

$$\lambda^4_{\text{scale}} \in \left(\frac{c^{\text{easy}}_{\text{low}}}{\exp(M_t)}, \ \frac{c^{\text{easy}}_{\text{high}}}{\exp(M_t)} \right). \tag{100}$$

In training process with HA-DW, to rectify the biased advantage estimation, there exists a specfic λ_{scale} supposing to satisfy:

$$\lambda_{\text{scale}} \in \lambda^1_{\text{scale}} \cup \lambda^2_{\text{scale}} \cup \lambda^3_{\text{scale}} \cup \lambda^4_{\text{scale}} \tag{101}$$

which denotes:

$$\begin{aligned} \lambda_{\text{scale}} \in & \left(\frac{1 + \frac{(1-c^{\text{hard}}_{\text{low}})\hat{p}_t}{1-\hat{p}_t}}{\exp(M_t)}, \ \frac{1 + \frac{(1-c^{\text{hard}}_{\text{low}})\hat{p}_t}{1-\hat{p}_t}}{\exp(M_t)} \right) \cup \\ & \left(\left(1 + \frac{(1 - c^{\text{easy}}_{\text{high}})\hat{p}_t}{1-\hat{p}_t} \right) \cdot \exp(M_t), \ \left(1 + \frac{(1 - c^{\text{easy}}_{\text{low}})\hat{p}_t}{1-\hat{p}_t} \right) \cdot \exp(M_t) \right) \cup \\ & \left(c^{\text{hard}}_{\text{low}} \cdot \exp(M_t), \ c^{\text{hard}}_{\text{low}} \cdot \exp(M_t) \right) \cup \\ & \left(\frac{c^{\text{easy}}_{\text{low}}}{\exp(M_t)}, \ \frac{c^{\text{easy}}_{\text{high}}}{\exp(M_t)} \right). \end{aligned} \tag{102}$$

Overall, since the difficulty does not affect the expressions, we can further derive Equation (102) as follows:

$$\lambda_{\text{scale}} \in \left(\frac{1 + \frac{(1-c_{\text{high}})\hat{p}_t}{1-\hat{p}_t}}{\exp(D_{t,i} \cdot M_t)}, \ \frac{1 + \frac{(1-c_{\text{low}})\hat{p}_t}{1-\hat{p}_t}}{\exp(D_{t,i} \cdot M_t)} \right) \cup \left(\frac{c_{\text{low}}}{\exp(D_{t,i} \cdot M_t)}, \ \frac{c_{\text{high}}}{\exp(D_{t,i} \cdot M_t)} \right). \tag{103}$$

When Equation (103) holds, our method HA-DW is efficient in compensating biased advantage estimation.

F.8 Bias Analysis of Advantage Estimation with Standard-Deviation Normalization

In this section, we will discuss group-relative advantage estimation with standard-deviation normalization. And the group-relative advantage with normalization is denoted as:

$$\hat{A}_{t,i} = \frac{r_{t,i} - \hat{p}_t}{\sqrt{\hat{p}_t \cdot (1 - \hat{p}_t)}}, \quad \hat{p}_t = \frac{1}{G} \sum_{i=1}^{G} r_{t,i}. \tag{104}$$

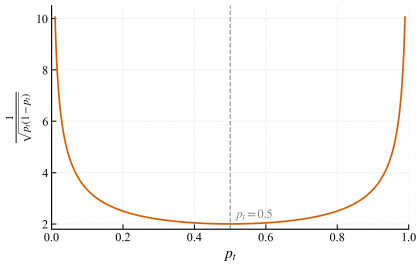

Figure 4: Curve of the normalization term $1/\sqrt{p_t(1-p_t)}$ with respect to p_t.

Firstly, based on our preceding theoretical analysis, we can observe that when $p_t < 0.5$, the expected reward p_t tends to be overestimated, whereas when $p_t > 0.5$, it tends to be underestimated. Meanwhile, as shown in Figure 4, when $p_t < 0.5$, the normalization term $\frac{1}{\sqrt{p_t \cdot (1-p_t)}}$ decreases monotonically; and when $p_t > 0.5$, the normalization term increases monotonically. Therefore, we can intuitively conclude that the advantage tends to be underestimated when $p_t < 0.5$, and the advantage tends to be overestimated when $p_t > 0.5$ which is consistent with the conclusion obtained above. We provide a detailed analysis of the expectation of the advantage estimator with standard-deviation normalization. Next, we present the main results.

Theorem 4 (Bias under Standard-Deviation Normalization)**.** *Under the retention condition $\mathcal{S} = \{1 \le R \le G-1\}$, for a correct response $r_{t,i} = 1$, the standard-deviation-normalized group-relative advantage satisfies:*

$$\mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \mid r_{t,i} = 1, \mathcal{S}\right] \neq \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}. \tag{105}$$

Moreover, the bias direction is deterministic in the extreme regimes:

$$\begin{cases} \mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \mid r_{t,i} = 1, \mathcal{S}\right] < \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}, & \text{if } p_t < \frac{1}{G}, \\ \mathbb{E}\left[\frac{\hat{A}_{t,i}}{\sqrt{\hat{p}_t(1-\hat{p}_t)}} \mid r_{t,i} = 1, \mathcal{S}\right] > \frac{A_{t,i}}{\sqrt{p_t(1-p_t)}}, & \text{if } p_t > \frac{G-1}{G}. \end{cases}$$

Thus, standard-deviation normalization rescales the group-relative advantage but does not eliminate its intrinsic conditioning bias.

Proof. Under the retention condition $\mathcal{S} \coloneqq \{1 \le R \le G-1\}$ of the group-relative paradigm, we separately analyze the relationship between the conditional expectation of the advantage estimation $\mathbb{E}\left[\hat{A}_t \mid S, r_{t,i}\right]$ with the expected reward p_t for correct and incorrect responses.

