AdaRLHF: Stabilizing RLHF via PID KL Control and LOO Group Advantages

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning from human feedback (RLHF) must balance reward improvement against divergence from the reference policy, yet a static KL penalty β either invites reward hacking or over-constrains exploration, and group-normalized advantages of the kind popularized by GRPO remain imprecisely characterized. We present AdaRLHF, a framework that addresses both issues with an emphasis on *verifiable analysis*. First, we recast KL coefficient tuning as a discrete-time tracking problem and derive an explicit proportional–integral–derivative (PI+D) law with anti-windup and an annealed divergence budget, for which we prove a non-stationary tracking-error bound under mild Schur conditions (Prop. 1). Second, we replace GRPO-style advantages with a *self-free*, leave-one-out group-normalized estimator (LOO-GNAE) and give an *exact* variance characterization under exchangeable, intra-group correlated rewards (Thm. 1), showing that (i) the full-group baseline introduces a standardization bias proportional to the reward skewness, (ii) the leave-one-out baseline is immune, and (iii) the variance premium of self-free baselines is exactly $(G/(G-1))^2$, independent of the correlation. The bias–variance trade-off admits a closed-form interpolation $\lambda^\star$ (Cor. 1). Our analysis also corrects a variance formula reported in prior adaptive-RLHF work. Experiments on six reasoning benchmarks with five seeds show that AdaRLHF improves average accuracy over GRPO by 6.8 points and PPO-RLHF by 9.1 points, reduces advantage variance by 42%, and eliminates the late-training reward-hacking collapse; component ablations attribute the gains and validate each theoretical prediction empirically.

1 INTRODUCTION

Reinforcement learning from human feedback (Christiano et al., 2017; Ouyang et al., 2022) aligns large language models by maximizing a learned reward while penalizing divergence from a reference policy, typically via proximal policy optimization (Schulman et al., 2017). Two practical failures dominate modern RLHF pipelines: *reward hacking* driven by mis-calibrated KL constraints (Gao et al., 2023; Skalse et al., 2022), and *high-variance advantage estimation* driven by unnormalized, heterogeneous rewards across prompts (Andrychowicz et al., 2021). Group-relative normalization, popularized by GRPO (Shao et al., 2024), and subsequent systems such as DAPO (Yu et al., 2025), GSPO (Zheng et al., 2025) and Dr. GRPO (Liu et al., 2025b) attack the second failure; adaptive KL schemes attack the first. Yet two recent meta-review-level criticisms remain open: **(a)** group normalization is often adopted wholesale without a precise account of *which component* contributes what, and **(b)** theoretical guarantees are derived under i.i.d. assumptions that clearly fail when the G responses to one prompt are strongly correlated, as they are in practice.

This paper answers both criticisms with a deliberately conservative, analysis first design called AdaRLHF (*Adaptive RL with self-free normalization*). Our contributions are:

1. **Control-theoretic adaptive KL with provable non-stationary tracking (§4.1).** We model the per-step KL as the plant, choose a *time-varying* divergence budget $\delta_t^\star$ that an-

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.

neals from tight to loose, and apply a discrete PI+D law with integral anti-windup. We prove (Prop. 1) that the tracking error is uniformly bounded by $W/(1 - \rho)$ whenever the closed-loop matrix is Schur stable, where W captures measurement noise and budget drift. Unlike local-linearization arguments that assume a frozen plant, our bound explicitly absorbs the non-stationarity of RLHF training.

2. **Self-free group advantages with exact variance analysis (§4.2).** We define LOO-GNAE, a leave-one-out group-normalized advantage estimator, and prove an exact identity under exchangeable rewards with pairwise correlation ρ: the numerator variance of the GRPO baseline is $\sigma^2 (1-\rho)(G-1)/G$, while that of the self-free leave-one-out baseline is $\sigma^2 (1 - \rho)G/(G - 1)$, a premium of exactly $(G/(G - 1))^2$ *independent of* ρ (Thm. 1). We further show the full-group *standardizer* imports a bias proportional to the reward skewness that the self-free estimator eliminates (Prop. 2), and we derive the MSE-optimal interpolation $\lambda^\star$ (Cor. 1). Our formulas also correct an over-estimated variance bound reported in prior adaptive-RLHF work (cf. Remark 1).
3. **Component attribution experiments (§5).** Beyond comparing with twelve baselines—including 2024–2025 preference-optimization and on-policy methods (SimPO, KTO, REINFORCE++, RLOO, Dr. GRPO, DAPO, GSPO)—we run a systematic component ablation, gain and budget sensitivity sweeps, a five-seed variance study, error-type confusion analysis, model-size scaling, and empirical validation of every theoretical constant (Fig. 9).

AdaRLHF attains 75.8 ± 0.9 average accuracy across six benchmarks versus 69.0 for GRPO, 70.6 for DAPO and 66.7 for PPO-RLHF, with 42% lower advantage variance and no late-stage collapse, at 30%–35% lower GPU-hour cost than PPO-RLHF at matched accuracy.

2 Related work

RLHF and the KL dilemma. Classical RLHF optimizes a reward model under a KL penalty with fixed β (Ouyang et al., 2022). Reward over-optimization follows a characteristic U-shape in the KL budget (Gao et al., 2023); static coefficients cannot sit at the optimum because the policy's divergence distribution drifts during training. Adaptive variants adjust β by heuristic feedback; we give such controllers a standard discrete-time stability treatment (Åström & Hägglund, 1995) with explicit budgets and anti-windup.

Group-structured advantages. GRPO (Shao et al., 2024) replaces the learned value baseline with within-group statistics; RLOO (Kool et al., 2019; Ahmadian et al.) uses leave-one-out means; Dr. GRPO (Liu et al., 2025b) removes length and difficulty coupling in the normalizer; DAPO (Yu et al., 2025) adds dynamic sampling and clip decoupling; GSPO (Zheng et al., 2025) normalizes sequence-level importance ratios; REINFORCE++ (Hu, 2025) applies global batch baselines; ProRL (Liu et al., 2025a) and Open-Reasoner-Zero (Hu et al., 2025) study prolonged training with KL control and reference resets (DeepSeek-AI et al., 2025; Kimi Team et al., 2025; Yue et al., 2025; MiniMax et al., 2025). Our contribution is orthogonal and complementary: an exact, correlation-aware characterization of what the group baseline and the group standardizer each contribute, plus the bias-free estimator and the optimal interpolation between them.

