COPC: COUPLED OFF-POLICY CORRECTION FOR ASYNCHRONOUS LLM REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Asynchronous RL improves throughput in large language model post-training by decoupling rollout generation from optimization, at the cost of training on stale trajectories. Existing methods primarily address token-level policy mismatch through importance-ratio control in the actor objective, providing *policy-side correction*. We show that this alone is insufficient because advantage estimates inherit mismatch from behavior-policy continuations, a phenomenon we term *advantage staleness*. This motivates a complementary *advantage-side correction*, raising the question: how do the two correction channels interact, and how should their strengths be coordinated? We derive exact bias and variance decompositions for a general two-channel actor update, revealing a nonseparable coupling between policy-weight and advantage-estimation errors: their interaction induces multiplicative bias terms, while squared policy weights amplify advantage uncertainty in gradient variance. This coupling motivates a testable hypothesis that the preferred settings of the two correction channels should be coordinated rather than chosen independently. Guided by this analysis, we introduce **C**oupled **O**ff-**P**olicy **C**orrection (**COPC**), an actor–critic method combining token-level ratio masking with a dedicated advantage-side correction based on two-sided clipped-ratio weighting of TD residuals for return and advantage estimation. Joint parameter sweeps across staleness levels provide empirical support for this hypothesis: the effect of varying one correction parameter depends on, and can even reverse with, the setting of the other. COPC achieves the highest reported performance on both tool-integrated mathematical reasoning and search, outperforming the strongest reported asynchronous baseline in each setting. COPC further exhibits strong practical robustness, with a broad high-performing parameter region and substantially improved training stability. In particular, in the search setting, it remains stable throughout training, while most evaluated asynchronous baselines exhibit late-stage performance collapse. These performance gains persist at 64-step policy staleness, while COPC incurs minimal step-time overhead over asynchronous PPO and retains a $1.7\times$ step-time speedup over synchronous PPO.

1 INTRODUCTION

Reinforcement learning (RL) has become a central paradigm for post-training large language models (LLMs) with preference feedback and verifiable rewards (Ouyang et al., 2022; Shao et al., 2024; Guo et al., 2025). Scaling RL, however, is often limited by costly rollouts whose completion times vary with response length and tool interactions, leading to substantial idle time under synchronous training. Asynchronous RL alleviates this bottleneck by decoupling rollout generation from policy optimization, allowing both stages to proceed concurrently (Wu et al., 2025; Zhu et al., 2025; Fu et al., 2025). This overlap improves hardware utilization but introduces policy staleness, as the learner updates on trajectories generated by earlier policy versions. Learning effectively under this behavior–learner mismatch is therefore central to realizing the benefits of asynchronous RL.

Policy lag can destabilize optimization as trajectories become increasingly off-policy relative to the current learner policy (Xu et al., 2026; Huang et al., 2026). Recent asynchronous LLM-RL methods primarily mitigate this mismatch through token-level importance-ratio control in the actor objective, using clipping, masking, or related variance-control mechanisms (Guan et al., 2026; Zhao et al., 2026); we refer to these mechanisms as *policy-side correction*. However, policy-side correction

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

does not correct advantage estimation: the advantage still depends on continuations generated by the stale behavior policy. Hence, a token can receive an appropriate policy-side weight while its advantage remains tied to stale behavior-policy continuations. We term this *advantage staleness*.

Advantage staleness motivates a complementary *advantage-side correction* that acts directly on advantage estimation. This yields two distinct channels in the actor update: policy-side correction controls how off-policy samples are weighted, whereas advantage-side correction controls how their learning signals are estimated. This leads to a central question:

Do the two correction channels interact, and how should their strengths be coordinated?

We answer the first part by deriving exact bias and variance decompositions for a general two-channel actor update. The resulting expressions reveal a nonseparable interaction: policy-weight and advantage-estimation errors couple multiplicatively in the bias, while squared policy weights amplify conditional advantage uncertainty in the gradient variance. Thus, the two correction channels generally cannot be calibrated independently. Although the analysis does not prescribe an optimal coordination rule, it motivates a concrete, testable hypothesis: stronger policy-side mismatch control should favor weaker advantage-side correction, and vice versa.

Guided by this analysis, we introduce **C**oupled **O**ff-**P**olicy **C**orrection (**COPC**), an actor–critic method that instantiates the two channels with token-level ratio masking on the policy side and two-sided clipped-ratio weighting of TD residuals on the advantage side. The latter produces corrected returns for critic training and actor advantage estimation, while exposing a separate strength parameter that can be jointly calibrated with the policy-side correction in practice.

We test the coordination hypothesis through systematic two-dimensional parameter sweeps across multiple staleness levels. Figure 1 shows that preferred configurations follow the hypothesized coordination pattern, while the effect of varying one correction parameter can improve performance under one setting of the other but degrade it under another. This pattern persists across staleness levels, providing empirical support for both the hypothesized coordination and the cross-channel coupling predicted by our analysis. At the same time, strong performance spans a broad region of the parameter space, indicating robustness to precise hyperparameter choices. Given this robustness, we use the same pair of correction parameters across all tasks in our main experiments.

COPC achieves the highest reported performance on tool-integrated mathematical reasoning and search, measured by avg@32 and greedy-decoding accuracy, respectively, surpassing the strongest reported asynchronous baseline by 2.03 and 2.77 points. It substantially improves training stability: in the search setting, COPC remains stable throughout training, while most evaluated asynchronous baselines exhibit late-stage performance collapse. These gains persist at 64-step policy staleness, with minimal overhead over asynchronous PPO and a $1.7\times$ speedup over synchronous PPO.

Our contributions are fourfold:

- **Advantage staleness and two-channel analysis.** We identify *advantage staleness* as a distinct off-policy mismatch that persists after policy-side correction. Exact bias and variance decompositions for a general two-channel actor update reveal a nonseparable coupling between policy- and advantage-side errors, motivating a testable coordination hypothesis.
- **Coupled off-policy correction.** We introduce COPC, an actor–critic method combining token-level ratio masking with a new advantage-side correction based on two-sided clipped-ratio weighting of TD residuals. Corrected returns are used for critic training and actor advantage estimation, with separate channel parameters enabling joint calibration.
- **Empirical validation of coordinated correction.** Joint parameter sweeps across staleness levels support the coordination hypothesis and reveal pronounced cross-channel interactions. They also identify a broad high-performing parameter region, supporting a single shared setting across tasks without task-specific retuning in our experiments.
- **Performance, stability, and efficiency.** With the shared setting, COPC achieves the highest reported performance on both tool-integrated mathematical reasoning and search, while substantially improving training stability; in the search setting, it remains stable throughout training while most evaluated asynchronous baselines exhibit late-stage performance collapse. These gains persist at 64-step policy staleness, with minimal overhead over asynchronous PPO and a $1.7\times$ speedup over synchronous PPO across both evaluation settings.

2 Related Work

Asynchronous RL for LLM post-training. Asynchronous actor–learner architectures improve throughput by decoupling data collection from optimization, at the cost of behavior–learner policy mismatch (Espeholt et al., 2018). Recent systems extend this paradigm to LLM post-training through distributed rollout generation and asynchronous policy updates, including asynchronous RLHF, AReaL, and LlamaRL (Noukhovitch et al., 2025; Fu et al., 2025; Wu et al., 2025). Our work focuses on how to reliably correct learner updates under the resulting policy staleness.

Policy-side stabilization in off-policy LLM optimization. Recent LLM-RL methods mitigate behavior–learner mismatch through importance-ratio control and related stabilization mechanisms. AReaL separates behavior-policy correction from proximal policy constraints, while SAO applies two-sided token-level ratio masking to suppress updates under large policy mismatch (Fu et al., 2025; Hou et al., 2026). Other approaches target variance under policy staleness: VCPO combines effective-sample-size-based learning-rate adaptation with a minimum-variance baseline, whereas M2PO constrains the second moment of importance weights to limit high-variance updates (Huang et al., 2026; Zheng et al., 2025). These methods primarily stabilize off-policy actor updates, whereas COPC additionally corrects continuation-dependent returns and advantages.

Off-policy return and advantage correction. Classical off-policy RL methods correct multi-step return estimation through importance weighting or truncation. Retrace constructs off-policy return targets using truncated importance ratios, while V-trace uses corrected returns for both critic learning and actor advantage estimation (Munos et al., 2016; Espeholt et al., 2018). Reactor studies how truncation and value-estimation error affect off-policy actor–critic updates, and C-trace characterizes trade-offs among update variance, fixed-point bias, and contraction rate (Gruslys et al., 2017; Rowland et al., 2020). COPC differs in both correction design and scope: it introduces an advantage-side correction based on two-sided clipped-ratio weighting of TD residuals, with a separate parameter from policy-side masking, and studies the interaction between the two correction channels.