For correct responses with $r_{t,i} = 1$, since we condition on the event that the current response is correct, we necessarily have $R \ge 1$. Therefore, under $r_{t,i} = 1$, the event $S = \{1 \le R \le G-1\}$ is equivalent to $R \le G-1$. Let K denote the number of correct responses among the remaining $G-1$ samples. Then:

$$K \sim \text{Binomial}(G-1, p_t), \tag{106}$$

and the total number of correct responses in the group is $R = 1 + K$. The condition $R \le G-1$ implies $1 + K \le G-1$, and hence $K \le G-2$. For a correct response, the normalized group-relative advantage is:

$$\hat{A}^+_{t,i} = \frac{1-\hat{p}_t}{\sqrt{\hat{p}_t(1-\hat{p}_t)}}. \tag{107}$$

Leveraging $\hat{p}_t = R/G = (1+K)/G$, we can obtain:

$$\hat{A}^+_{t,i} = \sqrt{\frac{G-1-K}{1+K}}. \tag{108}$$

Thus, the conditional expectation can be written as:

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] = \sum_{k=0}^{G-2} \mathbb{P}(K = k \mid K \le G-2) \sqrt{\frac{G-1-k}{1+k}}. \tag{109}$$

Since:

$$\mathbb{P}(K = k \mid K \leq G-2) = \frac{\binom{G-1}{k} p_t^k (1-p_t)^{G-1-k}}{1 - p_t^{G-1}}, \tag{110}$$

where

$$\mathbb{P}(K \leq G-2) = 1 - \mathbb{P}(K = G-1) = 1 - p_t^{G-1}, \tag{111}$$

we obtain:

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] = \frac{\sum_{k=0}^{G-2} \binom{G-1}{k} p_t^k (1-p_t)^{G-1-k} \sqrt{\frac{G-1-k}{1+k}}}{1 - p_t^{G-1}}. \tag{112}$$

Here, the denominator $1 - p_t^{G-1}$ corresponds to excluding the case where all of the remaining $G-1$ responses are also correct, i.e., $R = G$, for which the group standard deviation is zero and the normalized advantage is undefined.

Therefore, the estimation bias for correct responses can be written as:

$$\Delta^+(p_t, G) = \frac{\sum_{k=0}^{G-2} \binom{G-1}{k} p_t^k (1-p_t)^{G-1-k} \sqrt{\frac{G-1-k}{1+k}}}{1 - p_t^{G-1}} - \sqrt{\frac{1-p_t}{p_t}}. \tag{113}$$

Equivalently, it can be written as a single fraction:

$$\Delta^+(p_t, G) = \frac{\sum_{k=0}^{G-2} \binom{G-1}{k} p_t^k (1-p_t)^{G-1-k} \sqrt{\frac{G-1-k}{1+k}} - \left(1 - p_t^{G-1}\right) \sqrt{\frac{1-p_t}{p_t}}}{1 - p_t^{G-1}}. \tag{114}$$

Since $p_t \in (0,1)$ and $1 - p_t^{G-1} > 0$, the function $\Delta^+(p_t, G)$ is continuous on $(0,1)$. When $p_t \to 0^+$, the dominant term in the summation is $k = 0$. Hence,

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] \to \sqrt{G-1}. \tag{115}$$

However, the true advantage satisfies:

$$A_{t,i}^+ = \sqrt{\frac{1-p_t}{p_t}} \to +\infty. \tag{116}$$

Therefore,

$$\lim_{p_t \to 0^+} \Delta^+(p_t, G) = -\infty. \tag{117}$$

This indicates that, for extremely difficult prompts, the advantage of correct responses is severely underestimated. When $p_t \to 1^-$, the valid sampling condition excludes the all-correct case $R = G$. Conditioned on $r_{t,i} = 1$, the dominant valid case is that exactly one of the remaining $G-1$ responses is incorrect. Thus,

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] \to \frac{1}{\sqrt{G-1}}. \tag{118}$$

Meanwhile,

$$A_{t,i}^+ = \sqrt{\frac{1-p_t}{p_t}} \to 0. \tag{119}$$

Therefore,

$$\lim_{p_t \to 1^-} \Delta^+(p_t, G) = \frac{1}{\sqrt{G-1}} > 0. \tag{120}$$

This implies that, for extremely easy prompts, the advantage of correct responses is overestimated.

Next, we analyze whether the bias:

$$\Delta^+(p_t, G) = \mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] - \sqrt{\frac{1-p_t}{p_t}} \tag{121}$$

is positive or negative around $p_t = 0.5$. The key observation is that, under the retention condition $S = \{1 \leq R \leq G-1\}$, the zero point is generally shifted to the right of 0.5. Therefore, while

the estimator is provably downward biased for $p_t < 0.5$, it is not necessarily upward biased for all $p_t > 0.5$. Let

$$n = G - 1, \qquad q = 1 - p_t, \tag{122}$$

and define

$$B_k = \binom{n}{k} p_t^k q^{n-k}. \tag{123}$$

Then the expected estimated advantage for correct responses can be rewritten as

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] = \sqrt{\frac{q}{p_t}} \cdot \frac{\sum_{k=0}^{n-1} \sqrt{B_k B_{k+1}}}{1 - p_t^n}. \tag{124}$$

Therefore,

$$\Delta^+(p_t, G) = \sqrt{\frac{q}{p_t}} \left(\frac{\sum_{k=0}^{n-1} \sqrt{B_k B_{k+1}}}{1 - p_t^n} - 1 \right). \tag{125}$$

Since $\sqrt{q/p_t} > 0$, the sign of $\Delta^+(p_t, G)$ is determined by the term inside the parentheses.

By the Cauchy–Schwarz inequality,

$$\left(\sum_{k=0}^{n-1} \sqrt{B_k B_{k+1}} \right)^2 \leq \left(\sum_{k=0}^{n-1} B_k \right) \left(\sum_{k=0}^{n-1} B_{k+1} \right). \tag{126}$$

Moreover,

$$\sum_{k=0}^{n-1} B_k = 1 - p_t^n, \tag{127}$$

and

$$\sum_{k=0}^{n-1} B_{k+1} = 1 - q^n. \tag{128}$$

Thus,

$$\frac{\sum_{k=0}^{n-1} \sqrt{B_k B_{k+1}}}{1 - p_t^n} \leq \sqrt{\frac{1 - q^n}{1 - p_t^n}}. \tag{129}$$

When $p_t < 0.5$, we have

$$q = 1 - p_t > p_t, \tag{130}$$

Consequently, we can derive that:

$$\Delta^+(p_t, G) < 0, \qquad p_t < 0.5. \tag{131}$$

Equivalently,

$$p_t < 0.5 \quad \Longrightarrow \quad \mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 1\right] < A^+_{t,i}. \tag{132}$$

Thus, for difficult prompts with $p_t < 0.5$, the advantage of correct responses is provably underestimated. However, for $p_t > 0.5$, the above argument does not imply

$$\Delta^+(p_t, G) > 0. \tag{133}$$

Instead, due to the retention condition S, the zero point is shifted to the right of 0.5. Since

$$\Delta^+(0.5, G) < 0 \tag{134}$$

for $G > 2$, while

$$\lim_{p_t \to 1^-} \Delta^+(p_t, G) = \frac{1}{\sqrt{G - 1}} > 0, \tag{135}$$

there exists at least one zero point

$$p_G^+ \in (0.5, 1) \tag{136}$$

such that

$$\Delta^+(p_G^+, G) = 0. \tag{137}$$

Therefore, the estimator may remain downward biased even for moderately easy prompts with $p_t > 0.5$, and the bias becomes positive only when p_t is sufficiently large.