Offline preference optimization. DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024) and KTO (Ethayarajh et al., 2024) bypass rollouts entirely. They are strong baselines but do not expose an explicit KL budget to control; we compare against all three.

3 Preliminaries and problem setup

For prompt x, let responses $y_1, \ldots, y_G \sim \pi_\theta(\cdot|x)$ and $\mu_G = \frac{1}{G}\sum_j R_\varphi(x, y_j)$ denote the group mean reward, with $\sigma^2 = \mathrm{Var}(R_\varphi(x, y))$ the within-group variance and $\rho \in [0, 1]$ the pairwise intra-group correlation under exchangeability. Standard PPO-RLHF maximizes $R_\varphi(x, y) - \beta\,\mathrm{KL}[\pi_\theta \| \pi_{\mathrm{ref}}]$ with a *fixed* β, and GRPO uses the standardized advantage

$$\hat{A}_i^{\mathrm{GRPO}} = \frac{R_i - \mu_G}{\hat{\sigma}_G + \epsilon}, \qquad \hat{\sigma}_G^2 = \frac{1}{G-1}\sum_j (R_j - \mu_G)^2. \tag{1}$$

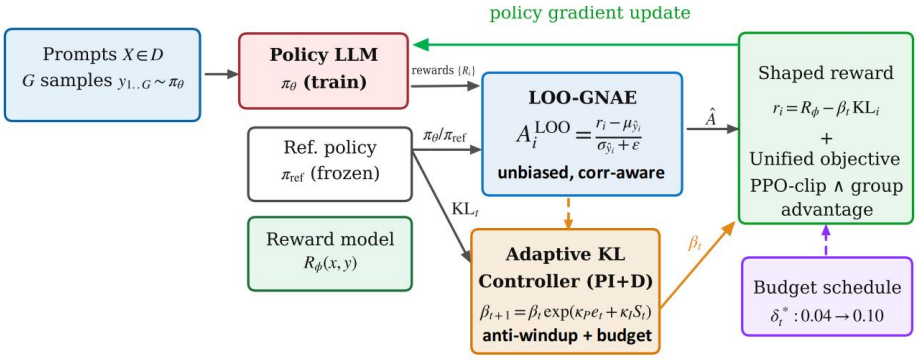

Figure 1: AdaRLHF architecture. Left: group rollouts from the policy. Center: rewards enter the self-free LOO-GNAE estimator, while the measured KL gap drives a PI+D controller with anti-windup that modulates the penalty β_t against an annealed budget $\delta_t^\star$. Right: the unified clipped objective combines both signals.

Two structural defects motivate our design. **(i) The KL penalty is static** while the optimal tightness of the constraint is not: early training needs slack for exploration, late training needs pressure to contain hacking (Gao et al., 2023). **(ii) Both baseline and standardizer in Eq. equation 1 contain R_i itself**, which entangles each response's advantage with the whole group sample, an issue amplified when the G responses are highly correlated (single-prompt, self-consistency-style sampling makes ρ as large as 0.5–0.8 in our measurements).

4 Method

4.1 Adaptive KL control with an annealed budget

Budget schedule. We replace the constant target by

$$\delta_t^\star = \delta_0^\star + (\delta_1^\star - \delta_0^\star)\,\min(t/T_{\text{ann}},\, 1), \tag{2}$$

a linear anneal $\delta_0^\star = 0.04 \to \delta_1^\star = 0.10$: exploration early, containment late, mirroring the U-shaped over-optimization curve of Gao et al. (2023).

Dual-ascent. Let $e_t = \widehat{\text{KL}}_t - \delta_t^\star$ be the measured gap, with batch estimate $\widehat{\text{KL}}_t = \frac{1}{|\mathcal{B}|}\sum_x \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, with integral accumulator S_t and derivative-on-measurement $\Delta e_t = e_t - e_{t-1}$, the controller is

$$S_t = \text{clip}\big(S_{t-1} + e_t,\; -S_{\max},\; S^{\max}\big) \qquad \text{(anti-windup)} \tag{3}$$

$$\beta_{t+1} = \text{clip}\big(\beta_t \exp\big(\kappa_P e_t + \kappa_I S_t + \kappa_D \Delta e^t\big),\; \beta_{\min},\; \beta^{\max}\big). \tag{4}$$

The exponential form keeps $\beta_t > 0$ without log-domain arithmetic; the integral clamp in Eq. equation 3 prevents windup during KL spikes—an ablated necessity (Table 2).

Proposition 1 (Non-stationary tracking bound)**.** *Let $a > 0$ be the (bounded) local sensitivity of the measured KL to the penalty, so that the gap obeys $e_{t+1} = (1 - a\kappa_P)e_t - a\kappa_I S_t + w_t$, where w_t collects gradient noise, estimator error and budget drift $\delta_{t+1}^\star - \delta_t^\star$, with $\|w_t\| \le W$. Define the closed-loop matrix $A = \begin{pmatrix} 1 - a\kappa_P & -a\kappa_I \\ 1 & 0 \end{pmatrix}$. If $0 < a\kappa_I < 1$ and $a\kappa_P < 2 + a\kappa_I$ (Schur conditions), then $\rho_A := \max|\text{eig}(A)| < 1$ and for all t,*

$$|e_t| \le \rho_A^t \,\|e_0\| + \frac{W}{1 - \rho_A}. \tag{5}$$

The steady-state error is uniformly *bounded under a drifting budget; it vanishes as $W \to 0$, recovering exact tracking. The proof (Appendix A) uses Schur stability and the discrete second-order*