3 Preliminaries and Motivations

3.1 Asynchronous RL Training

We model an LLM as an autoregressive policy π_θ generating a response $y = (y_1, \ldots, y_{|y|})$ to a prompt $x \sim \mathcal{D}$, where s_t denotes the context preceding y_t. In asynchronous RL, the current *learner policy* π_θ is updated using responses generated by a stale *behavior policy* μ. PPO-style optimization then combines importance ratios with advantage estimates in a clipped surrogate objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\mu(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\min\Big(\rho_t(\theta)\hat{A}_t, \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\,\hat{A}_t\Big)\right]. \tag{1}$$

Here, the token-level ratio $\rho_t(\theta) = \frac{\pi_\theta(y_t|s_t)}{\mu(y_t|s_t)}$ measures learner–behavior policy mismatch at token t, $\hat{A}_t$ denotes the advantage estimate used by the actor, and ϵ controls the clipping range.

Proximal Policy Optimization (PPO). PPO uses a learned critic $V_\phi(s_t)$ to estimate state values. Letting $V_t = V_\phi(s_t)$, generalized advantage estimation (GAE) (Schulman et al., 2015) constructs advantages from temporal-difference (TD) residuals:

$$\hat{A}_t^{\mathrm{GAE}} = \sum_{j=t}^{|y|}(\gamma\lambda)^{j-t}\delta_j, \qquad \delta_j = r_j + \gamma V_{j+1} - V_j. \tag{2}$$

Here, r_j denotes the reward at position j, γ is the discount factor, and λ controls trace decay and the associated bias–variance trade-off. We set the terminal value to $V_{|y|+1} = 0$.

3.2 Why Advantage-Side Correction Is Needed

To isolate the role of advantage-side correction, we reweight TD residuals along behavior-policy trajectories using two-sided clipped importance ratios. Let $\bar{\rho}_j = \mathrm{clip}\big(\rho_j(\theta), 1-\epsilon_\ell, 1+\epsilon_h\big)$ denote

the clipped ratio. We define the corrected return and induced advantage as

$$\hat{v}_t = V_t + \sum_{j=t}^{|y|} (\gamma\lambda)^{j-t} \bar{\rho}_j \delta_j, \qquad \hat{A}_t = \delta_t + \sum_{j=t+1}^{|y|} (\gamma\lambda)^{j-t} \bar{\rho}_j \delta_j. \tag{3}$$

Setting $\bar{\rho}_j = 1$ recovers standard GAE in equation 2, while the corresponding expected return operator has fixed point V^{μ}. Setting $\bar{\rho}_j = \rho_j(\theta)$ instead applies the exact, unclipped importance ratio and shifts the fixed point to V^{π_θ}, correcting the mismatch induced by stale behavior-policy continuations but potentially increasing estimator variance. Two-sided clipping bounds extreme ratios and controls this variance, at the cost of shifting the fixed point away from V^{π_θ}. The following proposition characterizes the resulting effective policy; its proof is provided in Appendix A.

Proposition 1 (Unique fixed point)**.** *Under the finite-horizon and support assumptions, the expected corrected-return operator associated with equation 3 has the unique fixed point $V^{\widetilde{\pi}}$, where*

$$\widetilde{\pi}(a \mid s) = \frac{\mu(a \mid s)\bar{\rho}(s,a)}{\sum_b \mu(b \mid s)\bar{\rho}(s,b)}. \tag{4}$$

Proposition 1 makes the source of advantage staleness explicit. Without advantage-side correction, $\bar{\rho}(s,a) = 1$ and hence $\widetilde{\pi} = \mu$, so the return target and induced advantage remain tied to stale behavior-policy continuations rather than the learner policy π_θ. Thus, policy-side correction alone cannot eliminate advantage staleness. Consistent with this prediction, COPC w/o RC replaces the corrected returns and advantages with standard GAE while keeping other settings fixed, and underperforms COPC on both tool-integrated mathematical reasoning and search (Tables 1 and 2).

3.3 Two-Channel Correction Setup

We formalize the two correction channels through an idealized reference actor update. Let $A_t^{\pi_\theta}$ denote the advantage under current-policy continuations and $\psi_t(\theta) = \nabla_\theta \log \pi_\theta(y_t \mid s_t)$ the token-level score function. Using the raw importance ratio and current-policy advantage, we define

$$g^\star(\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\mu(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} \rho_t(\theta)\psi_t(\theta)A_t^{\pi_\theta}\right]. \tag{5}$$

This reference isolates the two quantities that must be approximated or controlled in practice: the policy weight $\rho_t(\theta)$ and the continuation-dependent advantage $A_t^{\pi_\theta}$. We parameterize their respective corrections by η_P and η_A to study their interaction in the actor update.

Policy-side correction. Policy-side correction replaces the raw importance ratio $\rho_t(\theta)$ with a controlled weight $w_t(\eta_P)$. For example, SAO (Hou et al., 2026) uses ratio masking:

$$w_t(\eta_P) = \rho_t(\theta)\,\mathbb{I}\{1 - \epsilon_\ell^P \le \rho_t(\theta) \le 1 + \epsilon_h^P\}.$$

Here, $\eta_P = (\epsilon_\ell^P, \epsilon_h^P)$ defines the masking interval: $w_t = \rho_t(\theta)$ inside and $w_t = 0$ outside.

Advantage-side correction. Advantage-side correction replaces the ideal current-policy advantage $A_t^{\pi_\theta}$ with an estimate $\hat{A}_t(\eta_A)$ that accounts for mismatch from stale behavior-policy continuations, where η_A parameterizes the correction. For example, in equation 3, $\eta_A = (\epsilon_\ell^A, \epsilon_h^A)$ specifies the clipping interval for importance ratios used to weight TD residuals.

Combining the two correction channels yields the expected length-normalized actor update

$$g(\eta_P, \eta_A) = \mathbb{E}_{x\sim\mathcal{D},\, y\sim\mu(\cdot|x)}\left[\frac{1}{|y|}\sum_{t=1}^{|y|} w_t(\eta_P)\psi_t(\theta)\hat{A}_t(\eta_A)\right]. \tag{6}$$

4 Coupling Between Correction Channels

The preceding fixed-point analysis characterizes the effective value target induced by a concrete advantage-side correction. We now study, more generally, how advantage-side correction interacts

with policy-side ratio control in the actor update. We derive exact bias and variance decompositions relative to the reference update $g^\star(\theta)$ to characterize how errors from the two correction channels interact. The analysis treats $w_t(\eta_P)$ and $\hat{A}_t(\eta_A)$ abstractly under the conditions below, without assuming a particular correction mechanism. All quantities are evaluated at a fixed training iterate, with all required moments assumed finite. Proofs are provided in Appendix B.

4.1 Bias Coupling

For channel parameters (η_P, η_A), define the policy-weight and advantage-estimation errors as

$$\Delta_t(\eta_P) = w_t(\eta_P) - \rho_t(\theta), \qquad e_t(\eta_A) = \hat{A}_t(\eta_A) - A_t^{\pi_\theta}.$$

Proposition 2 (Bias coupling)**.** *The update bias relative to the reference satisfies*

$$g(\eta_P, \eta_A) - g^\star(\theta) = \mathbb{E}\Big[\frac{1}{|y|}\sum_{t=1}^{|y|}\psi_t(\theta)\Big(\underbrace{\Delta_t(\eta_P)A_t^{\pi_\theta}}_{\textit{policy-side}} + \underbrace{\rho_t(\theta)e_t(\eta_A)}_{\textit{advantage-side}} + \underbrace{\Delta_t(\eta_P)e_t(\eta_A)}_{\textit{interaction}}\Big)\Big]. \tag{7}$$

The first two terms capture the direct contributions of policy-weight and advantage-estimation errors. The interaction term $\Delta_t(\eta_P)e_t(\eta_A)$ reveals their multiplicative coupling: its expected contribution can reinforce or offset the direct terms. Consequently, the effect of adjusting either correction channel on the overall update bias generally depends on the residual error in the other.