Similarly, for an incorrect response, we can derive

$$\mathbb{E}\left[\hat{A}_{t,i} \mid S, r_{t,i} = 0\right] = -\frac{\sum_{k=1}^{G-1} \binom{G-1}{k} p_t^k (1-p_t)^{G-1-k} \sqrt{\frac{k}{G-k}}}{1-(1-p_t)^{G-1}}. \tag{138}$$

The bias pattern for incorrect responses mirrors that for correct responses. Since the true advantage of an incorrect response is negative, a positive bias indicates that the estimated advantage is less negative than the true one, meaning that the penalty is underestimated. Thus, for easy prompts, incorrect responses tend to be insufficiently penalized, whereas for sufficiently difficult prompts, they may be over-penalized. This behavior is symmetric to the correct-response case, with the corresponding zero point shifted to the opposite side of 0.5. Summarizing the above derivation, we can obtain the following conclusion.

Summarizing the above derivation, we show that standard-deviation normalization rescales but does not debias group-relative advantage estimation. The normalized estimator remains affected by the same retention condition as the unnormalized estimator, leading to difficulty-dependent bias for both correct and incorrect responses. In particular, correct responses are underestimated on difficult prompts and overestimated in sufficiently easy regimes, while incorrect responses exhibit a symmetric bias pattern. Therefore, standard-deviation normalization cannot eliminate the intrinsic conditioning bias of group-relative advantage estimation.

F.9 Non-binary Reward Analysis

In this section, we extend our analysis to *continuous bounded reward distributions* (e.g., Beta and truncated Gaussian scores), which better reflect the behavior of soft verifiers and learned reward models commonly used in practice. The conditional mean estimation error equals $p_t - \mathbb{E}[\hat{p}_t \mid S_\sigma]$. It is nonzero whenever conditioning on S_σ changes the expected group mean. Its magnitude and direction depend on the reward distribution and the update-selection event. Next, we show the main results.

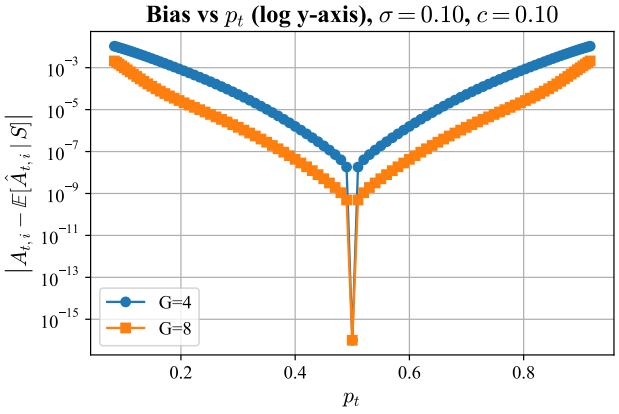

Figure 5: Illustration of advantage bias under truncated Gaussian rewards for different group sizes.

Theorem 6. *At training step t and let $G \geq 2$, with CDF F and PDF f. Given a prompt $x_t \sim \mathcal{D}$ and draw $G \geq 2$ i.i.d. rewards:*

$$r_{t,1}, \ldots, r_{t,G} \overset{\text{i.i.d.}}{\sim} \mathcal{D}(p_t). \tag{139}$$

And we extend the binary reward setting to non-binary rewards:

$$r_{t,i} \in \{0,1\} \to r_{t,i} \in [0,1]. \tag{140}$$

The group-relative advantage can be denoted as:

$$\hat{A}_{t,i} := r_{t,i} - \hat{p}_t, \quad \hat{p}_t = \frac{1}{G}\sum_{i=1}^{G} r_{t,i}, \tag{141}$$

while the expected advantage is defined as:

$$A_{t,i} \coloneqq r_{t,i} - p_t. \tag{142}$$

Fix a constant $\sigma \in [0,1]$ and define the update event:

$$S_\sigma := \big\{\exists\, i \neq j :\ |r_{t,i} - r_{t,j}| > \sigma\big\} \Rightarrow S_\sigma^c = \big\{\max_i r_{t,i} - \min_i r_{t,i} \leq \sigma\big\}. \tag{143}$$

For $u \in [0,1]$, define $u^+ := \min\{1, u+\sigma\}$, we have:

$$q(u) := F(u^+) - F(u), \tag{144}$$

and:

$$m(u) := \mathbb{E}[r_{t,1} \mid u \leq r_{t,1} \leq u^+] = \frac{\int_u^{u^+} x f(x)\,dx}{F(u^+) - F(u)} \quad \textit{(when } q(u) > 0\textit{)}. \tag{145}$$

Then the probability of a non-update *is:*

$$\mathbb{P}(S_\sigma^c) = G \int_0^1 f(u)\, q(u)^{G-1}\, du, \tag{146}$$

and:

$$\mathbb{P}(S_\sigma) = 1 - \mathbb{P}(S_\sigma^c). \tag{147}$$

Moreover, we have:

$$\mathbb{E}[\hat{p}_t \mid S_\sigma] = \frac{p_t - \mathbb{E}[\hat{p}_t \cdot \mathbf{1}_{\{S_\sigma^c\}}]}{\mathbb{P}(S_\sigma)} \tag{148}$$

with:

$$\mathbb{E}[\hat{p}_t \cdot \mathbf{1}_{\{S_\sigma^c\}}] = \int_0^1 \big(u + (G-1)m(u)\big)\, f(u)\, q(u)^{G-1}\, du. \tag{149}$$

