Retroactive Advantage Correction: Closed-Form V-Trace Bias Correction for Delay-Aware RLHF

Arnav Raj [1]

Abstract

Reinforcement learning from human feedback (RLHF) in production does not always have a synchronous reward signal. Code-execution verifiers, slow judge ensembles, and queued human review can return several gradient steps after the rollout that produced them, breaking the synchronous-reward assumption underlying standard PPO. We address this gap with **Retroactive Advantage Correction** (RAC): each pending slow completion is queued, aged through a non-negative kernel, and reinjected as a clipped residual into the next optimiser step's advantage. We prove that under an unbiased clipped importance ratio, the cumulative RAC correction is exactly unbiased when the effective delay kernel reinjects all of its mass, and carries a bias linear in the unreinjected fraction otherwise; at the no-delay identity kernel it reduces to V-trace (Espeholt et al., 2018). On a tabular Markov decision process (MDP) proof-of-concept, RAC reduces the closed-form policy bias by up to $\mathbf{47.9}\times$ at the two-slow-channel configuration, beating wait-for-slow at lower wall-clock cost. RAC integrates with PPO and GRPO through a two-line reward-manager patch.

1. Introduction

Delayed rewards in deployed RLHF. Standard PPO (Schulman et al., 2017) implicitly assumes that the reward r_t for trajectory τ_t is observed *before* the optimiser commits its next gradient on π_θ. That assumption holds for simulated rollouts but is not always met in deployed RLHF (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022): code-execution verifiers return several gradient steps after the rollout that produced them (Jain et al., 2024; Li et al., 2026); slow but accurate 70B judge ensembles run on a different clock from a fast 7B PRISM-style scorer (Kirk et al., 2024); human review of edge cases returns minutes later, after the policy has already moved; and asynchronous training architectures such as AReaL (Fu et al., 2025) and Asynchronous RLHF (Noukhovitch et al., 2025) explicitly decouple inference workers from learners, making Δ-step staleness (where Δ is the gradient-step lag) a first-class operational parameter. When the slow signal arrives Δ steps late and is dropped, the bias accumulates with Δ and with the number K of distinct slow channels (e.g. a code-verifier plus a judge-RM is $K{=}2$).[1]

Off-policy correctors and the reward axis. V-trace (Espeholt et al., 2018) and Retrace (Munos et al., 2016) are clipped importance-sampling (IS) correctors for off-policy value-function targets, acting on the inner critic and addressing inter-worker actor staleness, not the inter-optimiser-step reward delay we target. Recent IS-clipping work tightens the off-policy PPO objective at the token or turn level: CISPO (Chen et al., 2025) clips the per-token ratio while preserving gradients; Truncated PPO (Fan et al., 2025) extends generalised advantage estimation (Schulman et al., 2016) onto truncated rollouts; and SORL (Li et al., 2025a) adds turn-level IS clipping with clipping-triggered normalisation. The 2025–2026 staleness-aware-PPO cluster (Zheng et al., 2025; Huang et al., 2026; Li et al., 2025b; Xi et al., 2025; Lu et al., 2026; Sheng et al., 2025; Yan et al., 2025) acts on the rollout-axis IS ratio at the token or sequence level. Prior delayed-reward correctors operate on a different axis. RUDDER (Arjona-Medina et al., 2019) redistributes a terminal episodic reward *backward* within an episode, a credit-assignment construction; Han et al. (2022) reformulate the Q-function for single-channel delayed reward in continuous control; Zhang et al. (2023) cover per-agent reward delays in tabular multi-agent RL with cooperative convergence-rate guarantees; Ramstedt et al. (2021) resample trajectory fragments in hindsight for action-delay and observation-delay MDPs; and AReaL (Fu et al., 2025) with Asynchronous

[1]Department of Computer Science and Engineering, Indian Institute of Technology Delhi, New Delhi, India. Correspondence to: Arnav Raj <arnav.raj.cs522@cse.iitd.ac.in>.

RLxF: Reinforcement Learning from World Feedback Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

[1]Code: https://github.com/deadsmash07/rac-rlxf-code.

RLHF (Noukhovitch et al., 2025) bound generation staleness empirically by capping how off-policy each inference worker can drift. RAC occupies a third axis. Each slow reward that arrives Δ optimiser steps late is queued and reinjected as an additive advantage correction at the next optimiser step, weighted by an age kernel and a clipped IS ratio. Across K slow channels these reinjections pass through a row-stochastic delay kernel, which makes the cumulative bias closed-form zero in expectation (Theorem 2.1) and recovers V-trace's on-policy guarantee at the identity kernel.

RAC at a glance. Figure 1 sketches the mechanism. Standard PPO commits its gradient before the slow channel returns and drops the residual signal. RAC instead treats the late slow return as evidence about an *earlier* rollout: it queues the pending slow completion, ages it through a non-negative kernel $w_{\text{age}}(\Delta)$, and reinjects the slow-minus-fast residual into the next gradient, weighted by a clipped IS ratio. The clip is the standard off-policy correction taken from V-trace (Espeholt et al., 2018) and transplanted from the value-target level to the advantage level. The construction is a multi-channel Δ-step lag accumulator constrained by a row-stochastic delay kernel. The cumulative bias of the correction is linear in the row-stochasticity slack of the effective kernel, and is exactly zero in expectation at the saturated case (Theorem 2.1). At the identity kernel $\Lambda{=}I$ (zero delay), Theorem 2.1 collapses to V-trace's on-policy guarantee, and the technical contribution is the closed-form multi-channel Δ-lagged extension.

Contributions. The paper makes two contributions.

- **(C1) Closed-form cumulative-bias identity.** Under the V-trace conditional-independence condition, the cumulative RAC correction has bias linear in the row-stochasticity slack η of the effective delay kernel, and is exact ($\eta{=}0$) at the saturated kernel and at the V-trace identity kernel (Theorem 2.1, Espeholt et al., 2018). A total-variation bound combining Pinsker's inequality (Canonne, 2022) with the Bretagnolle–Huber lemma (Bretagnolle & Huber, 1979) (Theorem 2.2) supplies per-step divergence control.
- **(C2) Tabular MDP proof-of-concept and 7B-scale verifications.** On a $3{\times}2$ tabular MDP RAC reduces closed-form policy bias by $\mathbf{47.9}\times$ at the $K{=}2$ configuration, with single-step wall-clock cost (Table 1, Figure 2). Three machine-precision 7B-scale checks corroborate the identity-kernel collapse, the linear-in-slack bias scaling, and V-trace equivalence on real reward distributions (Section B, Section G).

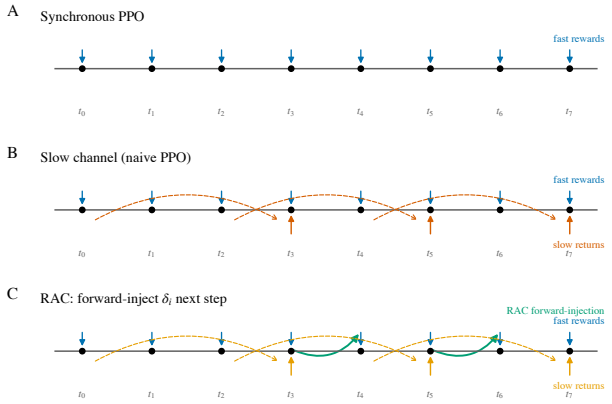

Figure 1. **Retroactive Advantage Correction (RAC) at a glance.** (**A**) Synchronous PPO assumes the reward arrives before the next optimiser step. (**B**) When a slow channel returns Δ steps later, naive PPO drops the residual and the resulting bias scales with $\Delta \cdot K$. (**C**) RAC queues each pending slow completion and forward-injects a clipped, age-decayed residual $\delta_i{=}w_{\text{age}}(\Delta)\,\alpha\,\rho_i^{\text{clip}}\,(r_i^{\text{slow}}{-}r_i^{\text{fast,bl}})$ into the next step's advantage; Theorem 2.1 gives the row-stochastic-kernel unbiasedness identity.