4.2 Variance Coupling

For a random vector X, let $\mathbb{V}[X] = \mathbb{E}[|X - \mathbb{E}[X]|_2^2]$ denote its total variance. Define the token-level gradient contribution as

$$Z_t = \frac{w_t(\eta_P)\psi_t(\theta)\hat{A}_t(\eta_A)}{|y|},$$

with $Z_t = 0$ beyond the response length. Let C_t denote conditioning information under which $|y|$, $\psi_t(\theta)$, and $w_t(\eta_P)$ are measurable for every policy-side setting η_P. Using the same conditioning information across correction settings, we define the conditional moments

$$m_t(\eta_A) = \mathbb{E}[\hat{A}_t(\eta_A) \mid C_t], \qquad \sigma_t^2(\eta_A) = \mathrm{Var}[\hat{A}_t(\eta_A) \mid C_t].$$

Proposition 3 (Variance coupling)**.** *The variance of each token contribution decomposes as*

$$\mathbb{V}[Z_t] = \mathbb{V}\left[\frac{w_t(\eta_P)\psi_t(\theta)m_t(\eta_A)}{|y|}\right] + \mathbb{E}\left[\frac{w_t^2(\eta_P)\|\psi_t(\theta)\|_2^2\sigma_t^2(\eta_A)}{|y|^2}\right]. \tag{8}$$

The first term captures conditional-mean variation. The second exposes coupling through $w_t^2(\eta_P)\sigma_t^2(\eta_A)$: squared policy weights rescale advantage uncertainty in gradient variance. The full-response decomposition, including cross-token covariances, is given in Appendix B.3.

4.3 Joint Gradient-Estimation Error

Consider N prompt groups, each containing K dependent responses. Let $Z_{i,k}$ denote the gradient contribution of response k to prompt i, and define group average and finite-sample update as

$$\bar{Z}_{i,K} = \frac{1}{K}\sum_{k=1}^{K} Z_{i,k}, \qquad \hat{g}_{N,K} = \frac{1}{N}\sum_{i=1}^{N}\bar{Z}_{i,K}.$$

We assume the group averages are independent and identically distributed while allowing dependence among responses within each group. Let $\bar{Z}_K$ denote a generic group average, with $\mathbb{E}[\bar{Z}_K] = g(\eta_P, \eta_A)$ under the sampling and advantage-estimation scheme defining g.

Corollary 1 (Joint gradient error)**.** *For fixed (η_P, η_A), the update has mean-squared error as follows*

$$\mathcal{E}(\eta_P, \eta_A) := \mathbb{E}\big[\|\hat{g}_{N,K} - g^\star(\theta)\|_2^2\big] = \|g(\eta_P, \eta_A) - g^\star(\theta)\|_2^2 + \frac{1}{N}\mathbb{V}[\bar{Z}_K]. \tag{9}$$

Remark 1 (Implications for joint calibration)**.** *The preceding decompositions expose cross-channel interactions in both update bias and variance through $\Delta_t e_t$ and $w_t^2\sigma_t^2$. Consequently, the gradient-estimation error in equation 9 need not be separable in η_P and η_A at a fixed training iterate, motivating joint rather than independent calibration of the two correction channels.*

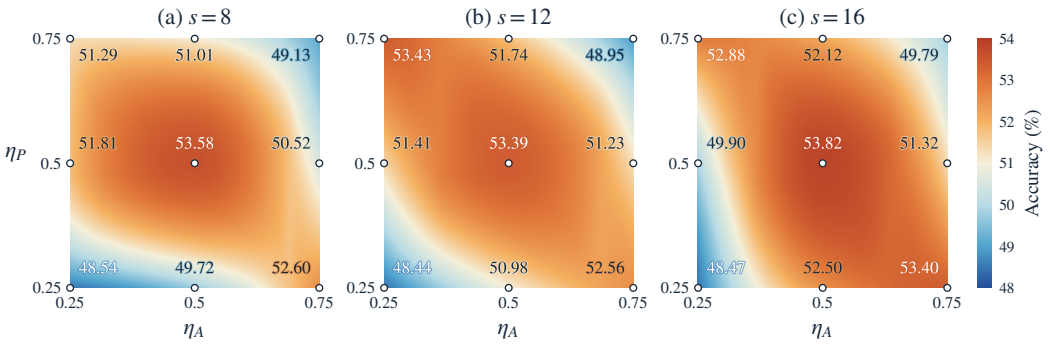

Figure 1: Mean avg@32 accuracy on AIME 2025 under joint policy- and advantage-side correction sweeps at staleness levels $s \in \{8, 12, 16\}$.

Coordination hypothesis. This nonseparable coupling motivates a concrete directional hypothesis: stronger correction in one channel should favor weaker correction in the other, and vice versa. We test this hypothesis through two-dimensional parameter sweeps across multiple staleness levels. Under the COPC parameterization in Section 5, smaller values of η_P and η_A correspond to stronger correction in their respective channels. As shown in Figure 1, the sweeps are consistent with the hypothesized coordination and reveal pronounced cross-channel interactions: changing one correction parameter can improve performance under one setting of the other yet degrade it under another.

5 COPC: Coupled Off-Policy Correction

We introduce COPC, an actor–critic method that combines our proposed advantage-side correction in equation 3 with the policy-side masking rule of SAO (Hou et al., 2026). The corrected returns serve as critic targets and construct actor advantages, while masking controls how these learning signals contribute to the policy update. Motivated by the coupling analysis in Section 4, COPC treats the two correction strengths as jointly selected hyperparameters.

Joint parameterization. We parameterize the policy- and advantage-side correction intervals using fixed base margins $\epsilon_\ell, \epsilon_h > 0$ and separate interval controls $\eta_P > 0$ and $\eta_A \geq 0$:

$$\ell_X = 1 - \eta_X \epsilon_\ell, \qquad u_X = 1 + \eta_X \epsilon_h, \qquad X \in \{P, A\}.$$

Here, correction strength refers to the restrictiveness of the ratio constraint: smaller η_X narrows the interval and represents stronger correction under this convention. The bounds (ℓ_P, u_P) control policy-side masking, while (ℓ_A, u_A) control advantage-side ratio clipping. Thus, smaller η_P yields stricter masking, whereas smaller η_A clips TD-residual importance weights more tightly around 1. We select (η_P, η_A) jointly to account for their coupled effect on the actor update.

Actor–critic update. For a rollout batch $\mathcal{B}$, before advantage normalization, the actor objective is

$$\mathcal{L}_{\text{actor}}(\theta) = -\frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \frac{1}{|y|} \sum_{t=1}^{|y|} \rho_t(\theta)\, \mathbb{I}\{\ell_P < \rho_t(\theta) < u_P\}\, \text{sg}\Big[\hat{A}_t(\eta_A)\Big]. \tag{10}$$

Here, $\text{sg}[\cdot]$ denotes stop-gradient. The indicator is treated as a fixed, non-differentiable mask: masked tokens contribute zero actor gradient, whereas retained tokens preserve the gradient path through the importance-ratio term $\rho_t(\theta)$, without differentiating through the masking decision.

The critic learns from the corrected returns via clipped value regression. Let $V_\phi(s_t)$ denote the current value prediction and V_t the frozen pre-update prediction. With threshold $\epsilon_V > 0$, define

$$V_t^{\text{clip}}(\phi) = V_t + \text{clip}(V_\phi(s_t) - V_t, -\epsilon_V, \epsilon_V).$$

The critic is optimized with the clipped regression objective

$$\mathcal{L}_{\text{critic}}(\phi) = \frac{1}{|\mathcal{B}|} \sum_{(x,y)\in\mathcal{B}} \frac{1}{|y|} \sum_{t=1}^{|y|} \max\left\{ \left(V_\phi(s_t) - \text{sg}[\hat{v}_t]\right)^2, \left(V_t^{\text{clip}}(\phi) - \text{sg}[\hat{v}_t]\right)^2 \right\}. \tag{11}$$

Table 1: Performance on tool-integrated mathematical reasoning measured by avg@32. Best in **bold** and second-best underlined within each backbone group. Δ rows report gains of COPC over PPO.

Method	AIME24	AIME25	AIME26	BRUMO25	BeyondAIME	Avg.
Qwen3-4B-Instruct						
PPO	54.65	43.75	47.40	44.83	27.53	43.63
V-trace	55.07	46.28	50.17	46.42	28.80	45.35
DAPO	56.42	48.99	51.11	46.08	29.17	46.35
SAO	58.06	50.31	53.13	47.29	30.48	47.85
AReaL	58.40	51.25	53.47	47.01	31.19	48.26
COPC w/o RC	58.33	50.83	53.33	47.22	30.82	48.11
COPC (ours)	**60.73**	**53.58**	**55.63**	**49.27**	**32.22**	**50.29**
Δ	+6.08	+9.83	+8.23	+4.44	+4.69	+6.65

Table 2: Performance on search measured by greedy-decoding accuracy. Best in **bold** and second-best underlined within each backbone group. Δ reports gains of COPC over PPO.