Finally, the conditional bias transferred to advantages satisfies, for all i, we have:

$$\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma] = p_t - \mathbb{E}[\hat{p}_t \mid S_\sigma]. \tag{150}$$

Proof. The complement event can be denoted as:

$$S_\sigma^c = \{\max - \min \leq \sigma\} \tag{151}$$

For absolutely continuous i.i.d. samples, the minimum has density:

$$g_{\min}(u) = G f(u) \big(1 - F(u)\big)^{G-1}. \tag{152}$$

Condition on $\min = u$. The remaining $G-1$ samples are i.i.d. with the original law conditioned on $[u,1]$; imposing $\max \leq u^+$ is equivalent to requiring each of those samples lies in $[u, u^+]$. Thus:

$$\mathbb{P}(S_\sigma^c \mid \min = u) = \Big(\frac{F(u^+) - F(u)}{1 - F(u)}\Big)^{G-1}, \tag{153}$$

and multiplying by $g_{\min}(u)$ gives:

$$\mathbb{P}(S_\sigma^c) = G \int_0^1 f(u)\, q(u)^{G-1} du. \tag{154}$$

On S_σ^c and $\min = u$, one sample equals the minimum and the other $G-1$ samples lie in $[u, u^+]$. By symmetry, the conditional mean of each of the $G-1$ non-minimum samples is $m(u)$, hence:

$$\mathbb{E}\Big[\sum_{i=1}^G r_{t,i} \;\Big|\; S_\sigma^c, \min = u\Big] = u + (G-1)m(u). \tag{155}$$

So we can derive:

$$\begin{aligned} \mathbb{E}[\hat{p}_t \cdot \mathbf{1}_{\{S_\sigma^c\}}] &= \int_0^1 \frac{u + (G-1)m(u)}{G}\, d\mathbb{P}(\min \in du, S_\sigma^c) \\ &= \int_0^1 (u + (G-1)m(u)) f(u) q(u)^{G-1} du. \end{aligned} \tag{156}$$

Corollary 6. *For* $\mathrm{Beta}(\alpha, \beta)$ *reward distribution, the Beta density is:*

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}, \tag{157}$$

and the CDF is:

$$F(x) = I_x(\alpha, \beta) \quad \textit{for } x \in [0, 1], \tag{158}$$

where $B(\cdot, \cdot)$ *is the Beta function and* $I_x(\alpha, \beta)$ *is the regularized incomplete beta function. In particular:*

$$p_t = \mathbb{E}[r_{t,1}] = \frac{\alpha}{\alpha + \beta}. \tag{159}$$

Moreover, we have:

$$q(u) = F(u^+) - F(u) = I_{u^+}(\alpha, \beta) - I_u(\alpha, \beta), \tag{160}$$

and the conditional mean over $[u, u^+]$ *admits the closed form:*

$$m(u) = \frac{\int_u^{u^+} x f(x)\, dx}{\int_u^{u^+} f(x)\, dx} = \frac{B_{u^+}(\alpha+1, \beta) - B_u(\alpha+1, \beta)}{B_{u^+}(\alpha, \beta) - B_u(\alpha, \beta)}, \tag{161}$$

where $B_x(\cdot, \cdot)$ *denotes the (unregularized) incomplete beta function.*

Consequently, substituting F, f, q, m into conclusions obtained earlier yields explicit one-dimensional integral expressions (in standard special functions) for $\mathbb{P}(S_\sigma^c)$ and $\mathbb{E}[\hat{p}_t \mid S_\sigma]$.

Corollary 7. *Let the reward* $Z_{t,1}, \ldots, Z_{t,G}$ *be i.i.d.* $\mathcal{N}(\mu, \xi^2)$ *with* $\xi > 0$*, and define* $r_{t,i}$ *to be* properly truncated *to* $[0, 1]$*, i.e.* $r_{t,i}$ *has the conditional law:*

$$r_{t,i} \overset{d}{=} Z_{t,i} \,\big|\, (0 \le Z_{t,i} \le 1), \qquad i = 1, \ldots, G. \tag{162}$$

Let $u^+ := \min\{1, u + c\}$ *and define, for* $u \in [0, 1]$ *with* $q(u) > 0$*, we have:*

$$q(u) := \mathbb{P}\big(u \le r_{t,1} \le u^+\big), \tag{163}$$

and:

$$m(u) := \mathbb{E}\big[r_{t,1} \mid u \le r_{t,1} \le u^+\big]. \tag{164}$$

Let Φ *and* φ *be the standard normal CDF and PDF, and set:*

$$a := \frac{0 - \mu}{\xi}, \qquad b := \frac{1 - \mu}{\xi}. \tag{165}$$

Then the truncated-normal density on $[0, 1]$ *is:*

$$f(x) = \frac{\varphi\Big(\frac{x-\mu}{\xi}\Big)}{\sigma\big(\Phi(b) - \Phi(a)\big)}\, \mathbf{1}_{[0,1]}(x). \tag{166}$$

Its CDF on $[0, 1]$ *is:*

$$F(x) = \frac{\Phi\Big(\frac{x-\mu}{\xi}\Big) - \Phi(a)}{\Phi(b) - \Phi(a)}. \tag{167}$$

The mean satisfies:

$$p_t = \mathbb{E}[r_{t,1}] = \mu + \xi\, \frac{\varphi(a) - \varphi(b)}{\Phi(b) - \Phi(a)}. \tag{168}$$

Moreover:

$$q(u) = F(u^+) - F(u), \tag{169}$$

and the conditional mean over $[u, u^+]$ *has the standard truncated-normal form:*

$$m(u) = \mu + \sigma\, \frac{\varphi\Big(\frac{u-\mu}{\xi}\Big) - \varphi\Big(\frac{u^+-\mu}{\xi}\Big)}{\Phi\Big(\frac{u^+-\mu}{\xi}\Big) - \Phi\Big(\frac{u-\mu}{\xi}\Big)}. \tag{170}$$

Consequently, substituting F, f, q, m *to yield explicit one-dimensional integral expressions for* $\mathbb{P}(S_\sigma^c)$ *and* $\mathbb{E}[\hat{p}_t \mid S_\sigma]$ *in terms of* Φ *and* φ*.*

Remark. Theoretical and numerical evaluations under properly truncated Gaussian reward distributions indicate that the group-relative advantage bias, $\left|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma]\right|$, tends to increase as p_t deviates further from $1/2$. Intuitively, near these extremes, a typical group of samples exhibits reduced dispersion. As a result, conditioning on the non-degenerate event S_σ preferentially selects groups with atypically large variability, which can shift the conditional expected group reward away from p_t and produce a nonzero conditional mean estimation error.