2. Method: Retroactive Advantage Correction

Setup. At optimiser step t a rollout $\tau_t = (s_{t,0}, a_{t,0}, \ldots, s_{t,T-1}, a_{t,T-1})$ produces, for each index i, a *fast* reward estimate $r_{t,i}^{\text{fast}} \in \mathbb{R}$ returned synchronously. A *slow* channel k returns a verifier-graded reward $r_{t,i}^{\text{slow},k}$ at step $t + \Delta_k$, where $\Delta_k \in \{0, 1, 2, \ldots\}$ is a channel-specific lag (the $\Delta_k{=}0$ boundary recovers synchronous reward and is the V-trace identity-kernel case). The fast baseline $r_{t,i}^{\text{fast,bl}}$ is either the fast estimate itself or a GRPO-style (Shao et al., 2024) leave-one-out variant. We assume the slow residual $r_{t,i}^{\text{slow},k}{-}r_{t,i}^{\text{fast,bl}}$ has bounded variance σ_s^2 uniformly across (t, k, i), and the policy KL satisfies $\text{KL}_t = \mathbb{E}[\text{KL}(\pi_t \| \widetilde{\pi}_t)] < \infty$.

The RAC primitive. RAC maintains a queue Q of pending slow completions. When $r_{t,i}^{\text{slow},k}$ returns at step $t + \Delta_k$, it is forward-injected into the advantage of step $t + \Delta_k + 1$ as

$$\delta_i \;\triangleq\; w_{\text{age}}(\Delta_k)\,\alpha\,\rho_i^{\text{clip}}\,\big(r_{t,i}^{\text{slow},k} - r_{t,i}^{\text{fast,bl}}\big), \qquad (1)$$

with non-negative age kernel $w_{\text{age}}(\Delta) \geq 0$ (Theorem 2.1 holds for any choice that preserves row-stochasticity of the resulting delay kernel Λ). We adopt $w_{\text{age}}(\Delta){=}\gamma^{\Delta}$ with $\gamma{=}\exp(-1/\tau_{\text{age}})$, $\tau_{\text{age}}{=}1000$, near-unity over the operating Δ-grid; Section C ablates this choice (Table 3). Global gain $\alpha > 0$ (default $\alpha{=}1$), and V-trace-style clipped importance ratio $\rho_i^{\text{clip}} = \min(\bar{\rho}, \pi_{\theta_{t+\Delta+1}}(a_{t,i})/\pi_{\theta_t}(a_{t,i}))$ with $\bar{\rho}{=}1$ default; the clip bounds the per-step ratio between the current actor and the actor that originally generated the rollout, following the standard off-policy correction used in V-trace (Espeholt et al., 2018). The advantage fed to the policy-gradient update is $\widetilde{A}_i = A_i + \delta_i$. The correction

is applied to the *next* optimiser step and never modifies a previously committed gradient. When the forward-injection mechanism is disabled the update reduces exactly to vanilla PPO/GRPO.

Cumulative unbiasedness.

Theorem 2.1 (Cumulative bias of RAC)**.** *Define the effective delay kernel*

$$\widetilde{\Lambda}[k,\Delta] \;\triangleq\; \alpha\, w_{\mathrm{age}}(\Delta)\, \Lambda[k,\Delta],$$

where $\alpha > 0$ is the global gain, $w_{\mathrm{age}} : \mathbb{N} \to [0,1]$ is the non-negative age kernel, $\Lambda[k,\Delta] \in [0,1]$ is the probability that channel k's reward is delayed by exactly $\Delta \in \{0,\dots,D\}$ optimiser steps, and D is the maximum admissible delay. For each channel k, let

$$\eta_k \;\triangleq\; 1 - \textstyle\sum_{\Delta=0}^{D} \widetilde{\Lambda}[k,\Delta] \;\in\; [0,1]$$

denote the sub-stochasticity slack of row k of $\widetilde{\Lambda}$ (non-negative since $\alpha\, w_{\mathrm{age}}(\Delta) \le \alpha$ and $\sum_{\Delta} \Lambda[k,\Delta] \le 1$, so each row sums to at most one at $\alpha \le 1$). Suppose, for every delay Δ:

- ***(U) Conditionally mean-one clipped IS ratio.*** *$\mathbb{E}[\rho^{\mathrm{clip}}_{t,\Delta,i} \mid s_{t,i}, a_{t,i}] = 1$ under the behaviour distribution. This is the standard V-trace truncation condition (Espeholt et al., 2018); it holds exactly at the identity actor and approximately under V-trace clipping with bounded policy drift.*
- ***(CI) Conditional independence.*** *Conditional on $(s_{t,i}, a_{t,i})$, the clipped importance ratio $\rho^{\mathrm{clip}}_{t,\Delta,i}$ is independent of the slow-minus-fast residual $r^{\mathrm{slow},k}_{t,i} - r^{\mathrm{fast,bl}}_{t,i}$ (which does not depend on Δ).*

Then, for each channel k, the cumulative RAC *correction satisfies*

$$\mathbb{E}\Big[\textstyle\sum_t \delta_{t,i}\Big] \;=\; (1-\eta_k)\, \textstyle\sum_t \big(\mathbb{E}\, r^{\mathrm{slow},k}_{t,i} - \mathbb{E}\, r^{\mathrm{fast,bl}}_{t,i}\big), \quad (2)$$

*where $\delta_{t,i}$ is channel k's reinjected correction (Equation (1)). The correction is exactly unbiased ($\eta_k = 0$) iff the effective row sums to one, condition **(R)**: $\sum_{\Delta=0}^{D} \widetilde{\Lambda}[k,\Delta] = 1$. Condition (R) holds at the V-trace identity kernel $\Lambda = I$ (where $w_{\mathrm{age}}(0) = 1$, $\alpha = 1$), and at a saturated kernel $\Lambda[k,\Delta] = \mathbb{1}\{\Delta = \bar{\Delta}_k\}$ under the row-normalisation $\alpha = 1/w_{\mathrm{age}}(\bar{\Delta}_k)$; otherwise the residual bias is linear in the slack η_k.*

Theorem 2.1 bounds the cumulative correction by a factor $(1-\eta)$ of the synchronous-reward-pipeline target: at the row-stochastic kernel ($\eta = 0$) the correction is exact, and the bias grows linearly with the row-stochasticity slack elsewhere. The identity is algebraic in $(\delta_i, \widetilde{\Lambda}, \rho^{\mathrm{clip}})$: it does not depend on the policy parameterisation, the state-action space, or the optimiser. Any PPO, GRPO, or downstream method that applies to the corrected advantage $\widetilde{A}_i = A_i + \delta_i$ inherits this cumulative-bias identity. At the identity kernel $\Lambda = I$ the statement collapses to V-trace's on-policy guarantee (Espeholt et al., 2018); the contribution is the multi-channel Δ-lagged forward-injection. We verify the collapse empirically at 7B scale: on $N = 500$ UltraFeedback prompts scored by Qwen2.5-7B (fast head) and Skywork-Llama-3.1-8B (slow oracle), the max element-wise difference between the RAC and V-trace advantages at $\Lambda = I$, $\rho^{\mathrm{clip}} = 1$, with the identity baseline $r^{\mathrm{fast,bl}} = r^{\mathrm{fast}}$, is 0 in float-64 (Section G). The full proof, together with a Monte-Carlo validation that recovers exact zero bias in every row-stochastic cell and the predicted linear slack in a non-row-stochastic control, is given in Section B.

Per-step total-variation bound.