Method	NQ	TriviaQA	PopQA	HotpotQA	2Wiki	MuSiQue	Bamboogle	Avg.
Llama-3.2-3B-Instruct								
PPO	41.00	52.20	37.20	37.60	27.00	26.00	40.80	37.40
V-trace	40.40	55.40	36.80	38.20	25.20	26.20	42.40	37.80
DAPO	44.00	54.60	36.80	39.40	39.80	28.80	45.60	41.29
SAO	39.80	48.20	39.20	37.40	40.20	27.20	46.40	39.77
AReaL	40.80	52.80	39.20	43.20	**41.00**	30.80	44.80	41.80
COPC w/o RC	43.00	56.20	**39.80**	42.00	29.40	28.80	46.40	40.80
COPC (ours)	**45.80**	**57.20**	38.80	**47.60**	40.40	**32.60**	**49.60**	**44.57**
Δ	+4.80	+5.00	+1.60	+10.00	+13.40	+6.60	+8.80	+7.17

Return targets and actor advantages are computed from frozen pre-update quantities and held fixed during optimization. Policy-side masking acts only on the actor gradient: masked tokens contribute to return estimation and critic training, and their TD residuals can influence earlier advantages.

6 EXPERIMENTS

We evaluate COPC along four dimensions. Joint parameter sweeps test the coordination hypothesis across staleness levels, while experiments on tool-integrated mathematical reasoning and search compare COPC with synchronous and asynchronous baselines and isolate the contribution of advantage-side correction. We further assess robustness to policy staleness and training efficiency through step time, throughput, and step-time speedup over synchronous PPO. Detailed training and inference configurations, including method-specific settings, are provided in Appendix C.

6.1 EXPERIMENTAL SETUP

All experiments are implemented in Slime (Zhu et al., 2025), which supports asynchronous RL by overlapping rollout generation with policy optimization. Within each task setting, all methods share the same rollout and evaluation implementations while retaining their method-specific objectives and return or advantage estimators.

We use a strengthened PPO training recipe incorporating selected practices from VAPO (Yue et al., 2025), including Clip-Higher, critic pretraining, group sampling, advantage normalization, and clipped value regression. We use $\gamma = \lambda = 1$ throughout all experiments. These choices provide a training backbone while keeping the comparison focused on policy- and advantage-side correction.

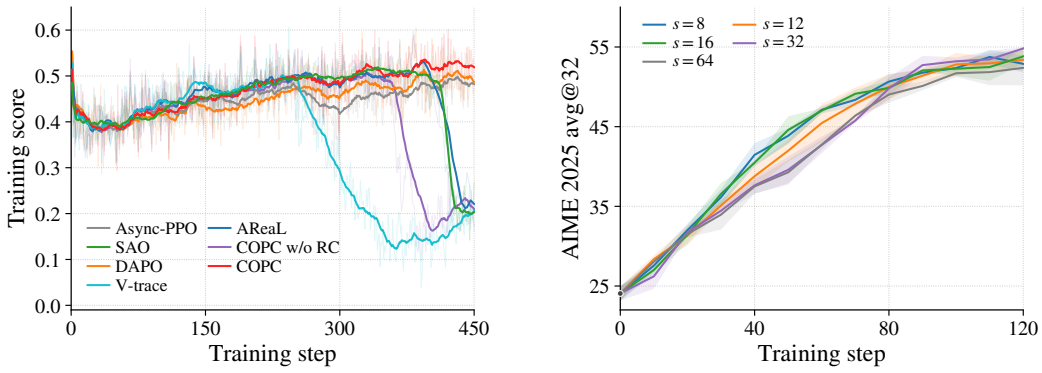

Figure 2: Training stability and robustness. Left: training dynamics of asynchronous methods on search. Right: COPC training dynamics on AIME 2025 under different policy-staleness levels.

Baselines. We compare COPC with representative asynchronous RL baselines and a controlled ablation. PPO (Schulman et al., 2017) uses the clipped surrogate objective with GAE advantages and follows our strengthened PPO recipe incorporating selected practices from VAPO (Yue et al., 2025), serving as a strong reference without explicit off-policy correction. DAPO (Yu et al., 2025) provides a critic-free alternative with group-relative advantages and asymmetric ratio clipping. AReaL (Fu et al., 2025) separates behavior-policy correction from proximal policy constraints, while SAO (Hou et al., 2026) applies token-level ratio masking to suppress updates under large policy mismatch. V-trace (Espeholt et al., 2018) uses truncated importance ratios for corrected return estimation and actor advantage construction. We additionally report COPC w/o RC, which replaces COPC's corrected returns and advantages with standard GAE while keeping all other settings fixed, isolating the contribution of advantage-side correction beyond policy-side masking.

Tool-integrated mathematical reasoning. We use Qwen3-4B-Instruct (Yang et al., 2025) as the policy backbone. Following the released ReTool training recipe (Feng et al., 2026), we initialize the policy with supervised fine-tuning on the ReTool SFT dataset and perform RL training on DAPO-Math-17K (Yu et al., 2025). We evaluate on AIME 2024/2025/2026 (MAA, 2025), BeyondAIME (ByteDance-Seed, 2025), and BrUMO (Balunovic et al., 2025).

Search. We use Llama-3.2-3B-Instruct (Grattafiori et al., 2024) as the policy backbone. Before RL, we fine-tune the policy on 2,000 examples following the SFT setup of Li et al. (2025). Following the Search-R1 setup (Jin et al., 2025), RL training uses Natural Questions (NQ) (Kwiatkowski et al., 2019) and HotpotQA (Yang et al., 2018). Retrieval uses the 2018 Wikipedia corpus (Karpukhin et al., 2020) indexed with E5-base-v2 embeddings (Wang et al., 2024), returning the top three documents per query. We evaluate on seven benchmarks: the single-hop datasets NQ, TriviaQA (Joshi et al., 2017), and PopQA (Mallen et al., 2023), and the multi-hop datasets HotpotQA, 2WikiMultiHopQA (Ho et al., 2020), MuSiQue (Trivedi et al., 2022), and Bamboogle (Press et al., 2023).

6.2 Joint Parameter Sweeps

Figure 1 reveals an interaction between policy- and advantage-side correction across all staleness levels. When $\eta_P = 0.25$, increasing η_A from 0.25 to 0.75 improves performance, whereas the change degrades performance when $\eta_P = 0.75$. This reversal persists across all staleness levels, demonstrating that the effect of adjusting one correction channel depends on the setting of the other.

The sweeps also reveal a compensatory structure: stronger correction in one channel tends to favor weaker correction in the other. Importantly, high performance is not confined to a narrowly tuned parameter pair; a broad range of correction strengths remains competitive across all tested staleness levels, indicating substantial hyperparameter robustness. Together, these results support the coordination hypothesis in Section 4 and motivate joint calibration of the two channels, while enabling a shared setting across subsequent experiments without staleness-specific retuning.

Table 3: Training efficiency on tool-integrated mathematical reasoning and search.

	Step time (s)	Throughput (tokens/s/GPU)	Step-time speedup
Tool-integrated mathematical reasoning			
Synchronous PPO	410.59	614.96	$1.00\times$
Asynchronous PPO	236.80	856.93	$1.73\times$
COPC (ours)	238.95	836.85	$1.72\times$
Search			
Synchronous PPO	213.18	65.35	$1.00\times$
Asynchronous PPO	124.90	96.44	$1.71\times$
COPC (ours)	126.18	94.92	$1.69\times$

6.3 Main Results and Ablations

Tables 1 and 2 report results on tool-integrated mathematical reasoning and search. COPC achieves the highest average performance across settings. On mathematical reasoning, it attains the best result on all five benchmarks, reaching an average of 50.29. On search, COPC achieves the highest average accuracy of 44.57 and outperforms the strongest asynchronous baseline by 2.77 points. This advantage is aligned with COPC's two-channel design: rather than correcting policy mismatch alone, COPC controls policy-side updates and continuation-dependent return and advantage estimation.

The ablation COPC w/o RC replaces corrected returns and advantages with standard GAE while retaining other settings. Return correction improves performance from 48.11 to 50.29 on mathematical reasoning and from 40.80 to 44.57 on search, outperforming the ablation on all five mathematical-reasoning benchmarks and six of seven search benchmarks. These gains show that policy-side masking alone cannot explain COPC's improvement; advantage-side correction provides a complementary benefit, consistent with the advantage-staleness motivation in Section 3.