Figure 5 illustrates two representative cases corresponding to group sizes $G = 4$ and $G = 8$, as predicted by Corollary 7. In both settings, the magnitude of the bias $|\mathbb{E}[\hat{A}_{t,i} - A_{t,i} \mid S_\sigma]|$ increases as p_t moves farther away from 0.5, corroborating our theoretical analysis.

G Supplementary Experiments

G.1 Advantage Distribution

Figure 1(b) compares the same prompts under 8 and 128 rollouts, using the latter as a lower-variance proxy for prompt difficulty. Small-group estimation underweights rare correct responses on hard prompts and under-penalizes rare incorrect responses on easy prompts. And we conducted an assessment of select prompts from the widely used training dataset MATH and DAPO-Math-17k [48] on Qwen3-4B-Base across different rollouts. At first, we evaluated the model's performance on the dataset at rollout=8. From these, we selected four groups of 50 prompts each: groups with single correct or incorrect response. We then evaluated the outcomes of these selected prompts at rollout=128 where enough rollouts can reflect intrinsic difficulty of these prompts.

For those groups with only 1 correct responses at rollout=8, the distribution of the number of correct responses within these groups is shown in Figure 6(a). For the MATH and DAPO-Math-17k datasets, 24 and 15 groups have fewer than 16 correct responses at rollout=128 respectively which suggests that the advantage of correct responses for these prompts are underestimated at rollout=8. And these distinct responses in these most challenging prompts are crucial for pushing the model's capability frontier, requiring more exploration. Similarly, for prompts with 1 incorrect answer at rollout=8 and we find that 12 and 21 groups have less than 16 incorrect responses with 128 rollouts on MATH and DAPO-Math-17k which may lead to over-exploitation as Figure 6(b).

G.2 Ablation Study on G

It is a widely accepted consensus that increasing the number of rollouts effectively mitigates estimation bias [45]. As the group size grows, the empirical advantage distribution converges closer to the true advantage distribution, thereby reducing the variance and bias inherent in the advantage estimation of group-relative RL algorithms. To rigorously validate the effectiveness of our dynamic adjusting approach HA-DW in mitigating estimation bias under constrained sampling conditions, we conducted a comparative analysis of model's training performance across varying rollout sizes. The results presented in Table 6 shows that increasing the number of rollouts can, to a certain extent, enhance model performance by providing a more stable baseline. Although scaling up the number of rollouts is a straightforward method to improve performance, its benefits are often capped by computational constraints. As the rollout size increases, the training time grows substantially and out-of-memory (OOM) will occur once it exceeds a certain threshold. Our method offers a more efficient alternative: dynamic advantage adjustment demonstrates superior efficacy even under the limited rollouts condition, and it effectively mitigates the estimation bias that typically plagues low-sample scenarios, achieving robust performance without the need for extensive sampling.

G.3 Ablation Study on λ_{scale}

As illustrated in Section F.7,there exists a specific scaling factor λ_{scale} satisfying Equation (103) to compensate biased advantage estimation. Although the theoretically optimal range of λ_{scale} may vary with prompt difficulty, the prompt-dependent variation is already captured by the difficulty-aware term, while the dynamic difficulty anchor accounts for changes during training. Therefore, λ_{scale} mainly serves as a global calibration factor controlling the overall magnitude of reweighting rather

than the relative adjustment across prompts. Table 10 demonstrates the performance of RL training under different values of λ_{scale}. When $\lambda_{\text{scale}} = 1.3$ or 1.5, the trained model achieves the best performance across five mathematical benchmarks. The results correspond to our analysis that there exists an optimal value that balances the adjustment across prompts of varying difficulties, thereby enhancing RL training performance. The consistently strong performance around 1.3–1.5 also suggests that precise prompt-level tuning of λ_{scale} is not required in practice.

G.4 Experiments of Different Domains

We further extend our experiments to instruction following and tool use. The results, presented in Table 11, show that HA-DW delivers robust improvements across diverse RLVR domains.

Dataset	8	16	8+HA-DW
MATH500	75.4	76.2	**78.0**
AIME25	19.6	19.2	**20.4**
AMC23	60.3	61.6	**63.4**
Minerva	33.8	34.2	**36.8**
OlympiadBench	43.5	43.9	**44.7**

Table 6: Performance of Qwen3-4B-Base trained with: Rollout=8 with GRPO, Rollout=16 with GRPO and Rollout=8 with GRPO+HA-DW. *Rollout=32 with GRPO is out of memory*.

G.5 Experiments on Different Group Sizes

We further evaluate the performance on five mathematical benchmarks with group sizes $G = 8$ and $G = 16$, and further apply HA-DW under $G = 8$. The results demonstrate in Table 6 show that our method substantially outperforms the gains obtained by merely increasing the group size.

G.6 Effect on Easy Prompts

For easy prompts, HA-DW is not primarily intended to further improve the accuracy of already-solved examples. Instead, it corrects distorted update signals that may otherwise over-reinforce correct responses and insufficiently penalize rare incorrect ones, thereby reducing unnecessary exploitation of easy prompts and reallocating learning signals toward more challenging examples. To further examine this effect, we compare the full HA-DW with a variant that applies reweighting only to hard prompts. As shown in Table 7, the full HA-DW achieves a higher average performance than the hard-prompt-only variant, suggesting that correcting the update distortion on easy prompts also contributes to the overall performance gains.

G.7 Ablation Study on Rollout Token Budget

As HA-DW may induce slightly longer responses, we further examine whether its performance gains can be attributed to additional rollout-generation compute. We train GRPO and GRPO+HA-DW with the same group size, maximum generation length, and a matched total rollout budget of 120M generated response tokens, stopping training once the budget is reached. As shown in Table 8, HA-DW consistently outperforms GRPO under the same rollout-token budget, improving the average accuracy from 46.5 to 48.0. This result suggests that the observed gains cannot be explained solely by longer reasoning trajectories or additional rollout-generation compute.