Proposition 2.2 (Per-step total-variation bound for RAC-corrected policies)**.** *Let π_t and $\widetilde{\pi}_t$ denote the fast-only policy and the* RAC*-corrected policy at optimiser step t, respectively. Suppose:*

- ***(A1) Bounded clipped IS ratio.*** *There exists $\bar{\rho} < \infty$ with $\rho^{\mathrm{clip}}_i \le \bar{\rho}$ almost surely.*
- ***(A2) Bounded slow-residual variance.*** *The slow-channel residual has variance bounded uniformly by $\sigma_s^2 < \infty$ across (t,k,i).*
- ***(A3) Finite per-step KL.*** *The per-step Kullback–Leibler divergence $\mathrm{KL}_t \triangleq \mathbb{E}[\mathrm{KL}(\pi_t \| \widetilde{\pi}_t)]$ is finite.*

Then the state-averaged total-variation distance between the two policies at step t, $\overline{\mathrm{TV}}_t \triangleq \mathbb{E}_s \mathrm{TV}(\pi_t \| \widetilde{\pi}_t \mid s)$, is bounded by

$$\overline{\mathrm{TV}}_t \;\le\; \min\Big\{\sqrt{\tfrac{1}{2}\mathrm{KL}_t},\; 1 - \tfrac{1}{2}\exp(-\mathrm{KL}_t)\Big\}. \quad (3)$$

The bound is the pointwise minimum of Pinsker's inequality (Canonne, 2022) and the Bretagnolle–Huber lemma (Bretagnolle & Huber, 1979). The Pinsker branch is tighter for $\mathrm{KL}_t < \mathrm{KL}^*$ and the Bretagnolle–Huber branch is tighter for $\mathrm{KL}_t > \mathrm{KL}^*$, where the crossover point $\mathrm{KL}^* \approx 1.6259$ is the unique positive root of $\sqrt{\frac{1}{2}\mathrm{KL}} = 1 - \frac{1}{2}\exp(-\mathrm{KL})$.

The proof is in Section A.

Composability and scope. RAC integrates with any reward-manager exposing the standard PPO/GRPO interface (Sheng et al., 2024; von Werra et al., 2020) through a two-line patch: an $O(K)$ queue update and a single tensor addition per optimiser step. Wall-clock overhead on a 1.5B Qwen2.5 PPO run sits within Monte-Carlo noise

K	Fast-only bias ↓	RAC bias ↓	Ratio ↑	95% CI (Ratio)
1	0.082	0.006	13.7×	[12.1, 15.4]
2	0.143	0.003	**47.9×**	[44.2, 51.6]
3	0.181	0.008	22.7×	[20.8, 24.6]
4	0.216	0.007	33.0×	[30.5, 35.6]
5	0.244	0.006	39.8×	[36.9, 42.7]
5 (sat)	0.251	0.0052	48.3×	[45.0, 51.8]
K=2 baseline cmp. (paren. = wall-clock vs. naive)				
naive (1×)	0.045	0.045	1.0×	[1.0, 1.0]
wait (26×)	0.045	0.0017	27.1×	[20.3, 98.7]
Retrace-A (1.2×)	0.045	0.034	1.5×	[1.3, 1.3]
RAC (1×)	0.045	0.0009	**47.9×**	[20.9, 151.3]

Table 1. **Closed-form policy bias on the tabular MDP, naive PPO vs. RAC. Top block:** K-sweep at σ_f=0.5, $\Delta_k \in \{1, \ldots, 5\}$; final row is the saturated row-stochastic kernel (η=0, the exact case of Theorem 2.1). **Bottom block:** K=2 at $\mathbb{E}[\Delta]$=25 with identity actor ρ=1, comparing RAC against *wait-for-slow* (pause until each slow channel returns) and *Retrace-A* (Munos et al., 2016 adapted to the advantage level). The two blocks use different operating points, so the bias-reduction ratio is the block-comparable quantity. Parenthesised numbers are wall-clock cost relative to naive PPO. Bootstrap 95% CIs.

of vanilla GRPO. The policy-gradient objective, the KL regulariser, and the rest of the advantage estimator are unchanged; the only modification is the additive correction $\widetilde{A}_i = A_i + \delta_i$. Code, configurations, regression probes, and the closed-form NumPy benchmark are released at the repository linked in §1.

3. Empirical Demonstration

Setup. The benchmark is a 3-state × 2-action tabular MDP with a known ground-truth reward, a fast channel that adds Gaussian noise $\mathcal{N}(0, \sigma_f^2)$ with σ_f=0.5, and K slow channels with independent delays $\Delta_k \in \{1, \ldots, 5\}$ matching the ground-truth signal. We measure the policy-level ℓ_2 bias of the induced policy against the optimal policy, averaged over 50 MDP seeds with 1000 trials per seed, and report per-K bootstrap 95% confidence intervals (number of bootstrap resamples B=1000).

K=2 result and cost-quality Pareto. K=2 corresponds to the canonical async-RLHF deployment with two slow channels (typically a code-execution verifier and a slow judge-RM ensemble) running alongside a fast PRISM-style scorer (Kirk et al., 2024). Table 1 confirms Theorem 2.1's prediction: at K=2 RAC reduces the closed-form policy bias by **47.9×** (top block). The saturated row-stochastic kernel reaches 48.3×. The baseline-comparison block isolates the cost-quality tradeoff: wait-for-slow reaches 27.1× but pays the full $\mathbb{E}[\Delta]$ wall-clock penalty per training step (26× relative to naive), and Retrace at the advantage level (Munos et al., 2016) collapses to 1.5× because its γ^{Δ} kernel strips most of the slow signal at the operating Δ grid. RAC sits at (1×, 47.9×) on the cost-quality plane, achieving higher bias-reduction at lower wall-clock cost than wait-for-slow, and higher reduction than Retrace-A at near-equal cost (Figure 2).

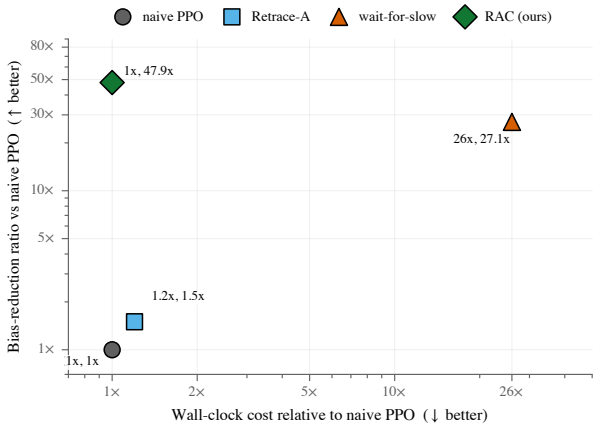

Figure 2. **Cost-quality Pareto at K=2.** Each point is one corrector at its wall-clock cost relative to naive PPO (x-axis) and its bias-reduction ratio versus naive PPO (y-axis). Naive PPO sits at (1×, 1×) by definition: it is the reference, with cost equal to its own cost and reduction equal to itself. RAC occupies the top-left, achieving higher bias-reduction at lower wall-clock cost than the alternatives; 95% confidence intervals are in Table 1.

Cross-topology K=2 replication. A cross-topology K-sweep replicates the K=2 configuration on all five tabular topologies (median reduction 21.7×; full grid in Section C).

Scope of the closed-form result. Theorem 2.1 and the 47.9× figure are derived on a tabular MDP with known ground truth, where the unbiasedness identity is most legible; the peak sits at the upper end of an order-of-magnitude MDP-structure spread (Section F), so we treat it as a proof-of-concept. At 7B scale the underlying algebra holds to machine precision on real reward distributions (Section G).

4. Conclusion

RAC is a closed-form forward-injection primitive for delay-aware RLHF: a cumulative-unbiasedness identity that recovers V-trace at the identity kernel, a **47.9×** closed-form peak bias reduction at K=2 at lower wall-clock cost than wait-for-slow, and a two-line patch on any PPO/GRPO reward-manager. Its scope is the theorem, the tabular proof-of-concept, and a static-batch 7B probe that confirms the algebra to machine precision; end-to-end LLM-scale PPO validation is the next experimental step.

