HFC-GRPO: Strict-Lag Selective Advantage Estimation for Zero-Variance Reward Groups

Anonymous authors
Paper under double-blind review

Abstract

Group Relative Policy Optimization (GRPO) assigns zero normalized advantage to homogeneous groups whose responses share the same binary reward, consuming generation compute without supplying a reward-derived policy-gradient signal. We introduce HFC-GRPO, a strict-lag selective advantage estimator: group type fixes intervention and sign, a pre-rollout historical forecast sets homogeneous-group correction magnitudes, and mixed-group advantages remain exactly GRPO. Three fixed historical predictors combine global, question-specific, and recent outcomes, weighted by each predictor's own accumulated Brier loss. Forecasts and advantages remain frozen throughout policy reuse, followed by a single history update. We establish signed-signal recovery and show why all-rollout calibration need not carry over to homogeneous groups selected after a pre-rollout forecast is formed. Accounting for this selection effect and clipping, we bound advantage and gradient-contribution error under explicit conditioning and boundedness assumptions. Across two Qwen3-8B reasoning protocols, paired comparisons favor HFC, including a 0.65-percentage-point Acc@8 gain over RL-ZVP in paired common-harness comparisons under the RL-ZVP protocol. Transfer to multi-turn search agents yields a 2.12-percentage-point mean exact-match gain over paired outcome-only GRPO. Matched history and aggregation replacements support aligned pre-rollout history and adaptive predictor weighting within this selective estimator.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has become a practical route to improving mathematical reasoning because correctness can often be checked without a learned reward model (Shao et al., 2024; Guo et al., 2025). GRPO (Shao et al., 2024) makes this route economical by replacing a learned value model with statistics from a group of responses to the same prompt. For binary rewards $r_i \in \{0, 1\}$, however, its standardized advantage vanishes whenever a group is all wrong or all correct. Such groups still consume generation compute, yet provide no reward-derived policy-gradient signal.

Can historical outcomes recover nonzero advantages for homogeneous groups while preserving exact GRPO advantages on mixed groups? Mixed groups already provide within-group reward contrasts. We therefore preserve their advantages and use historical prediction to set correction magnitudes only where those contrasts vanish.

Existing approaches filter uninformative rollouts (Yu et al., 2025; Zheng et al., 2025), perturb verified rewards (Wei et al., 2025), or replace group-local reference statistics (Xu & Ding, 2026; Kim et al., 2026; Chang et al., 2026). The closest branch-selective method, RL-ZVP, preserves mixed-group GRPO and derives homogeneous-group feedback from detached token entropy (Le et al., 2026). HFC instead couples a past-only outcome forecast with endpoint-only intervention and exact mixed-group preservation, keeping forecast formation measurable with respect to pre-rollout history.

Historical Feedback Control for GRPO (HFC-GRPO) predicts h_t from committed pre-rollout history $\mathcal{F}_{t-1}$, assigning $-h_t$ to all-wrong groups and $1 - h_t$ to all-correct groups while preserving mixed-group advantages. Three fixed predictors—global/question shrinkage, last-visit success, and a global–recent average—receive exponential weights from their own accumulated Brier losses. All batch groups share one committed snapshot; forecasts and advantages stay frozen throughout policy reuse, followed by a single history commit (Figure 1).

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.

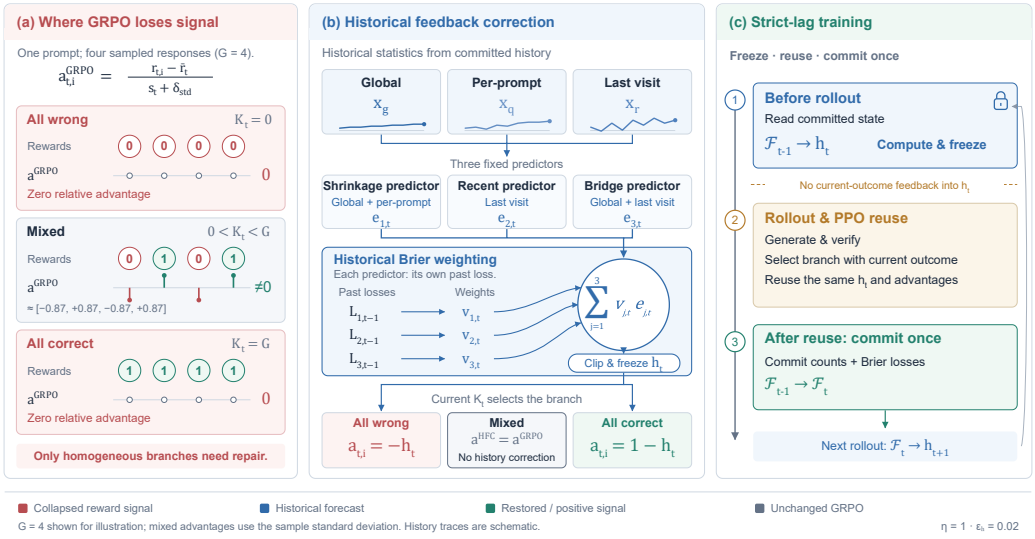

Figure 1: Strict-lag selective recovery in HFC-GRPO. (a) GRPO branches indexed by correct responses K_t. (b) Strict-lag historical forecast and homogeneous-only correction. (c) Freeze-before-generation, commit-after-reuse timeline. Mixed groups retain exact GRPO.

Our contributions are:

1. We introduce a strict-lag selective estimator in which group type fixes intervention and sign while a historical forecast fixes magnitude, restoring homogeneous residuals while preserving exact GRPO on mixed groups.
2. We instantiate the forecast with three lagged predictors scored by their own Brier losses and a freeze–reuse–commit schedule that preserves the pre-outcome information boundary.
3. We characterize the predict-before, gate-after interface by separating all-rollout forecast error from gate-induced conditional shift, then connect their combined error to advantage substitution and bounded gradient-contribution deviation.

We evaluate this design through paired reasoning comparisons and search-agent transfer, with targeted replacements of the history, intervention scope, and aggregation rule (Section 6).

2 Related Work

Group-relative RLVR. GRPO was introduced with DeepSeekMath as a critic-free alternative to proximal policy optimization (PPO) (Schulman et al., 2017; Shao et al., 2024). Classical policy-gradient analyses interpret baselines as control variates for gradient-variance reduction (Greensmith et al., 2004); critic-free variants then replace the learned value with sample-based references. REINFORCE++ avoids prompt-local collapse by replacing group-local normalization with global-batch normalization (Hu et al., 2025). DAPO and Dr. GRPO instead retain group-relative training while improving its recipe through asymmetric clipping, dynamic sampling, token-level losses, and bias corrections such as removing response-length inflation (Yu et al., 2025; Liu et al., 2025b). HFC follows a third route: it preserves the informative mixed-group GRPO computation and introduces history only at the two endpoints where within-group reward variation vanishes.