G.8 Robustness to Reward Noise

Since HA-DW increases the influence of hard prompts, we further evaluate its robustness when the reward signals of hard prompts are noisy. Specifically, we randomly select a subset of hard prompts and flip their binary verifier rewards during training, while keeping all evaluation labels and verifier signals unchanged. We compare GRPO and GRPO+HA-DW with and without injected reward noise on Qwen3-4B-Base. As shown in Table 9, reward noise degrades the performance of both methods. Nevertheless, HA-DW retains a modest advantage over GRPO under the noisy setting, achieving an average score of 44.1 compared with 43.5. These results suggest that HA-DW remains effective under moderate reward noise, while its performance is still sensitive to the reliability of reward signals on hard prompts.

Table 7: Ablation of HA-DW on easy prompts using Qwen3-4B-Base. The hard-prompt-only variant applies HA-DW reweighting only to hard prompts, while leaving easy-prompt updates unchanged.

Method	MATH500	AIME25	AMC23	Minerva	OlympiadBench	AVG
GRPO	75.4	19.6	60.3	33.8	43.5	46.5
GRPO + HA-DW (Hard Prompts Only)	77.4	20.6	62.8	35.7	44.3	48.2
GRPO + HA-DW	78.0	20.4	63.4	36.8	44.7	48.7

Table 8: Performance under a matched rollout-token budget. GRPO and GRPO+HA-DW use the same group size, maximum generation length, and total budget of 120M generated response tokens.

Method	Rollout Token Budget	MATH500	AIME25	AMC23	Minerva	OlympiadBench	AVG
GRPO	120M	75.4	19.6	60.3	33.8	43.5	46.5
GRPO + HA-DW	120M	77.4	19.8	62.7	35.9	44.3	48.0

G.9 Significance Analysis

We conduct a significance analysis on Qwen3-4B-Base across five independent runs and five mathematical benchmarks. The results is demonstrated in Table 12, showing that HA-DW robustly improves over its corresponding baselines on the evaluated benchmarks. The gains remain stable across random seeds, exhibiting low run-to-run variance, and are statistically significant with all p-values below 0.01. Moreover, HA-DW demonstrates robust effectiveness across different RLVR algorithms, including GRPO, GSPO, and DAPO, indicating that its benefits are reliable, broadly applicable, and not limited to a specific benchmark or algorithmic setting.

H Hard Evolving Difficulty Anchor

To simplify the update process of evolving belief C_t, thereby reducing algorithmic complexity. The synchronization of the model's state can be facilitated through a hard update mechanism, executed at every training step. Let h be a hyperparameter denoting the number of most recent training rounds considered. Let h be hyper-paramter to represent the last h training rounds. The Equation (15) can be rewritten as:

$$C_t^+ = \frac{h-1}{h} C_t^- + \frac{1}{h} b_t = \frac{1}{h} \left(\sum_{j=1}^{h-1} b_{t-j} + b_t \right), \tag{171}$$

which indicates that the belief update is effectuated by directly synthesizing the accuracy information derived from the preceding h batches with observations from the current iteration, and we leave the remaining update procedures intact. Although this formulation ignores short-term oscillations in belief updates, it significantly simplifies the overall algorithm.

I Prompt

Prompt Template

{question}. Let's think step by step and output the final answer within \\boxed{}.

J Broader Impact

This work identifies and mitigates a fundamental bias in group-relative advantage estimation for post-training large language models. By improving the stability and generalization of reasoning-oriented training, HA-DW may support more reliable LLMs for scientific research and complex problem solving. Potential negative impacts are generally aligned with those of advanced LLM reasoning technologies, including misuse and unsafe deployment. We therefore emphasize the importance of fair, safe, and responsible deployment practices.

Table 9: Robustness of HA-DW under injected reward noise on hard prompts using Qwen3-4B-Base.

Method	MATH500	AIME25	AMC23	Minerva	OlympiadBench	AVG
GRPO	75.4	19.6	60.3	33.8	43.5	46.5
GRPO + Reward Noise	72.6	16.7	56.9	30.9	40.6	43.5
GRPO + HA-DW	78.0	20.4	63.4	36.8	44.7	48.7
GRPO + HA-DW + Reward Noise	73.4	16.3	58.1	31.8	40.9	44.1

λ_{scale}	**MATH500**	**AIME25**	**AMC23**	**Minerva**	**OlympiadBench**	**AVG**
0.5	75.4	18.1	61.1	34.2	43.7	46.5
0.8	76.8	19.2	61.3	34.9	43.7	47.2
1.0	76.8	18.5	61.6	36.0	44.3	47.4
1.3	**78.0**	20.4	**63.4**	36.8	**44.7**	**48.7**
1.5	77.8	**20.8**	63.1	**37.1**	44.0	48.6
1.7	76.4	20.0	**63.4**	36.4	44.3	48.1
2.0	76.8	19.0	61.9	35.3	43.5	47.3

Table 10: Performance of Qwen3-4B-Base trained with GRPO+HA-DW on different λ_{scale}.

K Case Study

This appendix demonstrates some output examples generated by policy models trained with GRPO and GRPO+HA-DW. And the results are shown in Figure 7 and Figure 8.

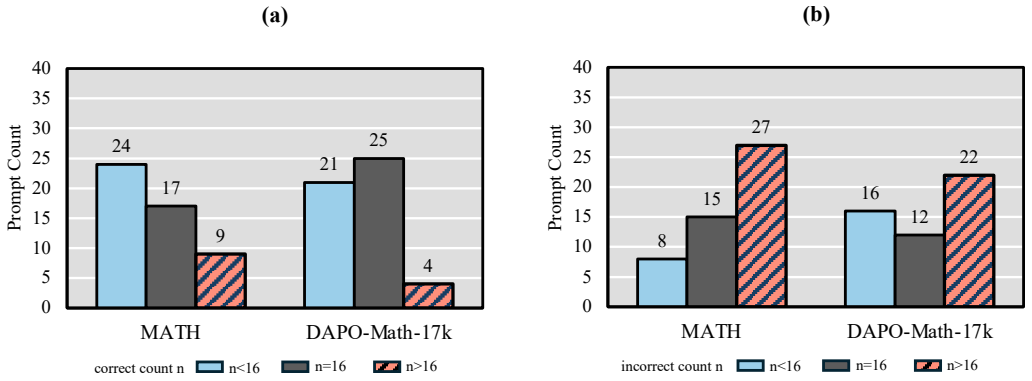

Figure 6: The distribution of prompts by the number of correct and incorrect responses on the MATH dataset and DAPO-Math-17k under 8 and 128 rollouts of Qwen3-4B-Base.