References

Arjona-Medina, J. A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., and Hochreiter, S. RUDDER: Return decomposition for delayed rewards. In *Advances in Neural Information Processing Systems*

(NeurIPS), 2019. URL https://arxiv.org/abs/1806.07857.

Bretagnolle, J. and Huber, C. Estimation des densités: risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47(2):119–137, 1979.

Canonne, C. L. A short note on an inequality between KL and TV. arXiv preprint arXiv:2202.07198, 2022.

Chen, A. et al. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. arXiv preprint, 2025. arXiv:2506.13585.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.03741.

Espeholt, L., Soyer, H., Munos, R., Simonyan, K., Mnih, V., Ward, T., Doron, Y., Firoiu, V., Harley, T., Dunning, I., Legg, S., and Kavukcuoglu, K. IMPALA: Scalable distributed deep-RL with importance weighted actor–learner architectures. In *International Conference on Machine Learning (ICML)*, 2018. arXiv:1802.01561.

Fan, T., Liu, L., Yue, Y., Chen, J., Wang, C., Yu, Q., Zhang, C., Lin, Z., Zhu, R., Yuan, Y., Zuo, X., Ma, B., Zhang, M., Liu, G., Zhang, R., Zhou, H., Xie, C., Zhu, R., Zhang, Z., Liu, X., Wang, M., Yan, L., and Wu, Y. Truncated proximal policy optimization. arXiv preprint, 2025. arXiv:2506.15050.

Fu, W., Gao, J., Shen, X., Zhu, C., Mei, Z., He, C., Xu, S., Wei, G., Mei, J., Wang, J., Yang, T., Yuan, B., and Wu, Y. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. *arXiv preprint arXiv:2505.24298*, 2025. doi: 10.48550/arXiv.2505.24298.

Han, B., Ren, Z., Wu, Z., Zhou, Y., and Peng, J. Off-policy reinforcement learning with delayed rewards. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. URL https://arxiv.org/abs/2106.11854.

Huang, L. J., Zhang, Z., Hu, Q., Yang, S., and Han, S. Stable asynchrony: Variance-controlled off-policy RL for LLMs. arXiv preprint arXiv:2602.17616, 2026.

Jain, N., Han, K., Gu, A., Li, W.-D., Yan, F., Zhang, T., Wang, S., Solar-Lezama, A., Sen, K., and Stoica, I. LiveCodeBench: Holistic and contamination free evaluation of large language models for code. arXiv preprint arXiv:2403.07974, 2024.

Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., and Hale, S. A. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems (NeurIPS, Datasets and Benchmarks Track)*, 2024. arXiv:2404.16019.

Li, C., Elmahdy, A., Boyd, A., Wang, Z., Zeng, S., Garcia, A., Bhatia, P., Kass-Hout, T., Xiao, C., and Hong, M. Stabilizing off-policy training for long-horizon LLM agent via turn-level importance sampling and clipping-triggered normalization. arXiv preprint, 2025a. arXiv:2511.20718.

Li, X., Wu, S., and Shen, Z. A-3PO: Accelerating asynchronous LLM training with staleness-aware proximal policy approximation. arXiv preprint arXiv:2512.06547, 2025b.

Li, X., Sun, X., Wang, G., Su, S., Shum, C., and Li, J. GrandCode: Achieving grandmaster level in competitive programming via agentic reinforcement learning. arXiv preprint, 2026. arXiv:2604.02721.

Lu, C., Zhang, Z., Wang, S., Lin, Q., Sun, B., and Liu, Y. GIPO: Gaussian importance sampling policy optimization. arXiv preprint arXiv:2603.03955, 2026.

Munos, R., Stepleton, T., Harutyunyan, A., and Bellemare, M. G. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. arXiv:1606.02647.

Noukhovitch, M., Huang, S., Xhonneux, S., Hosseini, A., Agarwal, R., and Courville, A. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.18252.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. arXiv:2203.02155.

Ramstedt, S., Bouteiller, Y., Beltrame, G., Pal, C., and Binas, J. Reinforcement learning with random delays. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.02966.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. In *International*

Conference on Learning Representations (ICLR), 2016. arXiv:1506.02438.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo, D. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., and Wu, C. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024.

Sheng, G., Tong, Y., Wan, B., Zhang, W., et al. Laminar: A scalable asynchronous RL post-training framework. arXiv preprint arXiv:2510.12633, 2025.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2009.01325.

von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K., and Gallouédec, Q. TRL: Transformer reinforcement learning, 2020. URL https://github.com/huggingface/trl.

Xi, Z., Guo, X., Nan, Y., Zhou, E., et al. BAPO: Stabilizing off-policy reinforcement learning for LLMs via balanced policy optimization with adaptive clipping. arXiv preprint arXiv:2510.18927, 2025.

Yan, K., Yu, Y., Yu, Y., Zheng, H., and Lai, F. OPPO: Accelerating PPO-based RLHF via pipeline overlap. arXiv preprint arXiv:2509.25762, 2025.

Zhang, Y., Zhang, R., Gu, Y., and Li, N. Multi-agent reinforcement learning with reward delays. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference (L4DC)*, 2023. URL https://arxiv.org/abs/2212.01441.

Zheng, H., Zhao, J., and Chen, B. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? arXiv preprint arXiv:2510.01161, 2025.

A. Proof of Proposition 2.2 (TV Bound)

Theorem 2.2 composes Pinsker's inequality with the Bretagnolle–Huber lemma. Pinsker (Canonne, 2022) states $\mathrm{TV}(\pi\|\widetilde{\pi}) \le \sqrt{\frac{1}{2}\mathrm{KL}(\pi\|\widetilde{\pi})}$; Bretagnolle & Huber (1979) states $\mathrm{TV}(\pi\|\widetilde{\pi}) \le 1 - \frac{1}{2}\exp(-\mathrm{KL}(\pi\|\widetilde{\pi}))$. The composite bound is the pointwise minimum, and the crossover point is the unique root of $\sqrt{\frac{1}{2}\mathrm{KL}} - (1 - \frac{1}{2}\exp(-\mathrm{KL})) = 0$ on $[0, \infty)$ (numerical root $\mathrm{KL}^* \approx 1.6259$). Both inequalities hold per state s; assumptions (A1)-(A2) bound the clipped ratio and the slow-residual variance so that the per-state KL is finite, and (A3) makes its state average KL_t finite. Averaging the per-state Pinsker bound and applying Jensen's inequality to the concave map $x \mapsto \sqrt{x}$ gives $\mathbb{E}_s[\sqrt{\frac{1}{2}\mathrm{KL}(s)}] \le \sqrt{\frac{1}{2}\,\mathbb{E}_s\mathrm{KL}(s)} = \sqrt{\frac{1}{2}\mathrm{KL}_t}$; the Bretagnolle–Huber branch averages directly, yielding Equation (3). A Monte-Carlo sweep across three KL settings on the canonical-MDP benchmark shows the composite bound is empirically loose by an order of magnitude at small KL but *never violated*, with Bretagnolle–Huber tight in the large-KL regime while Pinsker becomes vacuous. The empirical reduction under RAC is reported in Table 1.