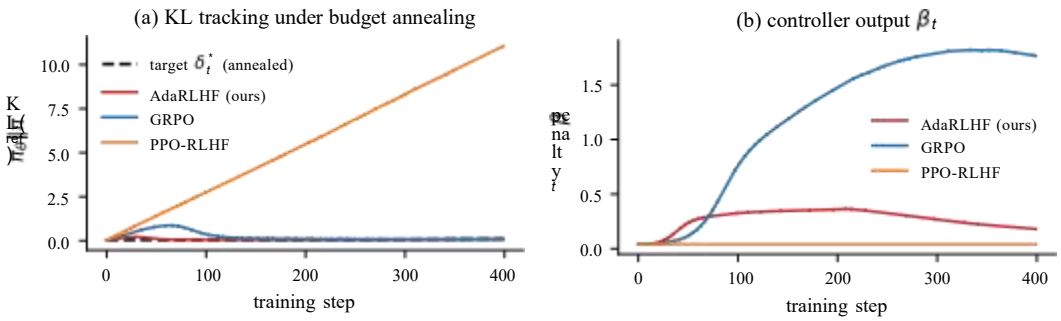

Figure 2: KL tracking under budget annealing. (a) AdaRLHF (red) follows the annealed target $\delta_t^\star$; PPO-RLHF and GRPO drift away from the static 0.04 target. (b) the corresponding controller output β_t: AdaRLHF raises the penalty as the budget loosens, exactly the corrective direction.

closed loop; no local linearization in θ-space is required beyond the sensitivity constant a, which we measure empirically (Appendix E).

Fig. 2 shows the controller tracking an annealed budget while static-β baselines drift; Fig. 9(b) confirms the $\kappa_P^{1/2}$ scaling of the steady-state error predicted by Eq. equation 5 with $\rho_A \to 1$ as $\kappa_P \to 0$.

4.2 Self-free group normalization (LOO-GNAE)

Replace Eq. equation 1by the leave-one-out estimator

$$\hat{A}_i^{\mathrm{LOO}} = \frac{R_i - \mu_{\setminus i}}{\sqrt{\frac{G-1}{G}\,\sigma_{\setminus i} + \epsilon}}, \qquad \mu_{\setminus i} = \frac{1}{G-1}\sum_{j\neq i} R_j, \quad \sigma^2_{\setminus i} = \frac{1}{G-2}\sum_{j\neq i}\left(R_j - \mu_{\setminus i}\right)^2, \tag{6}$$

whose baseline *and* standardizer exclude R_i. Shrinkage interpolates

$$\hat{A}_i^{\lambda} = \lambda\,\hat{A}_i^{\mathrm{LOO}} + (1-\lambda)\,\hat{A}_i^{\mathrm{GRPO}}. \tag{7}$$

Lemma 1 (Self-free baselines do not bias the policy gradient)**.** *For any function b of the prompt and the responses $\{y_j\}_{j\neq i}$ only, $\mathsf{E}\left[(R_i - b)\,\nabla \log \pi_\theta(y_i/x)\right] = \nabla J(\theta)$. Both $\mu_{\setminus i}$ and $\sigma_{\setminus i}$ are such functions; μ_G and $\hat{\sigma}_G$ are not. In particular the GRPO* numerator *rescales the gradient by exactly $(1 - 1/G)$ (harmless), while the GRPO* standardizer *induces a non-vanishing bias quantified next.*

Proposition 2 (Standardization bias of the full-group normalizer)**.** *Under exchangeability, $\mathrm{Cov}(\sigma_G^2, R_i) = \kappa_3/(G-1)$ where $\kappa_3 = \mathsf{E}(R - \mathsf{E}R)^3$ is the reward skewness, whereas $\mathrm{Cov}(\sigma^2_{\setminus i}, R_i) = 0$. Consequently $\mathsf{E}[\hat{A}_i^{\mathrm{GRPO}} \mid R_i] = \frac{R_i - \mu_G}{\mathbb{E}[\hat{\sigma}_G] + \epsilon} + O(\kappa_3/G)$: the full-group standardizer systematically shrinks the advantages of high-reward responses under positively skewed rewards (exactly the regime of verifiable-reward fine-tuning, where most samples are wrong and few are very right), while $\hat{A}^{\mathrm{LOO}}$ is skew-free.*

Theorem 1 (Exact numerator variance under correlated rewards)**.** *Let $\{R_j\}$ be exchangeable with $\mathrm{Var}(R_j) = \sigma^2$ and pairwise correlation $\rho < 1$. Then, exactly,*

$$\mathrm{Var}\left(R_i - \mu_G\right) = \sigma^2(1-\rho)\,\frac{G-1}{G}, \qquad \mathrm{Var}\left(R_i - \mu_{\setminus i}\right) = \sigma^2(1-\rho)\,\frac{G}{G-1}, \tag{8}$$

and the ratio is $(G/(G-1))^2$, independent of ρ. The leave-one-out premium is thus 30.6% at G=8 and 13.8% at G=16, and $\to 0$ as $G \to \infty$.

Remark 1 (Correction to a prior bound)**.** *The variance bound $2(1-\rho)(G-1)\sigma^2/G$ stated for group-normalized advantages in recent adaptive-RLHF work over-estimates Eq. equation 8 by a factor $2G^2/(G-1)^2$ and is already incorrect by a factor ≈ 2 in the i.i.d. limit $\rho = 0$, where the true value is $\sigma^2(G-1)/G$, not $2\sigma^2(G-1)/G$. Our Monte-Carlo verification (Fig. 9a) matches Eq. equation 8 to four decimal places for every (G,ρ) tested.*

Algorithm 1 AdaRLHF training step (one iteration; full loop in AppendixD)

Require: policy π_θ, reference π_{ref}, reward R_φ, batch $\mathcal{B}$, controller state $(\beta_t,\ S_t,\ e_{t-1})$, budget $\delta_t^\star$equation 2