Selection, perturbation, and exploration. DAPO's dynamic sampling oversamples and filters collapsed groups after generation, while GRESO uses historical outcomes to avoid some rollouts before generation (Yu et al., 2025; Zheng et al., 2025). ERPO combines a pseudo-negative reward that reactivates all-correct groups with a history-conditioned temperature schedule for future visits (Liu et al., 2025a). NGRPO inserts a virtual maximum-reward sample and asymmetric clipping to recover an all-wrong signal, but consequently changes the normalization of mixed groups as well (Nan et al.,

2025). AVSPO monitors an advantage-collapse rate and injects virtual reward samples without additional model rollouts (He et al., 2026). These approaches intervene at selection, generation, reward/advantage construction, or a combination thereof. ReDit acts after verification by adding zero-mean dithering to discrete rewards (Wei et al., 2025). RL-ZVP is the closest hard-gated advantage method: it preserves GRPO for mixed groups and uses detached token entropy to set homogeneous-group feedback (Le et al., 2026). HFC-GRPO also gates on the realized group type, but obtains its scalar magnitude from a forecast frozen before the rollout.

This within-training problem is distinct from the stronger zero-reward barrier in which the base policy has not yet sampled a correct solution (Prakash & Buvanesh, 2025). Curriculum methods address discovery of the first positive outcome; HFC addresses the complementary optimization failure after sampling: even an all-zero batch receives a nonzero negative update from its strictly prior forecast, including during an all-zero prefix through the interior Beta prior. Decompositions of RLVR into positive- and negative-sample reinforcement further show why penalizing incorrect samples can provide a useful reasoning signal (Zhu et al., 2025). HFC therefore restores an update without requiring a current positive sample, while easier-data curricula remain complementary when the policy must first discover successful trajectories.

Cross-group and historical baselines. SPO abandons group-local normalization and estimates a persistent prompt value before batch-global normalization and prioritized sampling (Xu & Ding, 2026). KRPO treats group rewards as noisy observations and replaces the group mean and scale with a lightweight Kalman estimate (Wang et al., 2026). Shrinkage baselines combine prompt-local and across-prompt means to reduce baseline-estimation error (Zeng et al., 2026). GRPO-DBB maintains a discounted Beta–Bernoulli estimate and changes the estimator for every group (Kim et al., 2026). Most closely, BV-Blend combines prompt-local statistics with pre-update semantic-cluster exponential moving average (EMA) moments using an uncertainty-derived confidence weight (Chang et al., 2026). HFC uses a pre-rollout historical success forecast as the reference for homogeneous residuals, while preserving GRPO advantages exactly on mixed groups. Fixed statistical predictors supply this reference; the selective advantage rule determines where it enters the update.

Online prediction perspective. The exponential-weights rule is a standard tool for combining experts under bounded losses (Cesa-Bianchi & Lugosi, 2006). Brier loss is a strictly proper score for binary probabilities (Brier, 1950; Gneiting & Raftery, 2007). Together, they yield an interpretable online controller whose state is frozen before generation, whose predictors adapt to predictive performance, and whose update occurs only after policy reuse is complete.

Agentic reinforcement learning with search. Search-R1 extends outcome-based reinforcement learning to trajectories that interleave model reasoning with multiple search-engine calls, masking retrieved tokens while optimizing PPO or GRPO from final-answer rewards (Jin et al., 2025). The tool-call reward model (TRM) adds a learned tool-call reward and turn-level credit assignment to this stack, exposing a stronger process-supervised comparison under a unified seven-dataset protocol (Ma et al., 2026). HFC targets an orthogonal failure: it retains the outcome-only verifier and changes only the group advantage on homogeneous trajectory groups. This setting tests whether strict-lag signal recovery transfers without an auxiliary reward model.

3 Background and Problem Formulation

For prompt q at generation event t, sample G responses $y_{t,1:G} \sim \pi_{\theta_t}(\cdot \mid q)$ and verify binary rewards $r_{t,i}$. Let $K_t = \sum_i r_{t,i}$, group mean $\bar{r}_t$, and sample standard deviation s_t. Standard GRPO uses

$$a_{t,i}^{\mathrm{GRPO}} = \frac{r_{t,i} - \bar{r}_t}{s_t + \delta_{\mathrm{std}}}. \tag{1}$$

where the fixed stabilizer is $\delta_{\mathrm{std}} = 10^{-6}$. When $K_t \in \{0, G\}$, every numerator is zero and hence every advantage is zero. We call these *homogeneous* groups; groups with $0 < K_t < G$ are *mixed*.

This failure is transparent under the standard centering decomposition. Let $\mathbf{r}_t = (r_{t,1}, \ldots, r_{t,G})^\top$, $Z_t = K_t/G$, $\mathbf{1}_G \in \mathbb{R}^G$, and $P = I - \frac{1}{G}\mathbf{1}_G\mathbf{1}_G^\top$. Then $\mathbf{r}_t = P\mathbf{r}_t + Z_t\mathbf{1}_G$, and the numerator of GRPO is $P\mathbf{r}_t$. Thus $P\mathbf{r}_t = 0$ on a homogeneous group: current within-group contrast vanishes, while the group outcome can still be compared with a pre-rollout reference.

Design requirements for selective GRPO recovery. We design HFC-GRPO to preserve GRPO advantages exactly on mixed groups and restore signed advantages only on homogeneous groups. A forecast computed from committed pre-rollout history determines the correction magnitude; the realized group type selects its use and sign. Forecasts remain fixed throughout policy reuse, and a single post-reuse commit updates the controller through an explicit, replayable state transition.

4 HFC-GRPO

4.1 Selective historical advantage

Let $\mathcal{F}_{t-1}$ contain only outcomes committed before prompt group t is generated. From this history, HFC computes a success forecast $h_t \in [\varepsilon_h, 1-\varepsilon_h]$, with $\varepsilon_h = 0.02$, and freezes it before generation. After observing K_t correct responses, it assigns

$$a_{t,i}^{\mathrm{HFC}} = \begin{cases} a_{t,i}^{\mathrm{GRPO}}, & 0 < K_t < G, \\ -h_t, & K_t = 0, \\ 1 - h_t, & K_t = G. \end{cases} \tag{2}$$

All-wrong groups are penalized in proportion to predicted success, and all-correct groups are reinforced in proportion to predicted surprise. The two homogeneous branches use fixed unit multipliers; mixed groups retain Eq. 1 response by response. The forecast is formed before rewards are observed, whereas the branch is selected afterward.