6.4 Training Stability, Robustness, and Efficiency

Figure 2 evaluates training stability and robustness to increasing policy staleness. On search, COPC remains stable throughout training, whereas most evaluated asynchronous baselines exhibit late-stage performance collapse. COPC w/o RC also degrades sharply late in training, suggesting that advantage-side correction contributes not only to final performance but also to training stability. On AIME 2025, COPC remains stable as staleness increases from $s = 8$ to $s = 64$, with comparable final avg@32 performance across all settings, demonstrating robustness to highly stale rollouts.

Across both task settings, COPC closely matches the efficiency of asynchronous PPO, incurring only about a 1% increase in step time and at most a 2.5% reduction in throughput, while retaining an approximately $1.7\times$ step-time speedup over synchronous PPO. Together, these results show that COPC improves the robustness of asynchronous training without sacrificing its efficiency advantage.

7 Conclusion

We show that correcting stale trajectories in asynchronous LLM reinforcement learning requires more than controlling policy mismatch in the actor update. We identify *advantage staleness*, arising from behavior-policy continuations embedded in return and advantage estimates, and formulate stale-trajectory correction through complementary policy- and advantage-side channels. Our bias and variance decompositions reveal nonseparable interactions between the two channels, motivating their joint calibration. Building on this insight, we introduce COPC, which combines token-level policy-side masking with advantage-side correction based on two-sided clipped-ratio weighting of TD residuals. Across tool-integrated mathematical reasoning and search, COPC improves performance and training stability, remains robust across correction strengths and substantial policy staleness, and incurs minimal additional overhead while retaining a $1.7\times$ step-time speedup over synchronous PPO. More broadly, our results suggest that off-policy correction in asynchronous LLM RL should be treated as a coupled estimation problem rather than solely as policy-ratio control.

AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs, propose or refine hypotheses, assist with translation. We have not used generative AI tools for Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. Additionally, we used generative AI tools for create or modify scientific figures or images, create or edit software code, summarize or analyse existing literature, and edit a research paper to improve readability. We reviewed all AI-assisted material, checked the mathematical proofs and their assumptions, verified the cited references, validated figures against the underlying experimental records, and reviewed all AI-generated code for correctness. We take full responsibility for the accuracy and originality of the final work, including all content and code prepared with AI assistance.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. The study involved neither human participants nor animal experimentation. All datasets were obtained and used in accordance with their applicable licenses and usage guidelines, and no personally identifiable information was used. We considered potential sources of bias and privacy or security risks throughout the research process and took appropriate measures to mitigate them. We are committed to maintaining transparency, integrity, and responsible research practices.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of the results presented in this paper. All models and datasets used in our experiments are publicly available, and the implementation is available at `https://anonymous.4open.science/r/COPC-04A4`. We provide detailed descriptions of the training procedures, model configurations, hyperparameters, and hardware specifications in the appendix, together with the information necessary to facilitate independent reproduction of our experiments.

REFERENCES

Mislav Balunovic, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. MathArena: Evaluating LLMs on uncontaminated math competitions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

ByteDance-Seed. BeyondAIME: Advancing math reasoning evaluation beyond high school olympiads. `https://huggingface.co/datasets/ByteDance-Seed/BeyondAIME`, 2025.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad 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 *Proceedings of the 35th International Conference on Machine Learning*, pp. 1407–1416. PMLR, 2018.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=tRk1nofSmz`.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=X9diEuva9R`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Audrunas Gruslys, Will Dabney, Mohammad Gheshlaghi Azar, Bilal Piot, Marc Bellemare, and Remi Munos. The reactor: A fast and sample-efficient actor-critic agent for reinforcement learning. *arXiv preprint arXiv:1704.04651*, 2017.

Zhong Guan, Yongjian Guo, Haoran Sun, Wen Huang, Shuai Di, Likang Wu, Xiong Jun Wu, and Hongke Zhao. Missing old logits in asynchronous agentic rl: Semantic mismatch and repair methods for off-policy correction, 2026. URL https://arxiv.org/abs/2605.12070.

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

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps, 2020. URL https://arxiv.org/abs/2011.01060.

Zhenyu Hou, Yujiang Li, Jie Tang, and Yuxiao Dong. Single-rollout asynchronous optimization for agentic reinforcement learning. *arXiv preprint arXiv:2607.07508*, 2026.

Luke J. Huang, Zhuoyang Zhang, Qinghao Hu, Shang Yang, and Song Han. Stable asynchrony: Variance-controlled off-policy RL for LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=kiC03dyH9P.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning, 2025. URL https://arxiv.org/abs/2503.09516.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6769–6781, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.550. URL https://aclanthology.org/2020.emnlp-main.550/.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, et al. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276.

Weizhen Li, Jianbo Lin, Zhuosong Jiang, Jingyi Cao, Xinpeng Liu, Jiayu Zhang, Zhenqiang Huang, Qianben Chen, Weichen Sun, Qiexiang Wang, Hongxuan Lu, Tianrui Qin, Chenghao Zhu, Yi Yao, Shuying Fan, Xiaowan Li, Tiannan Wang, Pai Liu, King Zhu, He Zhu, Dingfeng Shi, Piaohong Wang, Yeyi Guan, Xiangru Tang, Minghao Liu, Yuchen Eleanor Jiang, Jian Yang, Jiaheng Liu, Ge Zhang, and Wangchunshu Zhou. Chain-of-agents: End-to-end agent foundation models via multi-agent distillation and agentic rl, 2025. URL https://arxiv.org/abs/2508.13167.

MAA. American invitational mathematics examination (aime), 2025. https://maa.org/.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, 2023. URL https://arxiv.org/abs/2212.10511.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc Bellemare. Safe and efficient off-policy reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous rlhf: Faster and more efficient off-policy rl for language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FhTAG591Ve.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models, 2023. URL https://arxiv.org/abs/2210.03350.

Mark Rowland, Will Dabney, and Rémi Munos. Adaptive trade-offs in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 34–44. PMLR, 2020.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

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

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 2022.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training, 2024. URL https://arxiv.org/abs/2212.03533.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, et al. Llamarl: A distributed asynchronous reinforcement learning framework for efficient large-scale llm training. *arXiv preprint arXiv:2505.24034*, 2025.

Haofeng Xu, Junwei Su, Yukun Tian, Lansong Diao, Zhengping Qian, and Chuan Wu. Gac: Stabilizing asynchronous rl training for llms via gradient alignment control, 2026. URL https://arxiv.org/abs/2603.01501.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

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

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyuan Xu, Jiaze Chen, Chengyi Wang, TianTian Fan, Zhengyin Du, et al. Vapo: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Guanqun Zhao, Zijun Xie, Binbin Zheng, Enlei Gong, Jiafeng Lu, Yehan Yang, Aoqi Hu, and Zeyu Chen. Deconstructing off-policy ratios: Entropy-scaled trust regions for asynchronous reinforcement learning, 2026. URL https://arxiv.org/abs/2607.22186.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms?, 2025. URL https://arxiv.org/abs/2510.01161.

Zilin Zhu, Chengxing Xie, Xin Lv, and slime Contributors. slime: An llm post-training framework for rl scaling. https://github.com/THUDM/slime, 2025. GitHub repository. Corresponding author: Xin Lv.

A Proof of Proposition 1

Fix the learner policy π_θ and behavior policy μ. Let $(\gamma, \lambda) \in [0,1]^2$ and $(\epsilon_\ell, \epsilon_h) \in [0,1] \times [0,\infty)$ be fixed. Consider a finite-horizon process with time-indexed states and episode length $T = |y| \le H$ almost surely, where $H \in \mathbb{N}$. Value functions satisfy $V(s_{T+1}) = 0$ at terminal states. For every nonterminal state s, assume $\operatorname{supp} \pi_\theta(\cdot \mid s) \subseteq \operatorname{supp} \mu(\cdot \mid s)$. All expectations are assumed finite.

Proof. For $a \in \operatorname{supp} \mu(\cdot \mid s)$, define

$$\bar{\rho}(s,a) = \operatorname{clip}\left(\frac{\pi_\theta(a \mid s)}{\mu(a \mid s)}, 1-\epsilon_\ell, 1+\epsilon_h\right), \qquad q(s,a) = \begin{cases} \mu(a \mid s)\bar{\rho}(s,a), & a \in \operatorname{supp} \mu(\cdot \mid s), \\ 0, & \text{otherwise.} \end{cases}$$

By the support assumption,

$$q(s,a) = \operatorname{clip}(\pi_\theta(a \mid s), (1-\epsilon_\ell)\mu(a \mid s), (1+\epsilon_h)\mu(a \mid s))$$

for every action a. Define $\kappa(s) = \sum_a q(s,a)$. Since $\pi_\theta(\cdot \mid s)$ is a probability distribution and its support is contained in that of $\mu(\cdot \mid s)$, there exists at least one action with $q(s,a) > 0$. Hence $\kappa(s) > 0$ at every nonterminal state.