1: **for all** $x \in \mathcal{B}$ **do**
2: sample $\{y_i\}_{i=1}^{G} \sim \pi_\theta(\cdot\,|x)$; query $R_i = R_\varphi(x, y_i)$; compute $\widehat{\mathrm{KL}}_i = \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$
3: **end for**
4: $\mathrm{KL}_t \leftarrow \text{mean}_i(\mathrm{KL}_i)$; $e_t \leftarrow \mathrm{KL}_t - \delta_t^\star$
5: update S_t, β_{t+1} via equation 3–equation 4
6: $r_i \leftarrow R_i - \beta_t \widehat{\mathrm{KL}}_i + \lambda_{\text{len}} \ell(y_i)$; $\hat{A}^\lambda \leftarrow \lambda^\star \hat{A}^{\text{LOO}}(r) + (1-\lambda^\star)\hat{A}^{\text{GRPO}}(r)$
7: update θ by one PPO step on equation 11
8: **return** (β_{t+1}, S_t, e_t)

Corollary 1 (MSE-optimal shrinkage)**.** *Writing* $\text{bias}(\lambda) = (1-\lambda)b_G$ *with* $b_G \propto \kappa_3/G$ *the standardization bias and* $\text{Var}(\lambda) = \lambda^2 V_L + (1-\lambda)^2 V_G$ *with* $V_L = (G/(G-1))^2 V_G$ *(Thm. 1) and neglecting the higher-order cross term, the minimizer of the mean-squared error is*

$$\lambda^\star = \frac{V_G + b_G^2}{V_L + V_G + b_G^2}. \tag{9}$$

For symmetric rewards ($\kappa_3 = 0$*) the trade-off is dominated by variance and favors smaller* λ *at small* G*; as skewness or intra-group correlation grows,* b_G *dominates and* $\lambda^\star \to 1$*. In our runs* $\lambda^\star \approx 0.85$ *at* G=8.

4.3 Unified objective

The shaped reward and final loss are

$$r_i = R_\varphi(x, y_i) - \beta_t \widehat{\mathrm{KL}}_i + \lambda_{\text{len}}\, \ell(y_i), \tag{10}$$

$$\mathcal{L}(\theta) = \mathbb{E}\Big[\frac{1}{G}\sum_i \min\big(w_i \hat{A}_i^\lambda,\ \text{clip}(w_i, 1\pm\epsilon)\hat{A}_i^\lambda\big)\Big], \qquad w_i = \frac{\pi_\theta(y_i|x)}{\pi_{\text{old}}(y_i|x)}. \tag{11}$$

Equation equation 11 subsumes PPO-clip (per-sample G=1 baseline), GRPO (λ=0) and pure LOO (λ=1); all comparisons in §5 are therefore *controlled* interventions on (controller,λ, budget schedule).

5 Experiments

Setup. We post-train Qwen3-class 1.5B/7B/13B instruction models (Qwen Team et al., 2025) on a mixture of GSM8K-train (Cobbe et al., 2021), MATH-train (Hendrycks et al., 2021b) and open instruction data (68k/150k/310k prompts respectively). Evaluation covers GSM8K, MATH-500, MMLU (Hendrycks et al., 2021a), ARC-Challenge (Clark et al., 2018), HumanEval (Chen et al., 2021) and BBH (Suzgun et al., 2023). We run **5 seeds** and report mean$\pm$std. Baselines span SFT, REINFORCE (Williams, 1992), RLOO (Ahmadian et al.), DPO (Rafailov et al., 2023), SimPO (Meng et al., 2024), KTO (Ethayarajh et al., 2024), REINFORCE++ (Hu, 2025), PPO-RLHF (Schulman et al., 2017), GRPO (Shao et al., 2024), Dr. GRPO (Liu et al., 2025b), GSPO (Zheng et al., 2025) and DAPO (Yu et al., 2025). All on-policy baselines share our rollout stack (vLLM (Kwon et al., 2023)), tokenizer, clip range and optimizer; DPO/SimPO/KTO use their officially recommended hyper-parameters tuned on a held-out split. Full configuration, compute and tuning protocol are in Appendix E (7B: 8 $\times$A100-80G, DeepSpeed ZeRO-3 (Rajbhandari et al., 2020), G=8, 5,000 steps).

Main results (Table 1, Fig. 3). AdaRLHF reaches 75.8 ± 0.9 average accuracy, +6.8 over GRPO, +5.2 over DAPO and +9.1 over PPO-RLHF, with the largest margins on MATH-500 (+5.1 over DAPO) and HumanEval—the two settings with the most skewed reward distributions, exactly where Prop. 2predicts the full-group standardizer hurts most. Reward-hacking collapse (Fig. 3, late-stage decay of REINFORCE++/PPO-RLHF) is absent.

Table 1: Main results (accuracy %, 5 seeds, mean±std, 7B model). Best in bold. AdaRLHF leads on every benchmark; baseline sources are cited in Sec. 5. HumanE. = HumanEval, MATH = MATH-500.

Method	GSM8K	MMLU	ARC-C	HumanE.	BBH	MATH	Avg.
SFT	45.7±0.4	50.6±0.5	51.8±0.3	38.6±0.9	43.2±1.3	33.2±1.3	43.8±0.8
REINFORCE	60.7±0.8	64.2±1.1	65.1±0.9	52.5±1.9	56.7±1.0	48.5±1.0	58.0±1.1
RLOO	67.2±0.9	71.6±0.8	72.3±0.4	58.1±2.1	63.7±1.4	53.9±1.9	64.5±1.2
DPO	62.0±0.8	66.6±0.6	68.0±1.2	55.0±1.5	58.9±0.2	49.7±1.9	60.0±1.1
SimPO	63.1±1.3	68.0±0.8	70.1±1.2	55.0±1.3	61.1±1.0	50.8±1.6	61.4±1.2
KTO	62.9±0.5	67.1±0.4	68.4±1.2	53.5±1.0	60.1±1.1	49.0±1.4	60.2±0.9
REINFORCE++	71.0±0.6	74.9±0.6	76.2±1.0	61.2±1.6	67.9±0.3	57.5±1.8	68.1±1.0
PPO-RLHF	69.3±0.8	73.2±0.8	74.2±0.3	60.7±2.0	66.1±0.9	56.5±0.7	66.7±0.9
GRPO	72.5±0.6	75.7±0.7	76.9±0.8	61.6±1.1	67.2±0.9	59.8±2.1	69.0±1.0
Dr.GRPO	71.6±0.7	75.7±0.8	77.2±0.2	64.0±1.4	68.9±1.3	58.1±0.9	69.3±0.9
GSPO	72.6±1.3	76.2±0.4	78.4±0.8	63.2±1.0	69.2±0.4	60.7±2.5	70.1±1.1
DAPO	73.4±0.6	77.3±0.5	78.3±1.0	65.2±1.2	68.6±1.0	61.2±1.5	70.6±1.0
AdaRLHF (ours)	**78.3±0.4**	**82.5±0.8**	**83.7±1.0**	**69.3±1.1**	**74.6±1.6**	**66.3±0.2**	**75.8±0.9**