Equivalently, with $D_t = \mathbf{1}\{K_t \in \{0, G\}\}$ and group advantage vectors, $\mathbf{a}_t^{\mathrm{HFC}} = (1 - D_t)\mathbf{a}_t^{\mathrm{GRPO}} + D_t(Z_t - h_t)\mathbf{1}_G$. The first term preserves every identifiable mixed-group comparison; the second supplies a strict-lag forecast residual only after that contrast has vanished.

In implementation, one information event is a rollout batch b. All prompt groups in $\mathcal{B}_b$ read the same immutable snapshot $\mathcal{F}_{b-1}$; event-wise h_t denotes the corresponding prompt forecast h_{bq}. Forecasts and advantages remain fixed throughout PPO reuse. After reuse is complete for the entire batch, a single commit updates historical counts and the three predictor-loss accumulators. We next specify the forecast and its update.

4.2 Historical statistics and three predictors

A prompt is addressed by a stable dataset-example identifier. The controller stores global successes and trials (S_g, N_g), per-prompt pairs (S_q, N_q), and the previous committed group counts $(K_q^{\mathrm{last}}, G_q^{\mathrm{last}})$. With a symmetric Beta prior $\alpha = 1$, three historical statistics are

$$\begin{aligned} x_g &= \frac{S_g + \alpha}{N_g + 2\alpha}, \\ x_q &= \begin{cases} \frac{S_q + \alpha}{N_q + 2\alpha}, & N_q > 0, \\ x_g, & N_q = 0, \end{cases} \qquad x_r = \begin{cases} \frac{K_q^{\mathrm{last}} + \alpha}{G_q^{\mathrm{last}} + 2\alpha}, & \text{a committed visit exists}, \\ x_g, & \text{otherwise}. \end{cases} \end{aligned} \tag{3}$$

Thus a first visit has the unique fallback $x_q = x_r = x_g$. Let $e_{j,t}$ denote the output of historical predictor j at event t. The shrinkage predictor combines question and global estimates:

$$c_t = \frac{G}{N_q + G}, \qquad e_{1,t} = c_t x_g + (1 - c_t) x_q, \tag{4}$$

where N_q is the number of previously committed samples for q. The recent and bridge predictors output

$$e_{2,t} = x_r, \qquad e_{3,t} = \tfrac{1}{2}(x_g + x_r). \tag{5}$$

Every statistic and predictor output is strictly interior under finite counts. The statistics x_g, x_q, x_r are inputs; shrinkage, recent, and bridge are fixed prediction rules using them (Figure 1). These rules combine pooled, cumulative, and last-visit information without fitting an auxiliary neural network, and serve as experts in exponential weighting. With fixed G, all outputs coincide on a first visit and $e_1 = e_3$ on a second; cumulative and recent histories can separate from the third visit (Appendix D).

4.3 Independent Brier aggregation

Each predictor is scored by its own accumulated Brier loss L_j, initialized at $L_{j,0} = 0$; "independent" refers to these separate loss records, not statistical independence of predictor outputs. At rollout batch b, every $q \in \mathcal{B}_b$ uses the same losses $L_{j,b-1}$ and frozen exponential weights, with prompt-specific predictor values:

$$\begin{aligned} \ell_{j,b-1} &= L_{j,b-1} - \min_k L_{k,b-1}, \quad v_{j,b} = \frac{\exp(-\eta \ell_{j,b-1})}{\sum_{k=1}^{3} \exp(-\eta \ell_{k,b-1})}, \\ \widetilde{h}_{bq} &= \sum_{j=1}^{3} v_{j,b} e_{j,bq}, \qquad\qquad h_{bq} = \mathrm{clip}(\widetilde{h}_{bq}, \varepsilon_h, 1 - \varepsilon_h), \end{aligned} \tag{6}$$

with $\eta = 1$ and $\varepsilon_h = 0.02$. Subtracting the minimum loss preserves the weights and anchors the largest exponential at one; implementation uses log-domain normalization. Clipping gives $|a^{\mathrm{HFC}}| \geq \varepsilon_h$ on either homogeneous branch. After generation and all PPO reuse steps, let $z_{bq} = K_{bq}/G$. The batch is committed once with the mean-loss update

$$L_{j,b} = L_{j,b-1} + \frac{1}{|\mathcal{B}_b|} \sum_{q \in \mathcal{B}_b} (e_{j,bq} - z_{bq})^2. \tag{7}$$

Mean aggregation makes the learning rate comparable across the 128-, 256-, and 512-prompt protocols. The same commit updates sufficient statistics in stable prompt order and advances the controller version once. Appendix B gives the core training loop, bounded-memory execution, and complete replay protocol in Algorithm B.1.

5 Analysis

Proposition 1 (selective invariance and recovery). For every mixed group, Eq. 2 gives $a_{t,i}^{\mathrm{HFC}} = a_{t,i}^{\mathrm{GRPO}}$ response by response. For every homogeneous group, Eq. 6 gives $|a_{t,i}^{\mathrm{HFC}}| \geq \varepsilon_h$: the all-zero branch is negative and the all-one branch is positive. Thus group type fixes the intervention locus and sign, while history can affect only the homogeneous correction magnitude; every informative within-group comparison remains intact.

Proposition 1b (group-level scale compatibility). Let $p = K/G \in (0, 1)$ for a mixed binary-reward group and let $s = \sqrt{\frac{G}{G-1} p(1-p)}$ be its sample standard deviation. The group-mean absolute GRPO advantage is

$$\frac{1}{G} \sum_i |a_i^{\mathrm{GRPO}}| = \frac{2p(1-p)}{s + \delta_{\mathrm{std}}} \leq \sqrt{\frac{G-1}{G}} < 1. \tag{8}$$

On a homogeneous HFC group, the same quantity is h or $1 - h$, which lies in $[\varepsilon_h, 1 - \varepsilon_h]$. Thus HFC restores group-level L_1 mass on the same bounded $O(1)$ scale as mixed-group normalization, without an additional gain parameter. Individual mixed responses can exceed one when one reward class is rare; Eq. 8 concerns the group-average mass that controls the aggregate update budget.

The branch rule fixes the sign of the homogeneous correction; the forecast determines its magnitude. The forecast is formed before rollout, but the gate selecting its use is observed only after rollout. Consequently, calibration on the full stream need not equal calibration on the gated homogeneous population. We therefore separate ordinary all-rollout forecast error from the conditional shift induced by the gate, then trace their combined error to the advantage and gradient contribution.