Write $\bar{\rho}_j = \bar{\rho}(s_j, y_j)$ and $\delta_j^V = r_j + \gamma V(s_{j+1}) - V(s_j)$. The expected corrected-return operator is

$$(\mathcal{R}V)(s) = V(s) + \mathbb{E}_\mu\left[\sum_{j=t}^{T} (\gamma\lambda)^{j-t} \bar{\rho}_j \delta_j^V \,\middle|\, s_t = s\right]. \tag{12}$$

We first establish uniqueness. Suppose V_1 and V_2 are fixed points of $\mathcal{R}$, and let $U = V_1 - V_2$. Subtracting their fixed-point equations gives

$$\mathbb{E}_\mu\left[\sum_{j=t}^{T} (\gamma\lambda)^{j-t} \bar{\rho}_j \big(\gamma U(s_{j+1}) - U(s_j)\big) \,\middle|\, s_t = s\right] = 0. \tag{13}$$

We proceed by backward induction over the time index. At the latest possible nonterminal time H, all successor states are terminal, so equation 13 reduces to $-\kappa(s)U(s) = 0$. Since $\kappa(s) > 0$, we have $U(s) = 0$ for every state at time H. Now suppose U vanishes on all states at times later than t. For any state s at time t, every future-state term in equation 13 vanishes, again leaving $-\kappa(s)U(s) = 0$. Thus $U(s) = 0$. Repeating backward over all time indices yields $U \equiv 0$, so the fixed point is unique.

It remains to identify the fixed point. Define

$$\widetilde{\pi}(a \mid s) = \frac{q(s,a)}{\kappa(s)} = \frac{\mu(a \mid s)\bar{\rho}(s,a)}{\sum_b \mu(b \mid s)\bar{\rho}(s,b)},$$

where the second equality uses the zero extension outside $\operatorname{supp} \mu(\cdot \mid s)$. Since $\kappa(s) > 0$, $\widetilde{\pi}$ is well defined and coincides with the policy in equation 4.

Let $r(s,a)$ denote the expected immediate reward and $P(\cdot \mid s,a)$ the transition kernel. At $V^{\widetilde{\pi}}$,

$$\begin{aligned} \mathbb{E}_\mu\left[\bar{\rho}_t \delta_t^{V^{\widetilde{\pi}}} \,\middle|\, s_t = s\right] &= \sum_a q(s,a)\left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s,a)} V^{\widetilde{\pi}}(s') - V^{\widetilde{\pi}}(s)\right] \\ &= \kappa(s)\left[\sum_a \widetilde{\pi}(a \mid s)\left(r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot \mid s,a)} V^{\widetilde{\pi}}(s')\right) - V^{\widetilde{\pi}}(s)\right] \\ &= 0, \end{aligned}$$

where the final equality follows from the finite-horizon Bellman equation for $\widetilde{\pi}$.

Therefore, each weighted TD residual has zero conditional expectation at its corresponding state. Applying the tower property to each term in equation 12 gives

$$(\mathcal{R}V^{\widetilde{\pi}})(s) - V^{\widetilde{\pi}}(s) = \mathbb{E}_\mu\left[\sum_{j=t}^{T} (\gamma\lambda)^{j-t} \bar{\rho}_j \delta_j^{V^{\widetilde{\pi}}} \,\middle|\, s_t = s\right] = 0.$$

Hence $V^{\widetilde{\pi}}$ is a fixed point of $\mathcal{R}$. Uniqueness then completes the proof. □

Corollary 2 (Convergence of the expected return iteration)**.** *Under the assumptions of Proposition 1, suppose additionally that the time-indexed nonterminal state space is finite. Fix π_θ, μ, and the clipping thresholds. Then, for any initial value function $V^{(0)}$ satisfying the terminal condition, the exact iteration $V^{(n+1)} = \mathcal{R}V^{(n)}$ converges to $V^{\widetilde{\pi}}$:*

$$\lim_{n\to\infty} \|V^{(n)} - V^{\widetilde{\pi}}\|_\infty = 0.$$

Proof. Let $\ell = 1 - \epsilon_\ell \in [0, 1]$ and $u = 1 + \epsilon_h \in [1, \infty)$. A direct case analysis gives

$$\mathrm{clip}(\xi, \ell, u) \le \ell + \left(1 - \frac{\ell}{u}\right)\xi \qquad \text{for all } \xi \ge 0.$$

By the support assumption, $\mathbb{E}_\mu[\rho_t \mid s_t = s] = 1$. Therefore,

$$0 < \kappa(s) = \mathbb{E}_\mu[\bar{\rho}_t \mid s_t = s] \le 1 + \ell\left(1 - \frac{1}{u}\right) < 2. \tag{14}$$

Represent value functions on the finite set of nonterminal states as vectors ordered by increasing time index. Since $\mathcal{R}$ is affine in V, write $\mathcal{R}V = b + MV$, where b is independent of V. From equation 12, the update at each state depends only on its own value and values at later time indices. Hence M is upper triangular. The coefficient multiplying the current-state value is

$$1 - \mathbb{E}_\mu[\bar{\rho}_t \mid s_t = s] = 1 - \kappa(s),$$

so the diagonal entries are $M_{ss} = 1 - \kappa(s)$. Therefore, writing $r(M)$ for the spectral radius of M,

$$r(M) = \max_s |1 - \kappa(s)| < 1,$$

where the strict inequality follows from equation 14 and finiteness of the state space.

By Proposition 1, $V^{\widetilde{\pi}} = b + MV^{\widetilde{\pi}}$. Subtracting this fixed-point identity from the iteration yields

$$V^{(n)} - V^{\widetilde{\pi}} = M^n\big(V^{(0)} - V^{\widetilde{\pi}}\big).$$

Since $r(M) < 1$, $M^n \to 0$ in any matrix norm. Consequently,

$$\|V^{(n)} - V^{\widetilde{\pi}}\|_\infty \le \|M^n\|_\infty \|V^{(0)} - V^{\widetilde{\pi}}\|_\infty \to 0.$$

□

B Proofs for Coupling Between Correction Channels

This appendix proves the results in Section 4. We fix the learner parameters θ, the value estimator, and the behavior-policy sampling scheme, and consider an arbitrary fixed pair of correction settings (η_P, η_A). Expectations are taken over the sampled prompt $x \sim \mathcal{D}$, the response $y \sim \mu(\cdot \mid x)$, and any additional randomness used to construct the advantage estimates, including co-sampled responses when the estimator uses a prompt group.

Responses are nonempty and have finite length almost surely. We assume that all terms in the expectation decompositions are absolutely integrable and that all quantities appearing in the variance results have finite second moments. No independence is assumed between tokens within a response or between responses within the same prompt group unless explicitly stated.

B.1 Proof of Proposition 2

Proof. Consider a realization of a prompt, its response, and any additional samples used to construct the advantage estimates. For each valid token position $t \in \{1, \dots, |y|\}$, recall

$$\Delta_t(\eta_P) = w_t(\eta_P) - \rho_t(\theta), \qquad e_t(\eta_A) = \hat{A}_t(\eta_A) - A_t^{\pi_\theta}.$$

Substituting these definitions gives the samplewise identity

$$\begin{aligned} w_t(\eta_P)\hat{A}_t(\eta_A) - \rho_t(\theta)A_t^{\pi_\theta} &= \big(\rho_t(\theta) + \Delta_t(\eta_P)\big)\big(A_t^{\pi_\theta} + e_t(\eta_A)\big) - \rho_t(\theta)A_t^{\pi_\theta} \\ &= \Delta_t(\eta_P)A_t^{\pi_\theta} + \rho_t(\theta)e_t(\eta_A) + \Delta_t(\eta_P)e_t(\eta_A). \end{aligned}$$

The first two terms capture the direct errors from the policy- and advantage-side channels, respectively, while the third captures their interaction.