Algorithm	IFEval	GPQA	API-Bank	AVG
Qwen3-4B-Base				
GRPO	32.3	26.2	62.1	40.2
GRPO+HA-DW	34.0	27.8	63.6	41.8
GSPO	32.0	26.8	61.2	40.0
GSPO+HA-DW	33.1	26.3	60.7	40.0
DAPO	31.1	27.2	62.6	40.3
DAPO+HA-DW	33.8	29.3	64.5	42.5
Qwen3-8B-Base				
GRPO	39.4	33.8	69.6	47.6
GRPO+HA-DW	41.6	32.2	72.4	48.7
GSPO	39.0	34.3	70.1	47.8
GSPO+HA-DW	41.9	35.4	72.0	49.8
DAPO	40.5	33.3	71.0	48.3
DAPO+HA-DW	42.1	36.4	73.4	50.6

Table 11: Performance comparison on instruction following and tool-use benchmarks.

Benchmark	GRPO (mean±std, N=5)	HA-DW (mean±std, N=5)	Average Improvement	P-Value
MATH500	75.44 ± 0.30	77.88 ± 0.23	+2.44	3.08×10^{-5}
AIME25	19.04 ± 0.61	20.44 ± 0.30	+1.40	7.59×10^{-3}
AMC23	60.42 ± 0.34	63.58 ± 0.41	+3.16	6.71×10^{-6}
Minerva	33.68 ± 0.41	36.78 ± 0.25	+3.10	4.33×10^{-4}
OlympiadBench	43.42 ± 0.44	44.88 ± 0.20	+1.46	1.20×10^{-3}

Table 12: Significance analysis of HA-DW against GRPO on Qwen3-4B-Base over five independent runs.

	Hyperparameter	GRPO	GRPO+HA-DW	GSPO	GSPO+HA-DW	DAPO	DAPO+HA-DW
General	nnode	1	1	1	1	1	1
	gpus per node	8	8	8	8	8	8
	use kl in reward	False	False	False	False	False	False
	use kl loss	False	False	False	False	False	False
	tensor parallel size	1	1	1	1	1	1
	test frequency	5	5	5	5	5	5
Training	train batch size	256	256	256	256	256	256
	mini batch size	16	16	16	16	16	16
	micro batch size	4	4	4	4	4	4
	epoch	3	3	3	3	9	9
	gradient clip	1.0	1.0	1.0	1.0	1.0	1.0
	optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
	warmup steps	10	10	10	10	10	10
	weight decay	0.1	0.1	0.1	0.1	0.1	0.1
	learning rate	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}	1×10^{-6}
Clipping	clip-high	0.2	0.2	0.0004	0.0004	0.28	0.28
	clip-low	0.2	0.2	0.0003	0.0003	0.2	0.2
Rollout	max prompt length	1024	1024	1024	1024	1024	1024
	max response length	4096	4096	4096	4096	4096	4096
	rollout.n	8	8	8	8	8	8
	do sample	False	False	False	False	False	False
	filtering	False	False	False	False	False	False
	dynamic batch size	True	True	True	True	True	True

Table 13: Hyperparameter settings for group-relative methods.

Case Study

Question:

In the xy-plane, a circle of radius 4 with center on the positive x-axis is tangent to the y-axis at the origin, and a circle with radius 10 with center on the positive y-axis is tangent to the x-axis at the origin. What is the slope of the line passing through the two points at which these circles intersect? The final answer can be written in the form $\frac{m}{n}$, where m and n are relatively prime positive integers. What is $m+n$? Let's think step by step and output the final answer within \\ boxed{}.

Answer: 7

GRPO Solution:

To solve the problem, we need to find the points of intersection of the two circles and then determine the slope of the line passing through these points.

1. Equation of the Circles: The first circle has a radius of 4 and is centered on the positive x-axis. Since it is tangent to the y-axis at the origin, its center is at $(4, 0)$. The equation of this circle is: $(x-4)^2 + y^2 = 16$. The second circle has a radius of 10 and is centered on the positive y-axis. Since it is tangent to the x-axis at the origin, its center is at $(0, 10)$. The equation of this circle is: $x^2 + (y-10)^2 = 100$.

2. Finding the Points of Intersection: We need to solve the system of equations:

$$\begin{cases} (x-4)^2 + y^2 = 16 \\ x^2 + (y-10)^2 = 100 \end{cases}$$

Expanding both equations, we get:

$$\begin{cases} x^2 - 8x + 16 + y^2 = 16 \implies x^2 - 8x + y^2 = 0 \\ x^2 + y^2 - 20y + 100 = 100 \implies x^2 + y^2 - 20y = 0 \end{cases}$$

Subtract from the second equation: $(x^2 + y^2 - 20y) - (x^2 - 8x + y^2) = 0 \implies -20y + 8x = 0 \implies 8x = 20y \implies x = \frac{5y}{2}$.

Substitute $x = \frac{5y}{2}$ into the first equation : $\left(\frac{5y}{2}\right)^2 - 8\left(\frac{5y}{2}\right) + y^2 = 0 \implies \frac{25y^2}{4} - 20y + y^2 = 0 \implies \frac{25y^2+4y^2}{4} - 20y = 0 \implies \frac{29y^2}{4} - 20y = 0\ 29y^2 - 80y = 0 \implies y(29y - 80) = 0$ So, $y = 0$ or $y = \frac{80}{29}$. If $y = 0$, then $x = \frac{5 \cdot 0}{2} = 0$, which is the origin and not a point of intersection of the two circles. If $y = \frac{80}{29}$, then $x = \frac{5 \cdot \frac{80}{29}}{2} = \frac{200}{29}$.

So, the points of intersection are $\left(\frac{200}{29}, \frac{80}{29}\right)$ and $(0, 0)$, but we already know $(0, 0)$ is not a valid intersection point for the two circles. Therefore, the only valid intersection point is $\left(\frac{200}{29}, \frac{80}{29}\right)$.