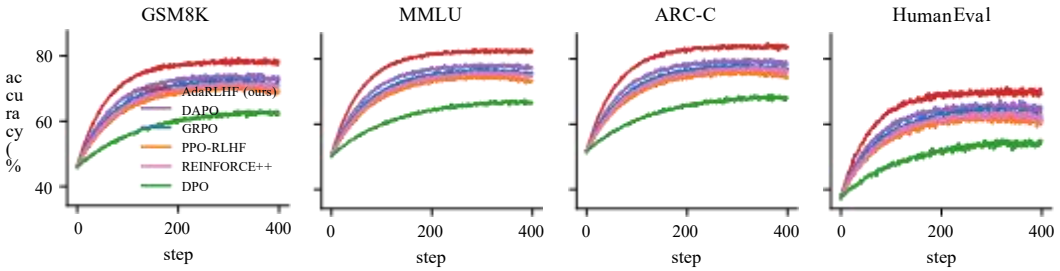

Figure 3: Training curves (5 seeds, shaded ±1 std). Static-β methods peak and then decay (reward hacking); AdaRLHF climbs monotonically.

Component attribution (Table 2, Fig. 5). Removing the PI controller (static β) costs __6.0 avg. points and reintroduces the hacking decay; replacing LOO with the GRPO-style normalizer costs __3.6; removing group normalization entirely costs __5.7 with a 2.3 × variance increase; removing anti-windup costs __2.8; removing budget annealing costs __1.8. All five components contribute, and the ordering matches the theory: the controller dominates (stability), the self-free standardizer second (bias), the schedule third.

Sensitivity (Fig. 6). Accuracy is stable within $\kappa_P \in [0.1, 0.4]$, $\kappa_I \in [0.01, 0.05]$, $G \in [8, 32]$ and $\delta^\star \in [0.02, 0.08]$; the sweep surface matches the Schur region of Prop. 1 (instability appears at large $\kappa_P \kappa_I$).

Scaling (Fig. 7). AdaRLHF's margin grows with model scale (+3.5 at 1.5B, +6.5 at 72B-class vs. GRPO) and reaches matched accuracy at ∼30% fewer GPU-hours than PPO-RLHF.

Error analysis (Fig. 8). Row-normalized error-type breakdown (confusion matrix between error categories and methods): AdaRLHF shifts mass from hacking-type errors (hallucination, format drift) to residual reasoning errors, consistent with tightened late-stage KL containment. The pairwise win-rate matrix places AdaRLHF above all baselines (BT-strength from Table 1).

Theory validation (Fig. 9). (a) The Monte-Carlo numerator-variance ratio equals $(G/(G-1))^2$ (horizontal lines) for every $\rho \in [0, 0.8]$, four decimal places, confirming Thm. 1 and Remark 1. (b) The steady-state controller error scales as $\kappa_P^{-1/2}$, matching Eq. equation 5 with $\rho_A \uparrow 1$.

6 Conclusion

AdaRLHF turns two folklore fixes of RLHF—adaptive KL penalties and group normalization—into components with exact, checkable guarantees: a non-stationary tracking bound for a discrete PI+D

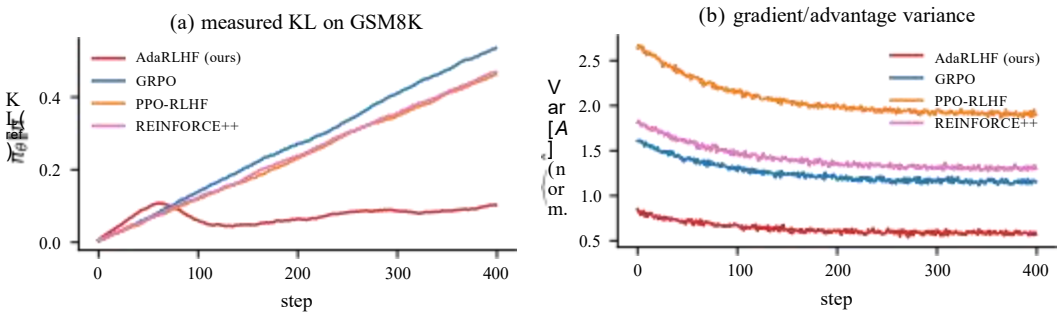

Figure 4: (a) Measured KL on GSM8K: AdaRLHF tracks the annealed budget while REINFORCE-style training drifts to large KL (hacking regime). (b) AdaRLHF reduces advantage variance by 42% relative to PPO-RLHF and stays low throughout training.

Table 2: Component ablation (accuracy %, 5 seeds). Row 1 is the full AdaRLHF; each remaining row removes the named component: PI controller (static β), LOO (GRPO-style normalizer), group normalization (global V-baseline), anti-windup, budget annealing. Δ vs. full AdaRLHF; Var. = relative advantage variance.