Multiplying by the score $\psi_t(\theta)$, summing over valid token positions, and applying the response-length normalization yields

$$\frac{1}{|y|}\sum_{t=1}^{|y|}\psi_t(\theta)\big[w_t(\eta_P)\hat{A}_t(\eta_A)-\rho_t(\theta)A_t^{\pi_\theta}\big]$$
$$=\frac{1}{|y|}\sum_{t=1}^{|y|}\psi_t(\theta)\big[\Delta_t(\eta_P)A_t^{\pi_\theta}+\rho_t(\theta)e_t(\eta_A)+\Delta_t(\eta_P)e_t(\eta_A)\big].$$

Taking expectations under the common sampling law and using the definitions of $g(\eta_P,\eta_A)$ and $g^\star(\theta)$ in equation 6 and equation 5 gives

$$g(\eta_P,\eta_A)-g^\star(\theta)=\mathbb{E}\left[\frac{1}{|y|}\sum_{t=1}^{|y|}\psi_t(\theta)\big(\Delta_t(\eta_P)A_t^{\pi_\theta}+\rho_t(\theta)e_t(\eta_A)+\Delta_t(\eta_P)e_t(\eta_A)\big)\right].$$

The random response length remains inside the samplewise expression, so the argument applies directly to responses of varying lengths. By the stated integrability assumptions, all terms are well defined. Since the decomposition holds before taking expectations, it requires no independence between the two correction errors, token contributions, or co-sampled responses. This proves Proposition 2. □

B.2 Proof of Proposition 3

Proof. Fix a token position t and correction settings (η_P,η_A). Recall that C_t denotes conditioning information under which $|y|$, $\psi_t(\theta)$, and $w_t(\eta_P)$ are measurable for every policy-side setting η_P. Define

$$Z_t=\frac{w_t(\eta_P)\psi_t(\theta)\hat{A}_t(\eta_A)}{|y|}.$$

For positions beyond the realized response length, we set $w_t(\eta_P)$, $\psi_t(\theta)$, and $\hat{A}_t(\eta_A)$ to zero and assign fixed dummy values to (s_t,y_t). Thus, all token-level quantities are well defined for a fixed index t across responses of varying lengths.

By construction, $w_t(\eta_P)\psi_t(\theta)/|y|$ is measurable with respect to C_t, and the same conditioning information is used across correction settings. Under the standing moment assumptions, $\hat{A}_t(\eta_A)$ and Z_t are square-integrable. Recall

$$m_t(\eta_A)=\mathbb{E}[\hat{A}_t(\eta_A)\mid C_t],\qquad \sigma_t^2(\eta_A)=\mathrm{Var}[\hat{A}_t(\eta_A)\mid C_t].$$

Because padded advantage values are identically zero, both conditional moments vanish at padded positions.

Using the C_t-measurability of the coefficient,

$$\mathbb{E}[Z_t\mid C_t]=\frac{w_t(\eta_P)\psi_t(\theta)m_t(\eta_A)}{|y|}.$$

Hence

$$Z_t-\mathbb{E}[Z_t\mid C_t]=\frac{w_t(\eta_P)\psi_t(\theta)}{|y|}\big(\hat{A}_t(\eta_A)-m_t(\eta_A)\big),$$

and therefore

$$\mathbb{E}\Big[\big\|Z_t-\mathbb{E}[Z_t\mid C_t]\big\|_2^2\mid C_t\Big]=\frac{w_t^2(\eta_P)\|\psi_t(\theta)\|_2^2\sigma_t^2(\eta_A)}{|y|^2}.$$

To obtain the unconditional variance, write

$$Z_t-\mathbb{E}[Z_t]=\big(Z_t-\mathbb{E}[Z_t\mid C_t]\big)+\big(\mathbb{E}[Z_t\mid C_t]-\mathbb{E}[Z_t]\big).$$

The first term has zero conditional mean given C_t, whereas the second is C_t-measurable, so their expected inner product is zero. Expanding the squared norm therefore gives the vector law of total variance:

$$\mathbb{V}[Z_t] = \mathbb{V}[\mathbb{E}[Z_t \mid C_t]] + \mathbb{E}\Big[\big\|Z_t - \mathbb{E}[Z_t \mid C_t]\big\|_2^2\Big]. \tag{15}$$

Substituting the conditional mean and applying the tower property yields

$$\mathbb{V}[Z_t] = \mathbb{V}\left[\frac{w_t(\eta_P)\psi_t(\theta)m_t(\eta_A)}{|y|}\right] + \mathbb{E}\left[\frac{w_t^2(\eta_P)\|\psi_t(\theta)\|_2^2\sigma_t^2(\eta_A)}{|y|^2}\right].$$

This is the claimed decomposition. The argument relies only on conditional measurability and the stated moment assumptions; no independence between the policy-side weight and the advantage estimate is required. □

B.3 Response-Level Variance Decomposition

Assume additionally that $|y| \le T_{\max}$ almost surely for some finite $T_{\max}$. Under the zero-padding convention, the response-level gradient contribution satisfies

$$Z = \sum_{t=1}^{|y|} Z_t = \sum_{t=1}^{T_{\max}} Z_t.$$

Its total variance decomposes as

$$\mathbb{V}[Z] = \sum_{t=1}^{T_{\max}} \mathbb{V}[Z_t] + 2 \sum_{1 \le t < j \le T_{\max}} \mathbb{E}\big[(Z_t - \mathbb{E}[Z_t])^\top (Z_j - \mathbb{E}[Z_j])\big]. \tag{16}$$

Proof. Since the sum is finite, linearity of expectation gives

$$\mathbb{E}[Z] = \sum_{t=1}^{T_{\max}} \mathbb{E}[Z_t], \qquad Z - \mathbb{E}[Z] = \sum_{t=1}^{T_{\max}} \big(Z_t - \mathbb{E}[Z_t]\big).$$

Expanding the squared Euclidean norm yields

$$\big\|Z - \mathbb{E}[Z]\big\|_2^2 = \sum_{t=1}^{T_{\max}} \big\|Z_t - \mathbb{E}[Z_t]\big\|_2^2 + 2 \sum_{1 \le t < j \le T_{\max}} \big(Z_t - \mathbb{E}[Z_t]\big)^\top \big(Z_j - \mathbb{E}[Z_j]\big).$$

The finite second-moment assumptions ensure that all terms are integrable, with cross terms controlled by Cauchy–Schwarz. Taking expectations and using $\mathbb{V}[X] = \mathbb{E}[\|X - \mathbb{E}[X]\|_2^2]$ gives equation 16. □

Each Z_t retains the random response-length normalization and vanishes beyond termination. Thus, the decomposition captures both cross-token dependence and variability induced by response length, without assuming independence among token contributions.

B.4 Proof of Corollary 1

Proof. For prompt group i, let

$$\bar{Z}_{i,K} = \frac{1}{K} \sum_{k=1}^{K} Z_i^{(k)}$$

denote its average response-level gradient contribution. By definition,

$$\hat{g}_{N,K} = \frac{1}{N} \sum_{i=1}^{N} \bar{Z}_{i,K}.$$

Each response contribution has marginal expectation $g(\eta_P, \eta_A)$, and hence

$$\mathbb{E}[\bar{Z}_{i,K}] = g(\eta_P, \eta_A), \qquad \mathbb{E}[\hat{g}_{N,K}] = g(\eta_P, \eta_A).$$

For brevity, write $g = g(\eta_P, \eta_A)$ and $g^\star = g^\star(\theta)$. Then

$$\hat{g}_{N,K} - g^\star = (\hat{g}_{N,K} - g) + (g - g^\star).$$

Since $\mathbb{E}[\hat{g}_{N,K} - g] = 0$, the cross term vanishes, giving

$$\mathbb{E}\big[\|\hat{g}_{N,K} - g^\star\|_2^2\big] = \mathbb{E}\big[\|\hat{g}_{N,K} - g\|_2^2\big] + \|g - g^\star\|_2^2. \tag{17}$$

The N prompt groups are i.i.d., while arbitrary dependence among the K responses within each group is retained in $\bar{Z}_{i,K}$. Therefore,

$$\mathbb{E}\big[\|\hat{g}_{N,K} - g\|_2^2\big] = \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N}(\bar{Z}_{i,K} - g)\right\|_2^2\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\big[\|\bar{Z}_{i,K} - g\|_2^2\big] = \frac{1}{N}\mathbb{V}[\bar{Z}_K].$$

Thus, independence is required only across prompt groups; no independence is assumed among responses within the same group.