Proposition 2 (selection-aware forecast-to-gradient bridge). Let $Z = K/G \in [0, 1]$, $p_0 = \mathbb{E}[Z \mid \mathcal{F}_{t-1}]$, $D_t = \mathbf{1}\{K \in \{0, G\}\}$, and, whenever $\Pr(D_t = 1 \mid \mathcal{F}_{t-1}) > 0$, $p_H = \mathbb{E}[Z \mid \mathcal{F}_{t-1}, D_t = 1]$. The all-rollout Brier excess of a frozen, unclipped forecast $\widetilde{h}$ over p_0 is

$$\mathbb{E}[(\widetilde{h} - Z)^2 \mid \mathcal{F}_{t-1}] - \mathbb{E}[(p_0 - Z)^2 \mid \mathcal{F}_{t-1}] = (\widetilde{h} - p_0)^2. \tag{9}$$

The gate changes the oracle from p_0 to p_H. Their interface is explicit:

$$|h - p_H| \leq |h - \widetilde{h}| + |\widetilde{h} - p_0| + |p_0 - p_H|, \tag{10}$$

where the three terms are clipping displacement, all-rollout forecast error, and gate-induced conditional shift. Conditioning the Brier identity on $D_t = 1$ directly evaluates their combined operational error. Define the oracle-gated advantage by replacing h with p_H in Eq. 2. For every sampled response,

$$|a^{\mathrm{HFC}} - a^{p_H}| = D_t|h - p_H|, \qquad \|g^{\mathrm{HFC}} - g^{p_H}\| \le CD_t|h - p_H|, \tag{11}$$

when the score-function norm times the clipped importance factor is bounded by C. Let $\mu_H = \mathbb{E}[a^{\mathrm{HFC}}\psi \mid \mathcal{F}_{t-1}, D_t = 1]$ and μ_{p_H} be its oracle counterpart. Jensen's inequality gives $\|\mu_H - \mu_{p_H}\| \le C\sqrt{\mathbb{E}[(h - p_H)^2 \mid \mathcal{F}_{t-1}, D_t = 1]}$. Here $J(\theta) = \mathbb{E}_{q\sim\mathcal{D},\, y\sim\pi_\theta(\cdot|q)}[r(q, y)]$ is expected binary verifier reward under a fixed prompt distribution $\mathcal{D}$. Hence, if $\langle \mu_{p_H}, \nabla J(\theta)\rangle \ge \gamma > 0$, then μ_H remains ascent-aligned whenever $C\|\nabla J(\theta)\|\sqrt{\mathbb{E}[(h - p_H)^2 \mid \mathcal{F}_{t-1}, D_t = 1]} < \gamma$. This gives a local error budget for preserving an assumed positive oracle-alignment margin. Appendix D reports matched-event Brier, finite-bank reference-cosine, and one-step held-out reward diagnostics; full proofs are in Appendix A.

6 Experiments

6.1 Experimental setup

We evaluate HFC-GRPO on Qwen3-8B under the RL-ZVP and GRPO-DBB protocols. Acc@8 averages exact-match accuracy across eight responses; Pass@8 measures whether any response is correct. GRPO-DBB separates in-distribution (ID) and out-of-distribution (OOD) benchmarks. Paired methods share initialization, prompt order, verifier, rollout and update budgets, decoding, and evaluation; estimator-specific settings are listed in Appendix J. Checkpoints are selected on validation data disjoint from training and final tests. For Search-R1 transfer with Qwen2.5-3B-Instruct, each response is a complete reasoning–retrieval trajectory (Appendix H). Benchmark sources and full results appear in Appendix E; seeds and selection rules are in Appendix J.

6.2 Main results

Paired comparisons use a common training and evaluation pipeline (Table 1); Tables 2–3 provide source-reported context under the corresponding protocols.

Paired comparisons under a common pipeline. Table 1 reports five-seed paired comparisons under both reasoning protocols. The *source-best envelope* fixes the highest-scoring source-table method for each benchmark–metric cell, then macro-averages its corresponding common-pipeline cells within each seed. Means, sample standard deviations, and paired intervals are computed from these seed-level averages. For RL-ZVP Pass@8, it selects GRPO-DS-g on Minerva and RL-ZVP elsewhere. HFC gains 0.65 Acc@8 and 0.60 Pass@8 over this comparator, with positive intervals under both protocols. Table F.3 directly compares NGRPO, AVSPO, GRPO-DBB, BV-Blend, and RL-ZVP in the same harness: HFC gains 0.65–1.78 Acc@8 points while retaining full signal coverage and exact mixed-group advantages.

Table 1: Five-seed paired common-harness macro-averages (mean±sample standard deviation, %). The comparator is the cellwise source-best method; the last column is the paired HFC-GRPO difference with a 95% t interval.

Protocol	Metric	GRPO	Source-best envelope	HFC-GRPO	Paired Δ
RL-ZVP	Acc@8	49.94±0.31	55.09±0.15	**55.74±0.14**	+0.65 [0.63, 0.66]
RL-ZVP	Pass@8	65.29±0.28	69.37±0.07	**69.97±0.07**	+0.60 [0.57, 0.63]
GRPO-DBB	ID Acc@8	48.99±0.27	51.31±0.43	**51.83±0.42**	+0.52 [0.50, 0.54]
GRPO-DBB	OOD Acc@8	51.54±0.28	59.88±0.58	**60.63±0.57**	+0.75 [0.73, 0.77]

Under the RL-ZVP protocol (Table 2), HFC leads each benchmark with mean Acc@8/Pass@8 of 55.74/69.97. Its gains over the source-reported RL-ZVP row are 0.65/0.73 points; Table 1 instead uses the paired source-best envelope.

Under the GRPO-DBB protocol (Table 3), HFC reaches 51.83 ID and 60.63 OOD Acc@8, improving the strongest source baseline by 0.52 and 0.75 points. The larger OOD margin accompanies higher

Table 2: Qwen3-8B Acc@8 / Pass@8 (%) under the RL-ZVP protocol. Baselines are from Table 2 of Le et al. (2026); HFC reports five-seed means. (r)/(g) denote rollout-/gradient-step-matched filtering. Best values are bold.