3. Finding the Slope: The other point of intersection can be found by symmetry or by solving the equations again, but we know the circles are symmetric about the line $y = x$. So, the other point of intersection is $\left(\frac{80}{29}, \frac{200}{29}\right)$. The slope of the line passing through the points $\left(\frac{200}{29}, \frac{80}{29}\right)$ and $\left(\frac{80}{29}, \frac{200}{29}\right)$ is: slope $= \frac{\frac{200}{29} - \frac{80}{29}}{\frac{80}{29} - \frac{200}{29}} = \frac{\frac{120}{29}}{-\frac{120}{29}} = -1$. So, the slope is -1, and in the form $\frac{m}{n}$, we have $m = -1$ and $n = 1$.

Therefore, $m + n = -1 + 1 = 0$.

The final answer is: $\boxed{-1}$

Figure 7: An example of GRPO

Case Study

GRPO+HA-DW Solution:

To solve the problem, we need to find the points of intersection of the two circles and then determine the slope of the line passing through these points.

1. Equation of the first circle:
The first circle has a radius of 4 and is centered on the positive x-axis. Since it is tangent to the y-axis at the origin, the center of the circle is at $(4, 0)$. The equation of the first circle is: $(x-4)^2 + y^2 = 16$.

2. Equation of the second circle:
The second circle has a radius of 10 and is centered on the positive y-axis. Since it is tangent to the x-axis at the origin, the center of the circle is at $(0, 10)$. The equation of the second circle is: $x^2 + (y-10)^2 = 100$.

3. Find the points of intersection: We need to solve the system of equations:

$$\begin{cases} (x-4)^2 + y^2 = 16 \\ x^2 + (y-10)^2 = 100 \end{cases}$$

First, expand both equations:

$$\begin{cases} x^2 - 8x + 16 + y^2 = 16 \implies x^2 - 8x + y^2 = 0 \\ x^2 + y^2 - 20y + 100 = 100 \implies x^2 + y^2 - 20y = 0 \end{cases}$$

Subtract the first expanded equation from the second: $(x^2 + y^2 - 20y) - (x^2 - 8x + y^2) = 0 - 0$ $-20y + 8x = 0 \implies 8x = 20y \implies x = \frac{5}{2}y$.

Substitute $x = \frac{5}{2}y$ into the first expanded equation: $\left(\frac{5}{2}y\right)^2 - 8\left(\frac{5}{2}y\right) + y^2 = 0$. This gives us two solutions for y: $y = 0$ or $\frac{29}{4}y = 20 \implies y = \frac{80}{29}$.

For $y = 0$: $x = \frac{5}{2}(0) = 0$. So, one intersection point is $(0, 0)$, which is the origin. For $y = \frac{80}{29}$: $x = \frac{5}{2}\left(\frac{80}{29}\right) = \frac{200}{29}$. So, the other intersection point is $\left(\frac{200}{29}, \frac{80}{29}\right)$.

4. Find the slope of the line passing through the points $(0, 0)$ and $\left(\frac{200}{29}, \frac{80}{29}\right)$: The slope m of the line passing through two points (x_1, y_1) and (x_2, y_2) is given by: $m = \frac{y_2 - y_1}{x_2 - x_1}$ Substituting the points $(0, 0)$ and $\left(\frac{200}{29}, \frac{80}{29}\right)$: $m = \frac{\frac{80}{29} - 0}{\frac{200}{29} - 0} = \frac{\frac{80}{29}}{\frac{200}{29}} = \frac{80}{200} = \frac{2}{5}$.

Therefore, the slope of the line is $\frac{2}{5}$, and $m + n = 2 + 5 = 7$.

The final answer is $\boxed{7}$.

Figure 8: An example of GRPO+HA-DW

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are aligned with the paper's theoretical and empirical contributions. Specifically, we clearly state the scope of our analysis, the assumptions under which the bias characterization holds, and the experimental settings used to validate the proposed method.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the dedicated Limitations section. We clarify that HA-DW targets biased advantage estimation in group-relative RL and is therefore mainly scoped to group-wise optimization methods. We also note that, although HA-DW may apply beyond standard RLVR, our analysis and experiments focus on RLVR reasoning tasks, leaving broader domains for future work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We state the assumptions for each theoretical result and provide complete derivations and proofs in the main paper and appendix. The theoretical results characterize the finite-sample bias of group-relative advantage estimation under the specified sampling setting, and the corresponding formulas, propositions, and proofs are numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the experimental settings needed to reproduce the main results, including the evaluated benchmarks, model backbones, group-relative training methods, reward definitions, sampling configurations, optimization hyperparameters, and evaluation protocols. We also report results across multiple runs where applicable, together with implementation details in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymized repository containing the code, data preparation instructions, environment setup, and commands needed to reproduce the main experimental results. The repository is anonymized for the submission stage and includes sufficient documentation to support faithful reproduction.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings necessary to understand the results, including the training and evaluation datasets, model backbones, data splits or evaluation protocols, optimization settings, hyperparameters, reward definitions, and group-wise sampling configurations. Additional training and implementation details are provided in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report statistical significance information for the main experiments, including results over multiple runs with mean and standard deviation where applicable, as well as significance tests and corresponding P-values. These analyses are used to assess whether the observed improvements of HA-DW over group-relative baselines are statistically reliable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the GPU resources used for the experiments, but does not provide complete information about memory usage, execution time, or total compute required for reproducing each experiment.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research conducted in this paper conforms to it. The work does not involve human subjects, private or sensitive personal data, or deceptive data collection, and we have taken care to preserve anonymity in the submission.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The Broader Impacts section discusses potential positive impacts, such as supporting more reliable LLMs for scientific research and complex problem solving, as well as potential negative impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release any high-risk models, pretrained language models, image generators, or scraped datasets. The work focuses on theoretical analysis and methodological improvements for group-relative reinforcement learning, so no additional safeguards for responsible release are applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use publicly available open-source assets, including benchmark datasets, model backbones, and relevant codebases, and properly cite their original sources in the paper. We respect the corresponding licenses and terms of use for these assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce or release new assets such as datasets, model checkpoints, or standalone code packages. Therefore, asset documentation requirements are not applicable to this work.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing experiments, human-subject studies, or participant-based data collection. Therefore, participant instructions, screenshots, and compensation details are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This work does not involve crowdsourcing experiments or research with human subjects. Therefore, IRB approval, participant risk disclosure, and related review procedures are not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing assistance, editing, and language polishing, and did not contribute to the core methodology, scientific claims, experiments, or theoretical results of this research. Therefore, declaration of LLM usage is not applicable under the checklist guideline.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.