By Proposition 2, the bias term $g(\eta_P, \eta_A) - g^\star(\theta)$ admits the exact decomposition in equation 7. Substituting the variance identity into equation 17 therefore yields

$$\mathcal{E}(\eta_P, \eta_A) = \big\|g(\eta_P, \eta_A) - g^\star(\theta)\big\|_2^2 + \frac{1}{N}\mathbb{V}[\bar{Z}_K(\eta_P, \eta_A)],$$

which proves the corollary.

□

C Detailed Training Configurations

C.1 Prompt Templates and Reward Design

Tool-integrated mathematical reasoning. For tool-augmented mathematical reasoning, we use the Qwen chat template with the built-in Python code-interpreter instructions, shown below, for training and evaluation. During rollout, the assistant may issue structured tool calls; each call is executed and its output is appended to the conversation context before generation resumes, allowing subsequent reasoning to incorporate intermediate computational results.

We use a binary correctness reward. Let z denote the text passed to the mathematical verifier and $a^\star$ the reference answer. We use `math_verify` to extract candidate mathematical expressions from z and determine their equivalence to $a^\star$:

$$r_{\text{math}}(z, a^\star) = \begin{cases} 1, & \text{if an extracted answer is verified as equivalent to } a^\star, \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

No separate formatting reward or tool-use bonus is applied; tool calls affect the reward only through their contribution to a verifiably correct final answer.

System Prompt

You are Qwen, created by Alibaba Cloud. You are a helpful assistant.
`# Tools`
You may call one or more functions to assist with the user query.
You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools>
{"type": "function",
 "function": {
   "name": "code_interpreter",
   "description": "A tool for executing code.",
   "parameters": {
     "type": "object",
     "properties": {
       "code": {"type": "string", "description": "The code to execute
    ."}
     },
     "required": ["code"]
   }
 }}
</tools>
```

For each function call, return a JSON object with function name and arguments within `<tool_call></tool_call>` XML tags:

```
<tool_call>
{"name": <function-name>, "arguments": <args-json-object>}
</tool_call>
```

Search. For search-augmented reasoning, we use the system prompt shown below during both training and evaluation. The prompt specifies the search-tool schema and the required format for structured tool calls.

We use a shaped reward for search-augmented reasoning. Let $c(s, \mathcal{Y}) \in \{0, 1\}$ denote normalized exact-match correctness, $f(s) \in \{0, 1\}$ indicate whether a valid boxed answer is present, and n denote the number of executed search calls. The reward is

$$r_{\text{search}}(s, \mathcal{Y}, n) = \begin{cases} 2.0, & c(s, \mathcal{Y}) = 1,\ n > 0, \\ 1.1, & c(s, \mathcal{Y}) = 1,\ n = 0, \\ -1 + 0.1 f(s) + \min(0.1n, 0.3), & c(s, \mathcal{Y}) = 0. \end{cases} \tag{19}$$

A correct answer receives reward 1.1 without search and 2.0 with search. For incorrect answers, valid formatting contributes 0.1 and the search-use bonus is capped at 0.3, so the total reward remains negative.

System Prompt

Given the following functions, please respond with a JSON for a function call with its proper arguments that best answers the given prompt.
Respond in the format "name": function name, "parameters": dictionary of argument name and its value. Do not use variables.

```
{
  "function": {
    "description": "Searches the web for relevant information based
    on the given query.",
    "name": "search",
    "parameters": {
      "properties": {
        "query_list": {
          "description": "A list of fully-formed semantic queries.
    The tool will return search results for each query.",
          "items": {"type": "string"},
          "type": "array"
        }
      },
      "required": ["query_list"],
      "type": "object"
    }
  },
  "type": "function"
}
```

C.2 RL Training Configuration

All experiments were conducted on 8 NVIDIA H200 GPUs. We use Qwen3-4B-Instruct for tool-integrated mathematical reasoning and Llama-3.2-3B-Instruct for search-augmented reasoning.

Supervised Fine-Tuning. Before reinforcement learning, we perform supervised fine-tuning (SFT) for each policy backbone. For tool-integrated mathematical reasoning, Qwen3-4B is fine-tuned on the ReTool SFT corpus (Feng et al., 2026). For search-augmented reasoning, Llama-3.2-3B-Instruct is fine-tuned on 2,000 examples sampled from the Chain-of-Agents SFT corpus (Li et al., 2025). The SFT configurations are summarized in Table 4.

Table 5 summarizes the default training configurations for the two task settings. Method-specific hyperparameters and deviations, including correction thresholds and update schedules, are specified below. Critic-related settings apply only to methods with a learned value function. Prompt templates and reward functions are described in Section C.1.

C.2.1 Method-specific configurations.

Table 5 summarizes the shared default configurations. Unless otherwise specified, losses are averaged per response for tool-integrated mathematical reasoning and per token for search. Method-specific settings and deviations are described below.

PPO. PPO uses the clipped surrogate objective with standard GAE and follows the default configuration in Table 5.

DAPO. DAPO uses group-relative advantages without a learned critic. Rewards are centered and standardized within each prompt group, without additional batch-level advantage normalization. Dynamic sampling retains groups with nonzero reward variance and replenishes discarded groups

Table 4: Supervised fine-tuning (SFT) configurations.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	1×10^{-5}
Weight decay	0.01
Learning-rate schedule	cosine
Warmup ratio	0.1
Global batch size	16
Maximum sequence length	16,384 (math) / 8,192 (search)
Training epochs	2

until the training batch is complete. For tool-integrated mathematical reasoning, DAPO additionally replaces response-level loss averaging with token-level averaging. Overlong reward shaping is disabled.

SAO. SAO replaces the PPO clipped surrogate with two-sided importance-ratio masking using the default masking interval in Table 5. We sample one response per prompt in both task settings, so each training batch contains responses from 512 distinct prompts.

AReaL. AReaL uses standard GAE and a decoupled actor objective, clipping the learner-to-proximal ratio to the default interval in Table 5 while applying a separate importance weight to correct proximal-to-behavior mismatch. Tokens whose proximal-to-behavior ratio exceeds 5.0 are excluded from the actor loss. Each rollout batch contains 1,024 responses and is split into two disjoint minibatches of 512 responses, yielding two actor updates with the proximal policy fixed across both. For search, the actor learning-rate warmup lasts 25 optimizer updates.

V-trace. Following Espeholt et al. (2018), V-trace uses truncation thresholds of 1 for value correction, trace coefficients, and policy-gradient correction, i.e., $\bar{\rho} = \bar{c} = \bar{\rho}_{\mathrm{PG}} = 1$. The actor uses the V-trace policy-gradient objective without advantage normalization, while the critic is trained against the corresponding V-trace target without PPO-style value clipping.

COPC. COPC uses base clipping margins $\epsilon_\ell = 0.4$ and $\epsilon_h = 0.56$, with interval controls $(\eta_P, \eta_A) = (0.5, 0.5)$ in the main experiments. This gives the interval $[0.80, 1.28]$ for both policy-side masking and advantage-side ratio clipping. COPC w/o RC sets $\eta_A = 0$, recovering standard GAE on the advantage side, while retaining $\eta_P = 0.5$ and all other settings.

C.2.2 Inference configurations.

Tool-integrated mathematical reasoning. We sample 32 responses per problem using a temperature of 0.6 and top-$p = 1.0$. Each response has a total token budget of 38,912 across all interaction turns, including tool observations. We report avg@32 accuracy on AIME24, AIME25, AIME26, BRUMO25, and BeyondAIME.

Search. We use greedy decoding with a temperature of 0 and top-$p = 1.0$, generating one response per question with a maximum response budget of 8,192 tokens. We report normalized exact-match accuracy on NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, and Bamboogle.

Table 5: Default training configurations for tool-integrated mathematical reasoning and search. Rollout generation and policy optimization are executed asynchronously on separate worker groups. Method-specific deviations are described in the accompanying text.

Configuration	**Math-with-Tool**	**Search**
General training hyperparameters		
Policy model	Qwen3-4B-Instruct	Llama-3.2-3B-Instruct
Samples per prompt	8	4
Rollout batch size (responses)	512	512
Training minibatch size (responses)	512	512
Actor updates per rollout batch	1	1
Max prompt length (tokens)	2048	512
Max response length (tokens)	16384	8192
Max assistant turns	8	8
Actor learning rate	1×10^{-6}	1×10^{-6}
Max staleness S	8	14
Actor ratio clip interval	$[0.80, 1.28]$	$[0.80, 1.28]$
Actor–critic settings		
Critic pretraining samples	512×16	512×5
Critic pretraining steps	50	50
Critic learning rate	5×10^{-6}	5×10^{-6}
Discount factor γ	1	1
Trace parameter λ	1	1
Value clipping margin	0.2	0.2