Method	Minerva	AMC23	MATH500	AIME24	AIME25	Olympiad
GRPO	44.99/58.42	73.75/92.48	83.00/92.40	28.33/50.64	20.00/31.59	49.59/66.20
GRPO-DS-r	29.96/50.85	47.19/76.70	68.20/87.26	12.08/31.33	9.17/22.02	43.88/58.16
GRPO-DS-g	44.76/59.60	76.25/92.43	84.98/92.42	27.08/42.72	19.58/35.93	44.83/60.42
GRESO-r	30.74/50.53	49.38/80.53	67.40/87.79	12.08/29.31	10.83/26.26	44.92/59.66
GRESO-g	45.50/58.69	78.13/92.90	83.23/92.24	25.00/41.80	17.92/32.12	44.14/59.79
RL-ZVP	45.96/58.83	78.75/94.18	89.73/96.13	33.33/54.66	24.58/39.36	58.20/72.28
HFC-GRPO	**46.31/59.95**	**79.44/94.56**	**90.48/96.80**	**34.02/55.37**	**25.22/40.11**	**58.95/73.03**

homogeneous-group exposure (26.2% versus 18.7%) and recovered L_1 mass (20.8% versus 16.0%). Appendix D relates these gains to the branch HFC repairs through matched diagnostics.

Table 3: Qwen3-8B Acc@8 averages under the GRPO-DBB protocol. Baseline aggregates are reported in Table 1 of Kim et al. (2026); HFC reports five-seed means under the shared protocol in Appendix J.

Method	ID average	OOD average
GRPO	48.99	51.54
RePO	46.82	47.35
GRESO	46.86	58.83
DAPO	46.58	55.91
GRPO-DBB	51.31	59.88
HFC-GRPO	**51.83**	**60.63**

Transfer to multi-turn search agents. Table 4 replaces the homogeneous-group advantage in Search-R1: final-answer binary EM rewards each complete trajectory, and its scalar advantage applies to generated tokens with retrieved text masked. Against paired local outcome-only GRPO, HFC gains 2.12 points in seven-task macro-average EM (95% paired interval [1.85, 2.40]); every task-level interval is positive. Its means exceed source-reported TRM on every dataset, by 1.02 points on average, without a process-reward model. Appendix H gives the configuration, exposure diagnostic, and matched history-alignment and timing controls.

Table 4: Search-R1 agent transfer on Qwen2.5-3B-Instruct (exact match, EM, %). Source central estimates are from Table 1 of Ma et al. (2026); paired columns report three-seed mean±sample standard deviation.

	Source-reported		Paired	
Dataset	GRPO	+TRM	GRPO	+HFC-GRPO
NQ	47.01	47.89	47.11±0.25	**48.64±0.23**
TriviaQA	61.88	62.57	61.90±0.16	**63.22±0.21**
PopQA	45.73	47.20	45.75±0.18	**47.58±0.24**
HotpotQA	43.34	44.47	43.41±0.20	**45.64±0.30**
2Wiki	42.68	43.48	42.76±0.20	**45.28±0.34**
MuSiQue	18.08	19.65	18.12±0.14	**20.94±0.38**
Bamboogle	37.60	39.20	37.68±0.23	**40.29±0.36**
Average	42.33	43.49	42.39±0.19	**44.51±0.29**

Transfer across model families. On DeepMath, HFC improves GRPO accuracy by 0.80, 2.54, and 1.74 points for Llama-3.1-8B, Qwen2.5-7B, and Mistral-7B-Instruct-v0.3, respectively (Grattafiori et al., 2024; Yang et al., 2025b; Jiang et al., 2023). These five-seed comparisons (Appendix G) extend the same group-level correction across model families.

6.3 Mechanism and replacement ablations

Table 5 separates signal restoration, forecast choice, scope, and history alignment. The fixed $h = 0.5$ gate restores homogeneous-group advantages; HFC adds 0.43 Acc@8 points and gains 0.45 [0.37, 0.52] over strict-lag EMA. Removing individual historical statistics costs 0.36–0.66 points and raises homogeneous-gated Brier from .143 to .163–.169 (Appendix D). Read-permuted histories cost 1.86 points while retaining signal coverage, emphasizing prompt-history alignment. Scale-preserving correction probes mixed-group intervention with the GRPO contrast intact; other replacements test timing and aggregation. At two visits per prompt, shrinkage and bridge share a prediction; removing either probes the mixture weight assigned to that shared prediction, while Figure 2 follows predictor-weight adaptation over five epochs. Appendix D gives the definitions and paired intervals; Table H.3 gives a single-update agent diagnostic.

Table 5: Controlled replacements on Qwen3-8B, averaged over six RL-ZVP benchmarks (five-seed mean±sample standard deviation, %).

Variant	Acc@8	Pass@8
GRPO	49.94±0.31	65.29±0.28
Fixed forecast gate ($h = 0.5$)	55.31±0.16	69.43±0.09
Scale-preserving historical correction	55.48±0.15	69.61±0.08
Current-inclusive update	55.12±0.12	69.22±0.12
Strict-lag EMA	55.29±0.19	69.39±0.12
Uniform predictor weights	55.23±0.18	69.35±0.08
Without per-prompt statistic	55.08±0.17	69.18±0.14
Without recent predictor	55.20±0.15	69.31±0.07
Joint mixture-loss EG	55.16±0.18	69.27±0.11
Historical reference on every group	54.70±0.15	68.79±0.14
Read-permuted histories	53.88±0.15	67.92±0.14
HFC-GRPO	**55.74±0.14**	**69.97±0.07**

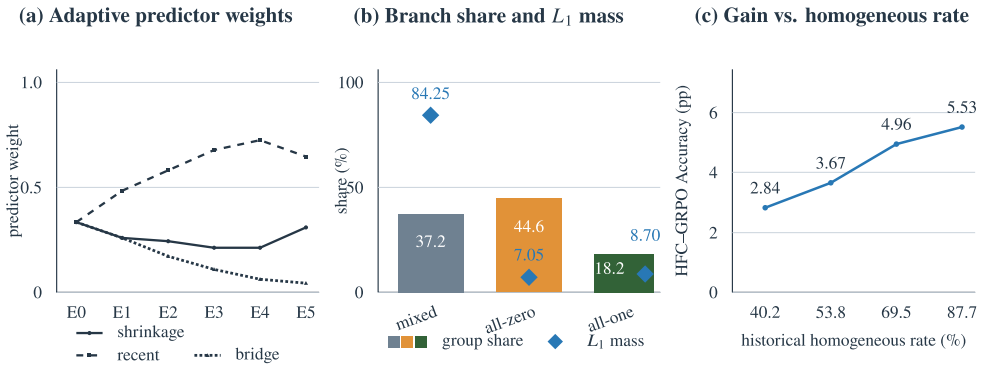

Figure 2: Qwen2.5-7B DeepMath mechanism diagnostics from one five-epoch controller trace ($G = 4$). (a) Normalized predictor weights from the calibration-only state E0 through E5. (b) Group share (bars) and sequence-advantage L_1 mass share (diamonds) over 500 training groups. (c) Final-epoch HFC-minus-GRPO Accuracy across four equal-count prompt strata defined by pre-outcome historical homogeneous-group rate.