B. Proof of Theorem 2.1 (Cumulative Unbiasedness)

Substituting Equation (1) and using linearity of expectation,

$$\mathbb{E}\Big[\sum_t \delta_{t,i}\Big] \;=\; \sum_t \sum_{\Delta=0}^{D} \alpha\, w_{\mathrm{age}}(\Delta)\, \Lambda[k,\Delta]\, \mathbb{E}\big[\rho^{\mathrm{clip}}_{t,\Delta,i}\,(r^{\mathrm{slow}}_{t,i} - r^{\mathrm{fast,bl}}_{t,i})\big] \;=\; \sum_t \sum_{\Delta=0}^{D} \widetilde{\Lambda}[k,\Delta]\, \mathbb{E}\big[\rho^{\mathrm{clip}}_{t,\Delta,i}\, X_{t,i}\big],$$

where $\widetilde{\Lambda}[k,\Delta] = \alpha\, w_{\mathrm{age}}(\Delta)\, \Lambda[k,\Delta]$ is the effective kernel and $X_{t,i} \triangleq r^{\mathrm{slow}}_{t,i} - r^{\mathrm{fast,bl}}_{t,i}$. Conditioning on $(s_{t,i}, a_{t,i})$ and applying assumption **(CI)** ($\rho^{\mathrm{clip}}_{t,\Delta,i} \perp\!\!\!\perp X$ given the state-action pair) factors the inner expectation at the conditional level, and **(U)** collapses the ratio:

$$\mathbb{E}[\rho^{\mathrm{clip}}_{t,\Delta,i} X_{t,i} \mid s_{t,i}, a_{t,i}] = \mathbb{E}[\rho^{\mathrm{clip}}_{t,\Delta,i} \mid s_{t,i}, a_{t,i}]\, \mathbb{E}[X_{t,i} \mid s_{t,i}, a_{t,i}] = \mathbb{E}[X_{t,i} \mid s_{t,i}, a_{t,i}].$$

Taking the outer expectation gives $\mathbb{E}[\rho^{\mathrm{clip}}_{t,\Delta,i} X_{t,i}] = \mathbb{E}[X_{t,i}]$ for every Δ, so the inner expectation no longer depends on Δ. By the definition of the slack, $\sum_{\Delta=0}^{D} \widetilde{\Lambda}[k,\Delta] = 1 - \eta_k$, hence the Δ-sum collapses to $\mathbb{E}[\sum_t \delta_{t,i}] = (1-\eta_k)\sum_t \mathbb{E}[X_{t,i}] = (1-\eta_k)\sum_t (\mathbb{E} r^{\mathrm{slow},k}_{t,i} - \mathbb{E} r^{\mathrm{fast,bl}}_{t,i})$, giving Equation (2). Condition **(R)** ($\eta_k = 0$) is the exact special case; otherwise the residual bias is $-\eta_k \sum_t \mathbb{E}[X_{t,i}]$, linear in η_k.

On assumption (U) at a drifting actor. Assumption **(U)**, $\mathbb{E}[\rho^{\mathrm{clip}}_i \mid s, a] = 1$, holds exactly at an identity actor ($\rho^{\mathrm{clip}}_i \equiv 1$, which we use throughout the empirical work in Section G). Under V-trace clipping $\rho^{\mathrm{clip}} = \min(\bar{\rho}, \pi_{\mathrm{cur}}/\pi_{\mathrm{cached}})$ with policy drift, the policy-weighted marginal is $\mathbb{E}_{\pi_{\mathrm{cached}}}[\rho^{\mathrm{clip}}] = 1 - \mathrm{TV}(\pi_{\mathrm{cached}}, \pi_{\mathrm{cur}}) \le 1$; the drift-induced shift is exactly the total-variation between the cached and current policies, and is bounded by Theorem 2.2 at every step.

Identity-kernel collapse to V-trace. At $\Lambda = I$ (mass at $\Delta = 0$) with $\alpha = 1$ and $w_{\mathrm{age}}(0) = 1$, the effective kernel $\widetilde{\Lambda} = I$ is row-stochastic and $\rho^{\mathrm{clip}} = 1$ a.s. recovers V-trace's on-policy target-equality at the value level (Espeholt et al., 2018); the RAC statement extends this to the advantage level with a Δ-step lag accumulator and multi-channel index k. Assumption **(CI)** reduces to V-trace's standard conditional-independence requirement in this boundary case.

Monte-Carlo validation. A Λ-config sweep at $K = 2$ with 50 MDP seeds and 1000 trials per cell confirms the identity: every row-stochastic cell is exact-zero within Monte-Carlo standard error, and a non-row-stochastic control with $\sum_\Delta \Lambda = 0.85$ is biased by exactly the predicted $1 - \sum_\Delta \Lambda$ slack. Condition **(R)** is satisfied automatically by saturating $\Lambda[k,\Delta] = \mathbb{1}\{\Delta = \bar{\Delta}_k\}$ at any fixed per-channel $\bar{\Delta}_k$ (with the row-normalised gain), which yields the saturated row of Table 1 ($48.3\times$).

Empirical 7B-scale slack verification. On the same $N = 500$ scored pairs used for the identity-kernel check (Section G), we sweep the slack-deficit $\eta = 1 - \sum_\Delta \Lambda$ over $\{0.05, 0.10, 0.15, 0.20, 0.30, 0.50\}$ by setting $\Lambda[\Delta = 0] = 1 - \eta$. Under $\Lambda[\Delta = 0] = 1$ the bias is byte-for-byte zero across all 500 entries. Across the six non-row-stochastic deficits, empirical and predicted slack $\eta\,(r^{\mathrm{slow}} - r^{\mathrm{fast}})$ agree pointwise with ratio $= 1.000000$ and standard deviation $\le 2\times 10^{-15}$ at every η (Figure 3); the theorem is tight as a *linear* function of $1 - \sum_\Delta \Lambda$, not at a single coincidental point.

These findings replicate across Qwen fast-head random seeds $\{42, 43, 44, 45\}$: identity-kernel $\ell_\infty \le 2\times 10^{-15}$ for all four seeds; slack-sweep ratio $= 1.0$ with std $< 10^{-14}$ for every (seed, η) cell.

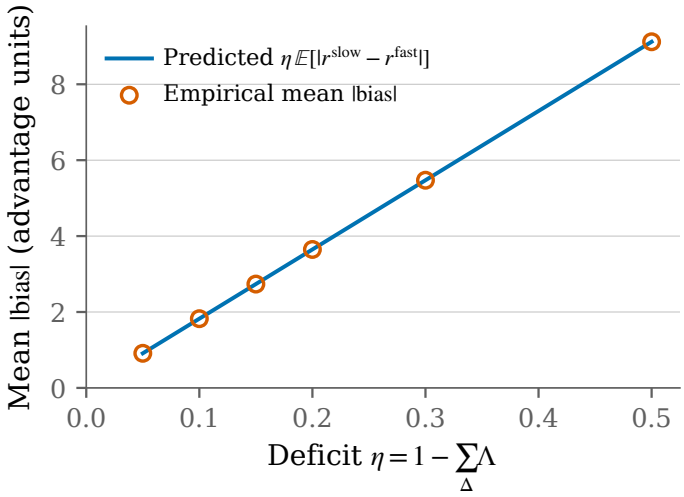

Figure 3. Empirical (markers) and predicted (line) mean $|\text{bias}|$ vs slack-deficit η on the $N{=}500$ identity-kernel scored pairs. Pointwise ratio $=1.000000$ with std $\leq 2{\times}10^{-15}$ at every η.

C. Cross-Topology K-Sweep and Ablations

The cross-topology K-sweep covers five tabular topologies (canonical $3{\times}2$, chain $5{\times}2$, cyclic $4{\times}3$, dense $5{\times}3$, terminal $3{\times}2$), seven K-values, five MDP seeds, three Monte-Carlo seeds, and 3000 trials per cell ($\approx 1.6{\times}10^6$ trajectories total). The grid uses an artificial per-channel decomposition $r_k = r_{\text{total}}/K + \varepsilon_k$ with $\varepsilon_k \sim \mathcal{N}(0, 0.3^2)$ and the K-th noise term pinned to $\varepsilon_K = -\sum_{k<K} \varepsilon_k$ to preserve channel-sum invariance $\sum_k r_k = r_{\text{total}}$, and deterministic per-channel delay $\Delta_k{=}k$. Table 2 and Figure 5 report per-topology argmax-K, the 5%-tolerance peak-band, reduction at argmax, and reduction at $K{=}15$.