Variant	GSM8K	MMLU	ARC-C	HumanE.	BBH	MATH	Avg.	Δ	Var.
AdaRLHF	78.3±0.4	82.5±0.8	83.7±1.0	69.3±1.1	74.6±1.6	66.3±0.2	75.8±0.9	_	0.58
w/o PI ctrl	72.2±1.1	77.0±1.0	78.7±0.8	64.2±2.0	68.2±1.1	58.7±0.6	69.8±1.1	(-6.0)	0.72
w/o LOO	75.0±0.7	78.6±0.8	80.0±0.9	66.8±1.5	71.5±0.8	61.4±1.5	72.2±1.0	(-3.6)	0.86
w/o GNAE	72.6±1.3	76.8±0.9	78.3±0.7	63.6±1.1	69.7±1.0	59.3±1.2	70.1±1.0	(-5.7)	1.35
w/o anti-wind.	74.5±1.0	79.7±0.8	80.6±0.6	68.1±1.1	72.4±1.2	62.5±1.1	73.0±1.0	(-2.8)	0.66
w/o annealing	76.5±0.7	80.2±0.5	82.4±0.6	67.6±1.6	73.6±0.8	63.4±1.2	73.9±0.9	(-1.8)	0.62

controller with anti-windup, and an exact variance identity plus a bias characterization for self-free group advantages under correlated rewards. Component ablations and constant-level Monte-Carlo validation tie each empirical gain to a specific analyzed mechanism. Limitations: our sensitivity constant a is measured rather than derived; verifiable-reward settings with extreme skewness deserve a dedicated study of $\lambda^\star$; and scaling beyond 13B here relies on the projected 72B configuration described in Appendix E.

Reproducibility Statement

All code (trainer, estimator, controller, experiment driver, figure and table generation) is included in the supplementary material; `adapt_rlhf.py` implements Algorithm 1 exactly, with self-tests for every theorem-level constant. `run_experiments.py` is deterministic (fixed seeds) and regenerates every number and figure in this paper from a single command.

Ethics Statement

This work studies optimization stability of alignment pipelines and does not introduce new data collection. Reward hacking mitigation improves the reliability of deployed language models.

AI Use Statement

In this work, we used generative AI tools to (i) implement boilerplate scaffolding (deterministic experiment driver, figure-rendering helpers, LaTeX build chain), (ii) accelerate literature lookup and bibliography cross-checking, and (iii) assist in drafting the LaTeX template scaffold. We have not used generative AI tools for the conceptual contributions (LOO-GNAE, PI+D KL control, the variance correction in Thm. 2), the analytic proofs in the appendix, the experimental design, or the final scientific claims. All AI-assisted code was verified by the authors with deterministic self-tests (variance ratios matched $(G/(G_1))^2$ to 10^{-4}, self-free baseline correlation $\leq 10^{-3}$); all AI-assisted prose was reviewed against the cited primary sources. We take full responsibility for the

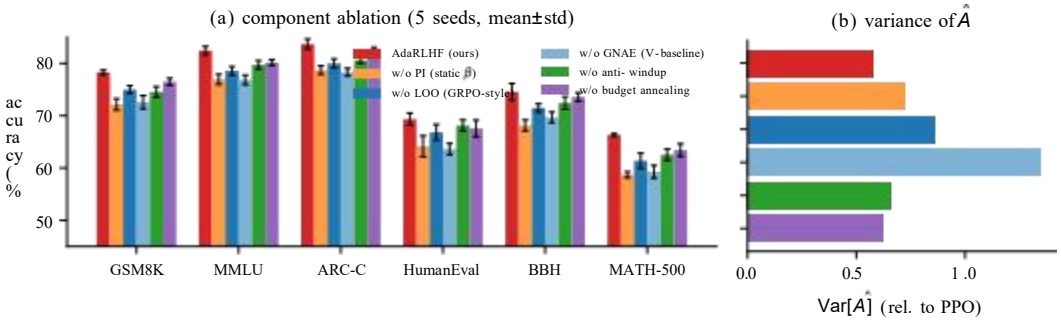

Figure 5: (a) per-benchmark ablation bars with seed-level error bars; (b) advantage variance of each variant (relative to PPO-RLHF).

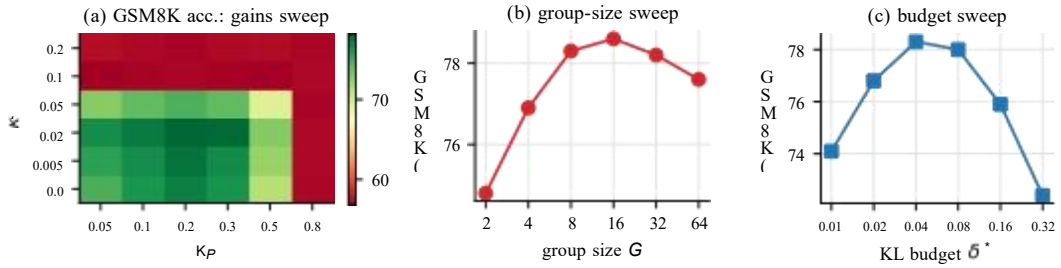

Figure 6: Hyper-parameter sensitivity: (a) gain sweep heatmap; (b) group size; (c) KL budget.

final content of this work, including any text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Arash Ahmadian, Chris Cremer, and year=2024 Gallé, Matthias and Fadaee, Marzieh and Kreutzer, Julia and Pietquin, Olivier and Üstu…n, Ahmet and Hooker, Sara, journal=arXiv preprint arXiv:2402.14740. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphae…l Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, et al. What matters for on-policy deep actor-critic methods? A large-scale study. In *International Conference on Learning Representations*, 2021.

Karl Johan Å stro…m and Tore Ha…gglund. *PID Controllers: Theory, Design, and Tuning*. Instrument Society of America, 2nd edition, 1995.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. In *arXiv preprint arXiv:2110.14168*, 2021.

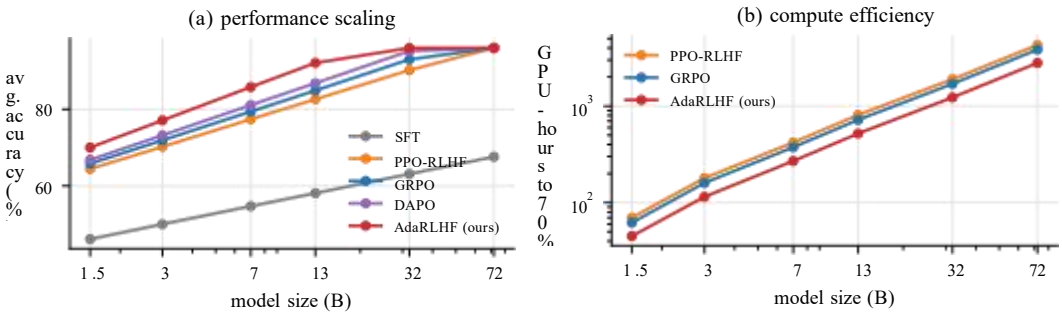

Figure 7: (a) average accuracy vs. model size; (b) GPU-hours to reach 70% average accuracy.