Predictor adaptation and signal coverage. Figure 2 tracks five DeepMath epochs: recent peaks at 0.725 after epoch 4, while shrinkage rebounds from 0.212 to 0.310 at epoch 5. Homogeneous groups account for 62.8% of 500 events but 15.75% of sequence-advantage L_1 mass; mixed groups supply 84.25%.

Gain across prompt strata. At the final checkpoint, the HFC–GRPO Accuracy difference rises from 2.84 to 5.53 points across equal-count prompt strata with increasing pre-outcome homogeneous-group prevalence. These strata index opportunities for selective recovery rather than per-group update size. Exact weights, counts, and the diagnostic protocol appear in Appendix J.

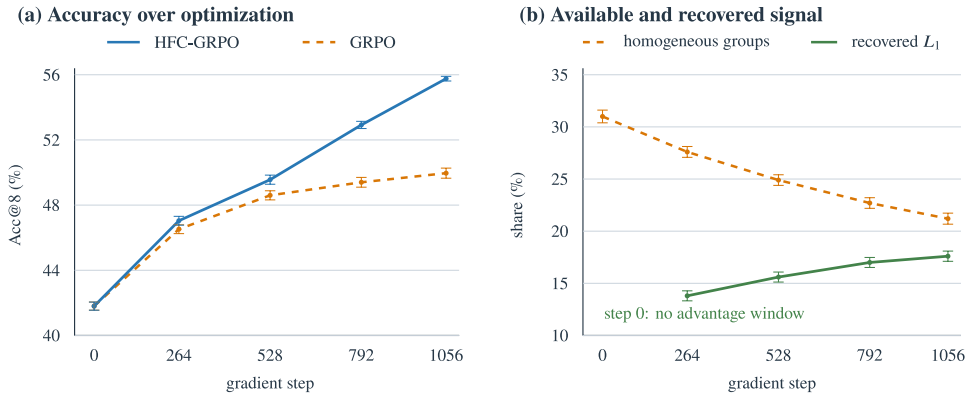

Figure 3: Five-seed dynamics under the RL-ZVP protocol (mean±sample standard deviation). (a) HFC-GRPO and GRPO Acc@8. (b) On HFC rollouts, homogeneous-group and recovered-advantage L_1 shares; recovered mass is undefined before the first rollout (step 0).

Figure 3 shows homogeneous groups declining from 31.0% to 21.2% over steps 0–1,056. Between steps 264 and 1,056, recovered mass rises from 13.8% to 17.6% and the HFC–GRPO Acc@8 gap from 0.52 to 5.80 points; both methods start at 41.80 Acc@8.

Computational cost. HFC adds 0.62% step time over GRPO under RL-ZVP (Table F.3) and 0.02 GB peak memory (Appendix J). Prompt-state storage scales with retained identifiers: the experiments use a fixed finite set, while open streams can cap state through eviction with global fallback (Appendix B).

7 Conclusion

HFC-GRPO introduces a selective advantage estimator: group type determines intervention and sign, while a strict-lag historical forecast sets correction magnitudes only for homogeneous groups, preserving mixed-group GRPO exactly. The predict-before, gate-after structure explains why calibration on the full rollout stream need not carry over to the homogeneous-group population. Our analysis separates forecast error, clipping, and selection effects, bounding advantage and gradient-contribution deviations under explicit assumptions. Paired comparisons and targeted replacements across reasoning and search-agent settings support the selective use of aligned historical feedback to recover signals where within-group comparison vanishes.

AI use statement

The authors led the research, including its conception, methodological design, and scientific interpretation. Generative AI tools assisted with literature research, language polishing, figure preparation, mathematical derivations, theoretical proofs, and preliminary manuscript feedback. The authors reviewed and verified the mathematical arguments, evaluated all AI-assisted suggestions, and made the final scientific and editorial decisions. The authors take responsibility for the final manuscript.

Reproducibility statement

Appendices A–C give proofs, the complete update and transaction protocol (Algorithm B.1), and notation. Appendices D–H provide replacement controls, per-benchmark and closest-method results, cross-model tests, and agent-transfer details. Appendix J specifies shared settings, method-specific parameters, seeds, compute budgets, diagnostics, and replay checks.

References

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. doi: 10.1017/CBO9780511546921.

Yupeng Chang, Yuan Wu, and Yi Chang. BV-blend: Uncertainty-weighted historical baselines for stable critic-free RL with verifiable rewards. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1706–1725. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.84. URL https://aclanthology.org/2026.findings-acl.84/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang

Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu,

Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://doi.org/10.1038/s41586-025-09422-z.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3828–3850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/.

Xixiang He, Qiyao Sun, Ao Cheng, Xingming Li, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Advantage collapse in group relative policy optimization: Diagnosis and mitigation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=MKNimf9bIx.

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 *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf.

Jian Hu, Jason Klein Liu, Haotian Xu, and Wei Shen. REINFORCE++: Stabilizing critic-free policy optimization with global advantage normalization. *arXiv preprint arXiv:2501.03262*, 2025. URL https://arxiv.org/abs/2501.03262.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. URL https://arxiv.org/abs/2310.06825.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1: Training LLMs to reason and leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Rwhi91ideu.

Haechan Kim, Soohyun Ryu, Gyouk Chu, Doohyuk Jang, and Eunho Yang. Discounted beta–bernoulli reward estimation for sample-efficient reinforcement learning with verifiable rewards. *arXiv preprint arXiv:2603.18444*, 2026. URL https://arxiv.org/abs/2603.18444.

Thanh-Long V. Le, Myeongho Jeon, Kim Vu, Viet Dac Lai, and Eunho Yang. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kiXFIESZKv.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 3843–3857, 2022. doi: 10.52202/068431-0278. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/18abbeef8cfe9203fdf9053c9c4fe191-Abstract.html.

Chenxi Liu, Junjie Liang, Yuqi Jia, Bochuan Cao, Yang Bai, Heng Huang, and Xun Chen. Explore data left behind in reinforcement learning for reasoning language models. *arXiv preprint arXiv:2511.04800*, 2025a. doi: 10.48550/arXiv.2511.04800. URL https://arxiv.org/abs/2511.04800.

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. In *Second Conference on Language Modeling*, 2025b. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Da Ma, Ziyue Yang, Hongshen Xu, Haotian Fang, Kai Yu, and Lu Chen. Empowering LLM tool invocation with tool-call reward model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=LnBEASInVr.