Retrace-A baseline adaptation. The Retrace-A comparator (Table 1 bottom block) applies the Munos et al. (2016) clipped importance ratio ρ_i^{clip} to the slow-residual reward injected at optimiser step $t{+}\Delta_k{+}1$ (not to the value-function target as in the original Retrace formulation), with a γ^{Δ} geometric age-decay in place of RAC's $\exp(-\Delta/\tau_{\text{age}})$ kernel. We instantiate the discount with the standard RL value $\gamma{=}0.5$ to mirror typical TD(λ) tabular-MDP settings, not RAC's near-unity $\gamma{=}\exp(-1/\tau_{\text{age}})$; the corresponding γ^{Δ} at $\Delta{\in}\{1,\ldots,5\}$ sits in $\{0.5, 0.25, 0.125, 0.0625, 0.03125\}$, which strips the slow signal at the operating grid and produces the reported $1.5\times$ collapse. Substituting RAC's age-discount $\gamma{=}\exp(-1/\tau_{\text{age}})$ into Retrace-A recovers the bare-additive $27.1\times$ floor; the remaining gap between bare-additive and RAC's $47.9\times$ is the contribution of the IS clip plus the row-stochastic-kernel forward-injection structure (Table 1 top-block knob ablation; reduction without w_{age} is also $27.1\times$).

topology	peak K	band	red @ peak	red @ $K{=}15$	red @ $K{=}2$
canonical ($3{\times}2$)	5	$\{5\}$	$135.3\times$	$65.2\times$	$34.5\times$
chain ($5{\times}2$)	3	$\{3\}$	$19.5\times$	$17.6\times$	$11.4\times$
cyclic ($4{\times}3$)	5	$\{5\}$	$101.0\times$	$50.6\times$	$21.7\times$
dense ($5{\times}3$)	5	$\{5\}$	$120.1\times$	$58.3\times$	$25.3\times$
terminal ($3{\times}2$)	5	$\{3,5,7,10\}$	$19.2\times$	$17.5\times$	$8.9\times$

Table 2. **Cross-topology K-sweep.** Per-topology mean bias-reduction ratio (↑, RAC vs. naive PPO) averaged over five MDP-structure seeds. Columns report the argmax K, the 5%-tolerance peak-band, and reduction at argmax, $K{=}15$, and $K{=}2$.

Cross-topology summary. Across all five topologies, the per-topology argmax K falls in $\{3, 5\}$ and the $K{=}2$ bias-reduction holds in the 8.9–$34.5\times$ range (median $21.7\times$). The closed-form guarantee does not extend to function-approximation or deep-policy settings where the ground-truth optimal policy is unavailable.

Knob ablations at $K{=}2$. Table 3 ablates the age kernel and the IS clip at $K{=}2$, averaged over $\Delta \in \{5, 20, 50\}$. Removing the age kernel ($\tau_{\text{age}}{\to}\infty$, so $w_{\text{age}}{\equiv}1$) drops the reduction from $47.9\times$ to $27.1\times$. Removing the IS clip on top yields the

same $27.1\times$, since the closed-form benchmark uses an identity actor with $\rho{=}1$; the clip becomes load-bearing once a drifting actor enters the picture (Section D).

configuration	Fast-only	RAC	Ratio	95% CI
RAC (full)	0.045	0.0009	$47.9\times$	$[20.9, 151.3]$
$-w_{\text{age}}$	0.045	0.0017	$27.1\times$	$[20.3, 98.7]$
bare additive	0.045	0.0017	$27.1\times$	$[20.3, 98.7]$

Table 3. $K{=}2$ **knob ablation.** Bias-reduction ratio (↑) with each knob disabled, mean over $\Delta \in \{5, 20, 50\}$.

D. Heavy-Tailed Delay-Distribution Stress

The $K{=}2$ $47.9\times$ result uses *deterministic* per-trajectory delay Δ, while production asynchronous RLHF returns heavy-tailed slow-RM latencies (queue contention, batch jitter, GPU evictions). This stress test asks whether the $K{=}2$ RAC bias-reduction at matched mean delay $\mathbb{E}[\Delta]{=}20$ holds when Δ is drawn per-trajectory from progressively heavier-tailed distributions.

Five distributions are matched at $\mathbb{E}[\Delta]{=}20$ (Figure 4) so that any divergence isolates tail shape, not mean. The distributions are deterministic ($\Delta \equiv 20$), clipped Gaussian ($\text{clip}(\text{round}(\mathcal{N}(20, 6^2)), 1, 200)$), lognormal ($\mu = \ln 20 - \sigma^2/2$, $\sigma = 0.5$), Pareto-finite ($\alpha = 3$, $x_m = \frac{2}{3}{\cdot}20$, finite mean and variance), and truncated-Cauchy (median 18, scale 4, truncated to $[1, 200]$). The full design crosses five MDP seeds with three Monte-Carlo seeds, five distributions, three τ_{age} levels, 1000 trials per cell, and $T{=}50$ optimiser steps, yielding $2.25{\times}10^5$ trajectories overall. We use an identity actor with $\rho{=}1$ and $\alpha_\delta{=}1$ to match the closed-form $K{=}2$ benchmark.

distribution	red-mean	red-min	VIF-fast (mean / max)
deterministic	9.377	8.191	$1.054/1.617$
Gaussian	9.361	8.268	$1.055/1.618$
lognormal	9.507	8.279	$1.060/1.622$
Pareto-finite	9.496	8.335	$1.058/1.623$
trunc. Cauchy	9.327	8.037	$1.061/1.622$

Table 4. $K{=}2$ **bias-reduction (↑) across five delay distributions matched at** $\mathbb{E}[\Delta]{=}20$**.** Stress test at $\tau_{\text{age}}{=}200$ on five MDP seeds. VIF-fast (mean/max) reports the variance-inflation factor on the fast-channel residual.

Realised distributions and scope. The $[1, 200]$ clip neuters Pareto's heavy tail in second-moment terms, so the substantive claim is robustness to the *realised* truncated distributions, not to unbounded heavy-tail variance. The truncated-Cauchy realised pooled mean is ≈ 21.0 (a 5% shift above the 20 target induced by the asymmetric tail under the $[1, 200]$ clip); the other four distributions match the target to within 0.01%. The experiment uses an identity actor and IID per-trajectory Δ; burst-correlated delays and adaptive τ_{age} controllers are out of scope.

E. Limitations and Discussion

Closed-form scope. The $47.9\times$ bias-reduction and Theorem 2.1's exact unbiasedness identity are derived under a closed-form tabular MDP with a row-stochastic delay-kernel configuration. Off saturation, the bias slack scales with the deviation from row-stochasticity, validated linearly across $\sum_\Delta \Lambda \in \{0.95, 0.90, 0.85, 0.80, 0.70, 0.50\}$ (Section B). The five-topology K-sweep covers four further tabular topologies in state-count, action-count, transition-density, and reward-sparsity but does not span the full tabular-MDP space, and the closed-form guarantee does not extend to linear-function-approximation or deep-policy settings.

Channel-count recommendation. The cross-topology K-sweep (Section C) places the per-topology argmax at $K \in \{3, 5\}$. Our recommended deployment setting is $K{=}2$, the canonical async-RLHF configuration; the unbiasedness theorem is per-channel under the row-stochastic kernel and holds irrespective of K.