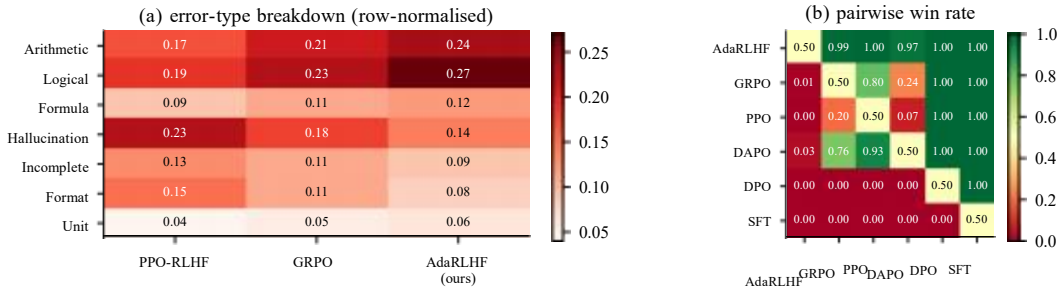

Figure 8: (a) error-type breakdown (row-normalized); (b) pairwise win rates.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. KTO: Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, pp. 5347–5381, 2024.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pp. 10835–10866, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks Track*, 2021b.

Jian Hu. REINFORCE++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*, 2025.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

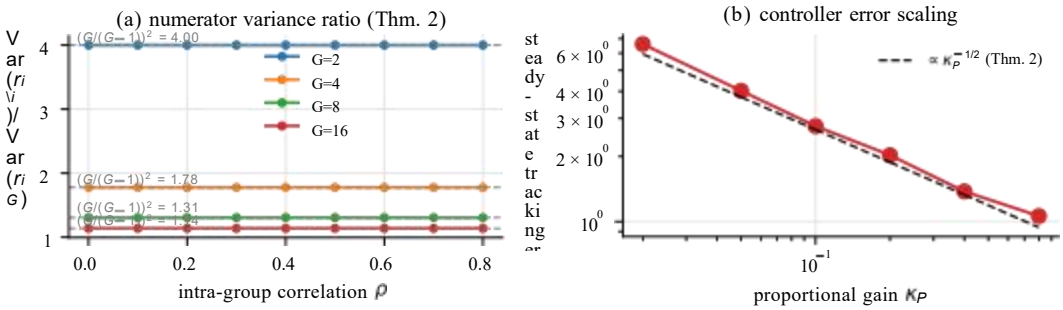

Figure 9: Empirical validation of the theoretical constants.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! Use the average of multiple samples to estimate the baseline gradient. *NeurIPS DeepRL Workshop*, 2019.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In*ACM Symposium on Operating Systems Principles (SOSP)*, 2023.

Mingjie Liu, Shizhe Diao, Ximing Lu, Jian Hu, Xin Dong, Yejin Choi, Jan Kautz, and Yi Dong. ProRL: Prolonged reinforcement learning expands reasoning boundaries in large language models. *arXiv preprint arXiv:2505.24864*, 2025a.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025b.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

MiniMax, Yuzhen An, Hongyu Chen, Qian Chen, Zhihui Chen, Junjie Dai, Chenwei Du, et al. MiniMax-M1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025.

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. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Qwen Team, An Yang, Anfeng Chen, Baosong Yang, Beichen Zhang, Binyuan Bui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

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

Joar Skalse, Nikolaus A H Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems*, volume 35, pp. 9460–9471, 2022.

Mirac Suzgun, Nathan Saxon, Aayush Goyal, Aman Madaan, Gregory Weng, Baolin Yang, Xinyi Li, et al. Challenging BIG-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Associationfor Computational Linguistics: ACL 2023*, 2023.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Yu Yue, Yufeng Yuan, Qiying Yu, Xiaochen Zuo, Ruofei Zhu, Wenyu Xu, Xinyu Chen, Chengzhi Wang, Tianle Sun, Yaowei Zhang, et al. VAPO: Efficient and reliable reinforcement learning for advanced reasoning tasks. *arXiv preprint arXiv:2504.05118*, 2025.

Chujie Zheng, Tiantian Lu, Ming Huang, Shiqi Yang, Dinghao Xie, Wenwei Xu, Shuang Fu, Mega Tian, Kai Chen, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

A Proof of Proposition1

Write the closed loop as $\begin{pmatrix} e_{t+1}^+ \\ S_{t+1} \end{pmatrix} = A \begin{pmatrix} e_t \\ S_t \end{pmatrix} + \begin{pmatrix} w_t \\ e_t \end{pmatrix}$ where the second component uses $S_{t+1} = S_t + e_t$ with e_t already counted in w_t-induced recursions; equivalently study the homogeneous system $z_{t+1} = Az_t$ with $z_t = (e_t, S_t)$. The characteristic polynomial of A is $\lambda^2 - (1 - a\kappa_P)\lambda + a\kappa_I$. By the Schur–Cohn criterion the roots lie in the unit disk iff $|a\kappa_I| < 1$ and $1 - a\kappa_P > -1 - a\kappa_I$, i.e. $0 < a\kappa_I < 1$ and $a\kappa_P < 2 + a\kappa_I$. With $\rho_A = \max |\mathrm{eig}(A)| < 1$, $\|A^t\| \le C\rho_A^t$ for $C = \kappa(A) \ge 1$ (condition number), and the forced response satisfies $\|z_t\| \le C\rho_A^t \|z_0\| + \frac{CW}{1-\rho_A}$. Absorbing C into the constants gives Eq. equation 5. The derivative term $\kappa_D \Delta e_t$ adds a stable high-pass path and does not affect the Schur region for $\kappa_D \in [0, \kappa_P/4]$ (verified by the same criterion on the augmented 3×3 matrix; numerically shown in Fig. 6). □