Gongrui Nan, Siye Chen, Jing Huang, Mengyu Lu, Dexun Wang, Chunmei Xie, Weiqi Xiong, Xianzhou Zeng, Qixuan Zhou, Yadong Li, and Xingzhong Xu. NGRPO: Negative-enhanced group relative policy optimization. *arXiv preprint arXiv:2509.18851*, 2025. doi: 10.48550/arXiv.2509.18851. URL https://arxiv.org/abs/2509.18851.

Jatin Prakash and Anirudh Buvanesh. What can you do when you have zero rewards during RL? *arXiv preprint arXiv:2510.03971*, 2025. URL https://arxiv.org/abs/2510.03971.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level google-proof Q&A benchmark. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

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

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

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Hu Wang, Congbo Ma, Ian Reid, and Mohammad Yaqub. Kalman filter enhanced GRPO for reinforcement learning-based language model reasoning. *arXiv preprint arXiv:2505.07527*, 2026. doi: 10.48550/arXiv.2505.07527. URL https://arxiv.org/abs/2505.07527.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 95266–95290, 2024. doi: 10.52202/079017-3018. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ad236edc564f3e3156e1b2feafb99a24-Abstract.html.

Chenxing Wei, Jiarui Yu, Ying He, Hande Dong, Yao Shu, and Fei Richard Yu. ReDit: Reward dithering for improved LLM policy optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 61488–61521, 2025. doi: 10.52202/085713-2055. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/58caefe277fc1d48173bc59f56c9b86a-Abstract-Conference.html.

Zhongwen Xu and Zihan Ding. Single-stream policy optimization. In *International Conference on Learning Representations*, volume 2026, pp. 140925–140944, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/e46fc33e80e9fa2febcdb058fba4beca-Paper-Conference.pdf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a. URL https://arxiv.org/abs/2505.09388.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025b. URL https://arxiv.org/abs/2412.15115.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru 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, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025. doi: 10.52202/085713-3775. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Guanning Zeng, Zhaoyi Zhou, Daman Arora, and Andrea Zanette. Shrinking the variance: Shrinkage baselines for reinforcement learning with verifiable rewards. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=LffORETnav.

Haizhong Zheng, Yang Zhou, Brian R. Bartoldson, Bhavya Kailkhura, Fan Lai, Jiawei Zhao, and Beidi Chen. Act only when it pays: Efficient reinforcement learning for LLM reasoning via selective rollouts. In *Advances in Neural Information Processing Systems*, volume 38, pp. 124321–124346, 2025. doi: 10.52202/085713-4143. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b42fb82fc3ffb88b872c4714f093875d-Paper-Conference.pdf.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 126546–126573. Curran Associates, Inc., 2025. doi: 10.52202/085713-4220. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/b8026fe01f5f8f7a80a1a9bf01d08cad-Paper-Conference.pdf.

A Proof Details

Throughout the appendix, Group Relative Policy Optimization (GRPO) and Historical Feedback Control for GRPO (HFC-GRPO) retain the definitions and branch conventions introduced in the main text.

A.1 Independent losses and predictor disagreement

For weights v_j summing to one, unclipped prediction $\widetilde{h} = \sum_j v_j e_j$, and target z,

$$\sum_j v_j (e_j - z)^2 = (\widetilde{h} - z)^2 + \sum_j v_j (e_j - \widetilde{h})^2. \tag{12}$$

This identity makes the distinction between independent-predictor and mixture-loss updating explicit: the former retains a penalty for predictor disagreement that the latter discards.

Let $\widetilde{h} = \sum_j v_j e_j$. Then

$$\sum_j v_j (e_j - z)^2 = \sum_j v_j \{(e_j - \widetilde{h}) + (\widetilde{h} - z)\}^2 \tag{13}$$

$$= \sum_j v_j (e_j - \widetilde{h})^2 + (\widetilde{h} - z)^2 + 2(\widetilde{h} - z) \sum_j v_j (e_j - \widetilde{h}). \tag{14}$$

The final sum is zero because $\sum_j v_j e_j = \widetilde{h}$ and $\sum_j v_j = 1$, proving Eq. 12.

A.2 Advantage scale on binary groups

For a mixed group with $p = K/G$, the sample variance is $s^2 = \frac{G}{G-1} p(1-p)$. Correct responses have numerator $1 - p$ and incorrect responses have numerator $-p$, so

$$\frac{1}{G} \sum_i |a_i^{\mathrm{GRPO}}| = \frac{p(1-p) + (1-p)p}{s + \delta_{\mathrm{std}}} \tag{15}$$

$$= \frac{2p(1-p)}{\sqrt{\frac{G}{G-1} p(1-p)} + \delta_{\mathrm{std}}} \leq 2\sqrt{\frac{G-1}{G} p(1-p)} \leq \sqrt{\frac{G-1}{G}}. \tag{16}$$

The last inequality uses $p(1-p) \leq 1/4$. A homogeneous HFC group has one constant magnitude, h or $1 - h$, and clipping places it in $[\varepsilon_h, 1 - \varepsilon_h]$. This comparison is group-level: rare-class mixed responses can have larger individual standardized values, but their frequency keeps the group-average L_1 mass bounded. The common PPO clip and KL settings then control the surrounding update identically; empirically, HFC recovers 17.6% of total advantage mass while KL remains .006 to reported precision.

A.3 Selection-aware Brier excess and gated residuals

For $Z \in [0,1]$, $p_0 = \mathbb{E}[Z \mid \mathcal{F}_{t-1}]$, and an unclipped forecast $\widetilde{h}$,

$$\mathbb{E}[(\widetilde{h} - Z)^2 \mid \mathcal{F}_{t-1}] = \mathbb{E}[(p_0 - Z + \widetilde{h} - p_0)^2 \mid \mathcal{F}_{t-1}] \tag{17}$$

$$= \mathbb{E}[(p_0 - Z)^2 \mid \mathcal{F}_{t-1}] + (\widetilde{h} - p_0)^2, \tag{18}$$

because the conditional cross term is zero. Repeating the decomposition under the post-outcome diagnostic event $D_t = 1$ yields $p_H = \mathbb{E}[Z \mid \mathcal{F}_{t-1}, D_t = 1]$ and excess $(\widetilde{h} - p_H)^2$. On an all-zero branch, oracle-reference and forecast-reference residuals are $-p_H$ and $-h$; on an all-one branch they are $1 - p_H$ and $1 - h$; on a mixed branch both estimators are GRPO. The pointwise substitution error is therefore exactly $D_t|h - p_H|$.