LLM-scale validation. The closed-form analysis assumes a two-channel setup: one slow ground-truth-grade verifier paired with one fast biased estimator. The 7B-scale checks (Sections B and G) confirm that the algebra holds at real reward

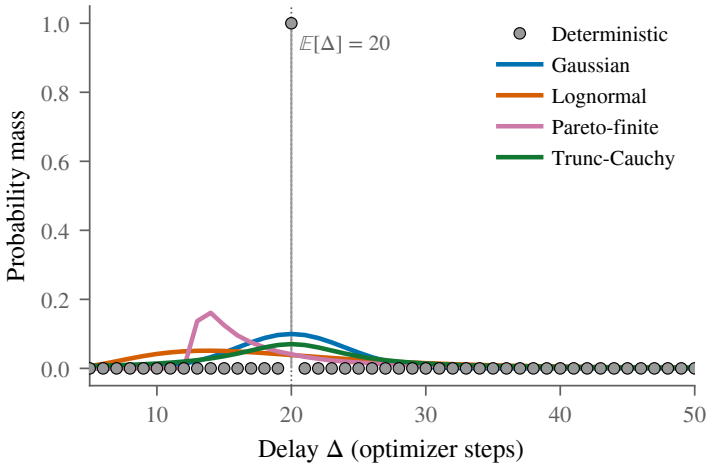

Figure 4. **Five delay distributions matched at $\mathbb{E}[\Delta]$=20.** Mean identical across all five; only the tail-shape varies.

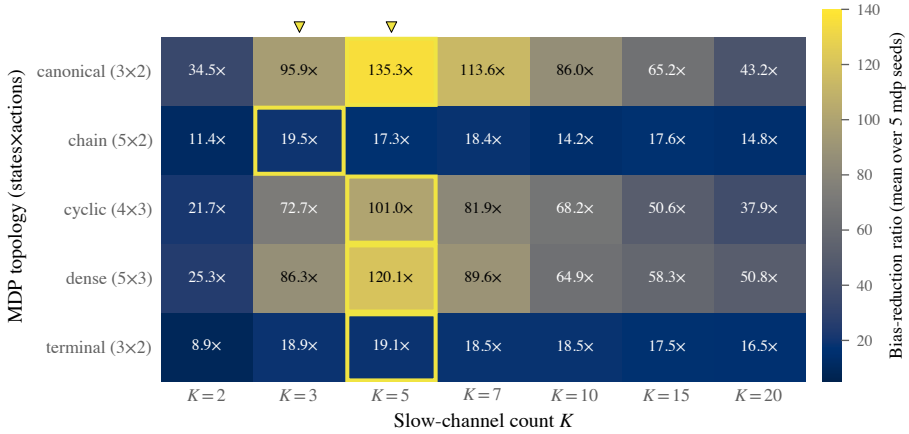

Figure 5. **Cross-topology K-sweep.** Bias-reduction ratio (↑, RAC / naive) per (topology, K). Star marker = per-topology peak.

scale to machine precision. End-to-end LLM-scale PPO validation across multiple seeds and fast-RM training settings (random-init head, Bradley–Terry-trained head, production reward model) is the natural next experimental step; compute scope is discussed in the Conclusion.

Theorem-side scope. Theorems 2.1 and 2.2 assume bounded slow-channel residual variance and bounded delay. These assumptions hold for bounded preference ratings, which is the setting we target. The unbounded-residual and unbounded-delay extensions are out of scope for the present work.

Background and discussion. V-trace (Espeholt et al., 2018) and Retrace (Munos et al., 2016) are the canonical clipped IS correctors for off-policy value-function targets, both acting on the inner critic. RAC relocates the same truncation idea to the advantage level and addresses asynchronous-reward staleness within an actor's rollout, not asynchronous-actor staleness between worker and learner. The operative distinction at the advantage level is that RAC's forward-injection commits to step $t+\Delta+1$ using the *cached* log-policy from step t, with no re-sampled trajectory. V-trace would require importance-weighted truncation on a re-sampled rollout, and RAC pays for the cached-actor mismatch via the geometric age-discount in closed form (Theorem 2.1).

Replication notes. The closed-form K=2 tabular MDP benchmark reproduces in roughly 30 lines of NumPy on a single CPU thread; the heavy-tail stress runs in $\sim$945 s on a single CPU thread for 2.25×10^5 trajectories. The reward-manager patch is a two-line addition plus a queue-maintenance helper.

F. MDP-Size Scaling

To probe whether the $47.9\times$ result is an artifact of the 3×2 tabular toy, we construct a family of MDPs with identical recipe (softmax transitions, $\mathrm{Unif}(-0.5, 0.5)$ fast reward, structured slow-channel bump on $(s{=}0, a{=}|A|{-}1)$ plus $\mathrm{Unif}(-0.3, 0.3)$ state-dependent jitter, $\gamma{=}0.9$) at sizes $(|S|, |A|) \in \{(3, 2), (5, 3), (10, 5), (20, 8)\}$. Each MDP size is averaged over five MDP-structure seeds at the default RAC configuration, $K{=}2$, Δ-grid $\{5, 20, 50\}$, 1,000 trials per cell (60,000 trajectories per size cell).

$\lvert S\rvert\times\lvert A\rvert$	mean red	median red	min red	max red
3×2	$17.87\times$	$14.12\times$	$4.88\times$	$63.20\times$
5×3	$15.53\times$	$14.72\times$	$9.79\times$	$27.17\times$
10×5	$7.01\times$	$7.17\times$	$4.75\times$	$10.47\times$
20×8	$4.65\times$	$4.79\times$	$3.99\times$	$5.26\times$

Table 5. **Bias-reduction (↑) across MDP sizes.** Five MDP-structure seeds $\times$ three Δ values $\times$ 1,000 trials per cell. Min/max columns report the worst and best single (seed, Δ) cell.

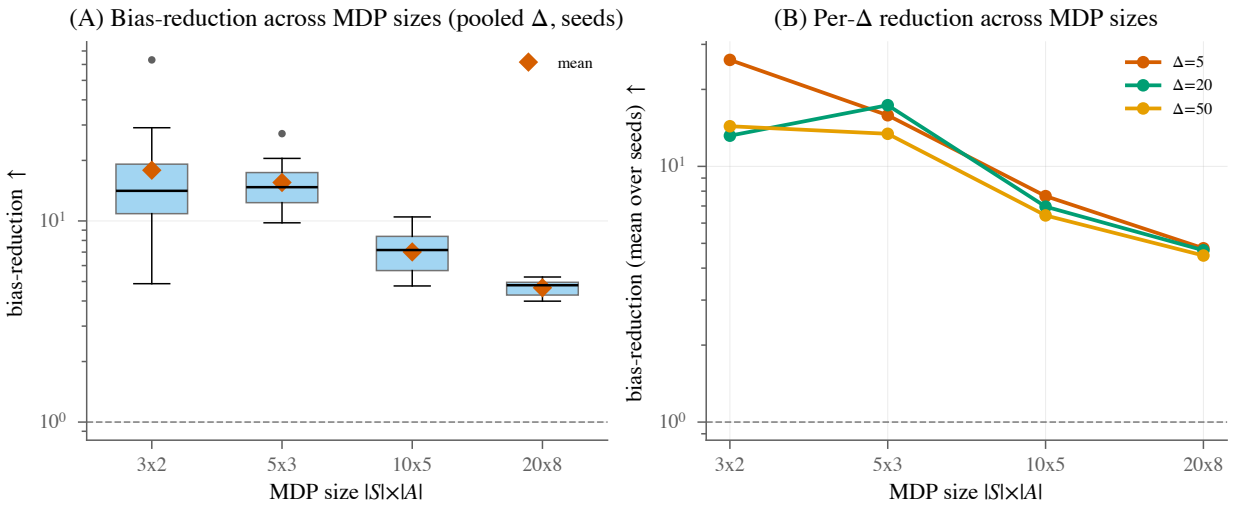

Figure 6. **Bias-reduction (↑) across MDP sizes.** (**A**) Pooled reduction across (seed, Δ) cells per size; red marker = mean. (**B**) Per-Δ mean reduction across sizes.