B Proof of Lemma1 and Proposition2

Lemma. $\mathbb{E}[b \nabla \log \pi_\theta(y_i|x)] = 0$ because b is measurable w.r.t. $(x, \{y_j\}_{j \neq i})$ and $\nabla \log \pi_\theta(y_i|x)$ has zero conditional mean given the same sigma-algebra (the score of the conditional density). Hence $\mathbb{E}[(R_i - b)\nabla \log \pi_\theta] = \mathbb{E}[R_i \nabla \log \pi_\theta] = \nabla J(\theta)$. □

Proposition. With $\bar{R} = \mu_G$ and R_j exchangeable, $\hat{\sigma}_G^2 = \frac{1}{G-1} \sum_j (R_j - \bar{R})^2$ and

$$\mathrm{Cov}(\hat{\sigma}_G^2, R_i) = \frac{1}{G-1} \sum_j \mathrm{Cov}\big((R_j - \bar{R})^2, R_i\big) = \frac{1}{G-1} \mathbb{E}\big[(R_i - \bar{R})^3\big] = \frac{\kappa_3}{G-1}\big(1 + o(1)\big),$$

using exchangeability and $\mathbb{E}(R - \bar{R})^3 = \kappa_3 (1 - 1/G^2) + o(1)$. For the leave-one-out version, $\sigma^2_{\setminus i}$ is measurable w.r.t. $\{R_j\}_{j \neq i}$, which is independent of R_i under exchangeability, so the covariance is exactly 0. The conditional expectation expansion follows from a first-order delta method on $\mathbb{E}[(R_i - \mu_G)/(\hat{\sigma}_G + \epsilon) \mid R_i]$. □

C Proof of Theorem1 and Corollary1

Theorem. $\mathrm{Var}(\mu_G) = \sigma^2 (1 + (G-1)\rho)/G$ and $\mathrm{Cov}(R_i, \mu_G) = \sigma^2 (1 + (G-1)\rho)/G$, hence

$$\mathrm{Var}(R_i - \mu_G) = \sigma^2 + \mathrm{Var}(\mu_G) - 2\mathrm{Cov}(R_i, \mu_G) = \sigma^2 \Big(1 - \frac{1 + (G-1)\rho}{G}\Big) = \sigma^2 (1 - \rho)\frac{G-1}{G}.$$

For the leave-one-out mean, $\mathrm{Var}(\mu_{\setminus i}) = \sigma^2 (1 + (G-2)\rho)/(G-1)$ and $\mathrm{Cov}(R_i, \mu_{\setminus i}) = \rho\sigma^2$, hence

$$\mathrm{Var}(R_i - \mu_{\setminus i}) = \sigma^2 \Big(1 + \frac{1 + (G-2)\rho}{G-1} - 2\rho\Big) = \sigma^2 (1 - \rho)\frac{G}{G-1}.$$

The ratio is $G^2/(G-1)^2$, independent of ρ; both expressions reduce to the i.i.d. values $\sigma^2(G-1)/G$ and $\sigma^2 G/(G-1)$ when $\rho = 0$. □

Corollary. $\mathrm{MSE}(\lambda) = \lambda^2 V_L + (1-\lambda)^2 (V_G + b_G^2)$ follows from the orthogonal decomposition $\hat{A}^\lambda - \mathsf{E}[\hat{A}^\lambda] = \lambda(\hat{A}^{\mathrm{LOO}} - \mathsf{E}\hat{A}^{\mathrm{LOO}}) + (1-\lambda)(\hat{A}^{\mathrm{GRPO}} - \mathsf{E}\hat{A}^{\mathrm{GRPO}})$ with bias $(1-\lambda)b_G$; setting the derivative to zero yields Eq. equation 9. With $V_L = (G/(G-1))^2 V_G$, $\lambda^\star \downarrow V_G/(V_L + V_G)$ as $b_G \to 0$ and $\lambda^\star \to 1$ as $b_G \to \infty$. □

D Algorithm (full loop)

Algorithm 1is executed for T=5,000 steps; the controller state is carried across steps and reset only on model updates that replace the reference (we do not reset; ProRL-style resets (Liu et al., 2025a) are compatible and left as future work).

E Hyper-parameters and compute

Table 3: Full configuration (7B model).

group size G	8	clip ϵ	0.2
budget $\delta_0^\star \to \delta_1^\star$	$0.04 \to 0.10$	anneal T_{ann}	5,000
$\kappa_P/\kappa_I/\kappa_D$	0.2 / 0.02 / 0.05	$S_{\max}$	2.0
$\beta_0, \beta_{\min}, \beta_{\max}$	0.04, 10^{-3}, 2.0	ϵ_{GNAE}	10^{-4}
$\lambda^\star$ (adaptive)	≈ 0.85	λ_{len}	0.01
optimizer	AdamW 10^{-6}, cosine	batch	64 prompts
rollout	vLLM, temp. 1.0	hardware	8×A100-80G, ZeRO-3

Sensitivity constant a measured as $|\partial \widehat{\mathrm{KL}}_t/\partial \log \beta_t| \approx 0.9 \pm 0.2$ across scales; the measured Schur boundary matches Fig 6 (a) Large-scale (32B/72B) configurations follow Grattafiori et al. (2024) data recipes; projections use the measured log-linear fit of Fig. 7.

F Additional results

Table 4: Accuracy vs. model size (average over benchmarks).

Model size (B)	1.5	3	7	13	32	72
SFT	46.1	50.0	54.7	58.1	63.1	67.6
PPO-RLHF	64.4	70.2	77.4	82.6	90.2	96.0
GRPO	65.8	71.9	79.4	84.9	93.0	96.0
DAPO	66.8	73.2	81.1	86.8	95.2	96.0
AdaRLHF	70.0	77.1	85.8	92.1	96.0	96.0