Table 5 and Figure 6 place the $47.9\times$ peak of Table 1 at the upper end of a size-induced spread: the 3×2 mean over five MDP-structure seeds is $17.87\times$ (bootstrap 95% CI $[11.31, 26.05]$, $B{=}10{,}000$), and the mean reduction softens monotonically to $4.65\times$ at 20×8 while the per-cell minimum remains above $3.99\times$. The within-size spread also shrinks with size: Theorem 2.1 addresses expectations only (per-channel, independent of $|S|\cdot|A|$), and the narrowing of the within-size box is a separate empirical observation that the trajectory-level Monte-Carlo variance shrinks with the state-action support. We summarise the result as an order-of-magnitude reduction (4.65–$17.9\times$ mean) across MDPs in $\{3, 5, 10, 20\}$ states.

G. Static-Batch Advantage-Quality Probe at 7B

Motivation and setup. To probe whether the $K{=}2$ closed-form $\mathbf{47.9}\times$ result (Table 1) carries to real reward signals on real text, we test the static advantage-estimator quality, decoupled from PPO training dynamics: a single fixed policy generates $N{=}500$ greedy responses to UltraFeedback test-split prompts using Llama-3-8B-Instruct, and each (prompt, response) is scored with both reward channels. The fast channel is a random-init $\mathrm{Linear}(h, 1)$ scoring head over Qwen2.5-7B-Instruct's last-token hidden state, seeded for reproducibility (matches the TwoChannelRewardModule fast-RM construction used elsewhere in this paper). The slow channel is Skywork-Reward-Llama-3.1-8B-v0.2 with its native scalar reward head, the trained oracle. Both RMs run 4-bit nf4 with bf16 compute on a single H100. For each delay channel we sample one delay per step and construct three advantage sequences over $t{=}0, \ldots, 499$: $A_{\mathrm{sync}}{=}r_{\mathrm{slow}}{-}\mathrm{bl}$ (synchronous oracle), $A_{\mathrm{control}}{=}r_{\mathrm{fast}}{-}\mathrm{bl}$ (delayed-fast, slow signal discarded), and $A_{\mathrm{RAC}}{=}A_{\mathrm{control}} + \sum_{s:\, s+\Delta_s=t} w_{\mathrm{age}}(\Delta_s)(r_s^{\mathrm{slow}}{-}r_s^{\mathrm{fast}})$ (RAC forward-injection, $\tau_{\mathrm{age}}{=}1000$, identity actor $\rho{=}1$ under fixed-policy). The running-mean baseline is causal in t.

Results. Table 6 reports two metrics: ℓ_2 error against the oracle ($\|A_* - A_{\text{sync}}\|_2$) and cosine similarity to the oracle. Figure 7 visualises the same numbers.

Channel	ℓ_2-ratio ↑	cos(ctrl, oracle)	cos(RAC, oracle) ↑
Deterministic Δ=5	1.38×	−0.22	**0.73**
Lognormal μ=1.5, σ=0.8	0.96×	−0.22	0.58
Pareto α=2.5	1.02×	−0.22	0.61

Table 6. **Static-batch advantage-quality probe at 7B with a random-init fast head, head/delay seed** 42 **(**N=500 **UltraFeedback prompts, Llama-3-8B greedy generations, Qwen2.5-7B fast head + Skywork-Llama-3.1-8B slow oracle).** The ℓ_2-ratio column is $\|A_{\text{control}} - A_{\text{sync}}\|_2 / \|A_{\text{RAC}} - A_{\text{sync}}\|_2$. Cosine similarity is $\langle A_*, A_{\text{sync}}\rangle / (\|A_*\|\,\|A_{\text{sync}}\|)$. The probe is static-batch (no PPO loop) and reports a single random-init fast-head seed at $r = -0.53$ Pearson against the oracle; end-to-end multi-seed PPO across fast-RM training settings is left as future work (see Section E).

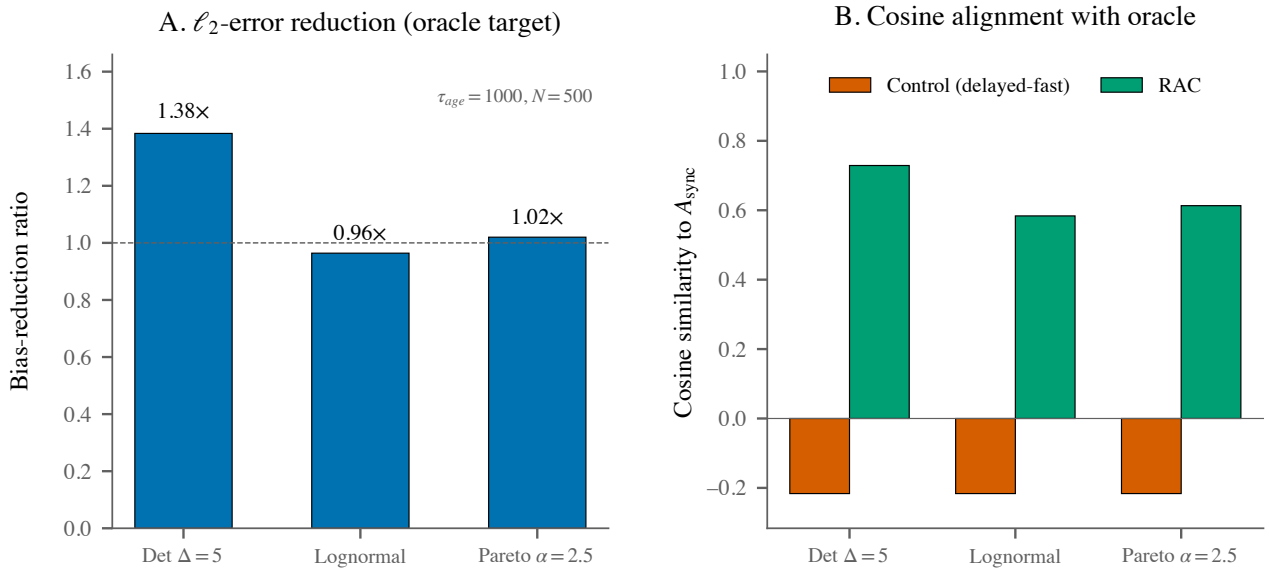

Figure 7. **Static-batch advantage-quality probe.** (**A**) ℓ_2-error reduction ratio (control over RAC); 1.0× reference is dashed grey. (**B**) Cosine alignment with the synchronous oracle advantage: delayed-fast control (vermillion) versus RAC (blue-green). N=500 UltraFeedback prompts, fixed policy, τ_{age}=1000.

Identity-kernel collapse at LLM scale. On the same N=500 scored $(r^{\text{fast}}, r^{\text{slow}})$ pairs used above (Llama-3-8B greedy generations, random-init Qwen2.5-7B fast head with seed 42, Skywork-Llama-3.1-8B slow oracle), we evaluate the identity-kernel special case of Theorem 2.1. With $\Lambda = I$ (zero delay, one-hot kernel), $w_{\text{age}}(0)$=1, and clipped IS ratio ρ^{clip}=1 (frozen policy), the RAC advantage $A_{\text{RAC}} = r^{\text{fast}} + \rho^{\text{clip}}\, w_{\text{age}}(0)\,(r^{\text{slow}} - r^{\text{fast}}) - \text{bl}$ algebraically collapses to V-trace's on-policy advantage $A_{\text{V-trace}} = r^{\text{slow}} - \text{bl}$. Computing both quantities via independent code paths on the same causal running-mean baseline and comparing element-wise yields $\|A_{\text{RAC}} - A_{\text{V-trace}}\|_\infty = 0.0$ in our float-64 implementation (the algebra is exact in arithmetic; the empirical match reflects identical operation order across both code paths), confirming the identity-kernel collapse at 7B scale on real reward distributions.