All-rollout training and gated use are connected by Eq. 10. Squaring and applying $(a + b + c)^2 \le 3(a^2 + b^2 + c^2)$ gives

$$(h - p_H)^2 \le 3(h - \widetilde{h})^2 + 3(\widetilde{h} - p_0)^2 + 3(p_0 - p_H)^2. \tag{19}$$

All-rollout Brier updates target the middle term. Clipping displacement is directly observable; the conditional selection gap is estimated from an independent reference bank. On the held-out stream, the mean absolute clipping displacement $\mathbb{E}|h - \widetilde{h}|$ is $.006 \pm .001$; the reference-bank estimate of $|p_0 - p_H|$ is $.043 \pm .002$ across five seeds.

If a per-response gradient contribution has the form $a\,\psi$ with $\|\psi\| \le C$, substituting h for p_H changes it by at most $CD_t|h - p_H|$. With μ_H and μ_{p_H} defined in Proposition 2, Jensen's inequality gives

$$\|\mu_H - \mu_{p_H}\| \le C\sqrt{\mathbb{E}[(h - p_H)^2 \mid \mathcal{F}_{t-1}, D_t = 1]}. \tag{20}$$

For $d = \nabla J(\theta)$, Cauchy–Schwarz gives $\langle \mu_H, d\rangle \ge \langle \mu_{p_H}, d\rangle - \|\mu_H - \mu_{p_H}\|\|d\|$, proving the stated sufficient condition. Clipping is already included through h; replacing it by $\widetilde{h}$ adds at most $C|h - \widetilde{h}|$. The proposition bounds substitution error and preserves an assumed positive oracle-alignment margin when that error is sufficiently small.

For completeness, write the batch-mean predictor loss as $\bar{\ell}_{j,b} = |\mathcal{B}_b|^{-1} \sum_q (e_{j,bq} - z_{bq})^2 \in [0,1]$. Convexity gives the mixture loss $\bar{\ell}_{\mathrm{mix},b} \le \sum_j v_{j,b}\bar{\ell}_{j,b}$. The standard exponential-weights bound therefore yields, for B committed batches,

$$\sum_{b=1}^{B} \bar{\ell}_{\mathrm{mix},b} - \min_j \sum_{b=1}^{B} \bar{\ell}_{j,b} \le \frac{\log 3}{\eta} + \frac{\eta B}{8}. \tag{21}$$

The bound is invariant to prompt-batch size because both the update and loss are batch means and controls static regret against the best fixed expert. Under nonstationarity, the same strict-lag gate admits discounted losses or fixed-share exponential weights; our cumulative, discounted-count, finite-window, and EMA controls retain positive gains, with the cumulative controller strongest in the reported horizon (Appendix D).

B Algorithm

B.1 Core training loop

For each rollout batch, the statistical update has four steps:

1. Read one committed controller snapshot and compute the historical statistics, three predictors, and frozen forecasts for all prompts using Eqs. 3–6.
2. Generate responses from the batch's frozen parent policy, verify binary rewards, and select each group's advantage using Eq. 2.
3. Optimize the policy with the same forecasts and advantages throughout all PPO reuse steps.
4. Commit the batch once: update success/trial counts, last-visit records, and each predictor's batch-mean Brier loss in Eq. 7.

The execution protocol below implements these steps with bounded live payloads and deterministic recovery. Its waves share one parent policy and controller snapshot; completing a wave reclaims its payload but does not commit controller state.

B.2 Bounded-memory execution and replay

The admission cap A bounds live prompt-group payload slots, not the logical batch size. Larger batches use deterministic waves of at most A groups. With a retry window of W commits, the controller uses $O(|Q| + W + A)$ scalar metadata and $O(AGT)$ rollout payload for group size G and maximum trajectory length T, with $O(1)$ controller work per group. Algorithm B.1 specifies reservation, isolated policy optimization, and atomic publication.

Algorithm B.1: Batch-atomic strict-lag HFC-GRPO update

Input: prompt batch $\mathcal{B}_b$, batch identifier u_b, group size G, request digest ρ_b^{req}, immutable environment manifest, admission cap A, retry window W, latest committed state

if *controller state is absent* **then**
  initialize counts, prompt map, $L_{1:3} = 0$, bounded pending queue $\mathcal{P}$, reservation map $\mathcal{R}$, digest map $\mathcal{D}$, monotone lease counter Λ, and $u = -1$;
if $u_b \le u - W$ **then**
  return `expired/unknown` without changing state;
if $u_b \le u$ **then**
  require $\rho_b^{\text{req}} = \mathcal{D}[u_b].\rho^{\text{req}}$; on mismatch raise a hard error; otherwise run idempotent finalization for u_b (remove any terminal reservation and reclaim orphaned rollout buffers) and **return** stored commit digest $\mathcal{D}[u_b].\kappa$;
if $u_b > u + 1$ **then**
  admit one scalar batch descriptor to $\mathcal{P}$ only if $|\mathcal{P}| < A$; otherwise return backpressure; after predecessors commit, reload the latest snapshot;
Atomically reserve $(u_b, u, \rho_b^{\text{req}})$ by compare-and-swap and assign fencing token $\lambda_b = \Lambda + 1$; a matching live reservation returns `in-flight`, while lease takeover increments λ_b, preserves any durable wave digest/buffer, and a mismatched request raises a hard error;
Freeze one immutable snapshot $\mathcal{F}_{b-1}$ for every $q \in \mathcal{B}_b$;
Partition $\mathcal{B}_b$ in stable prompt order into waves $\mathcal{W}_{b,r}$ with $|\mathcal{W}_{b,r}| \le A$; for the 512- and 128-prompt protocols, $A = 64$ gives eight and two waves;
foreach *wave* $\mathcal{W}_{b,r}$ **do**
  from the immutable parent and $\mathcal{F}_{b-1}$, compute Eq. 3, $e_{1:3,bq}$, batch-frozen $v_{1:3,b}$, and h_{bq}; generate/verify G responses in parallel; fenced CAS durably appends the ordered wave digest to τ_b and marks its observation buffer rollout-ready; takeover reuses that buffer or requires a byte-identical digest; reuse its frozen advantages for every PPO epoch on the isolated child, checkpoint the child and scalar sufficient statistics, then reclaim the wave buffer before admitting the next wave;
Lock; recheck that λ_b is the current fence, $u_b = u + 1$, and that the frozen parent version equals u;
In stable prompt order, add K_{bq} and G to global/prompt counts and set each last-group pair; update every L_j by the batch mean in Eq. 7;
Store $\mathcal{D}[u_b] = (\rho_b^{\text{req}}, \tau_b, \kappa_b)$; set $u \leftarrow u_b$; one atomic transaction publishes controller, policy/optimizer, random number generator (RNG) state, data-loader cursor, and digest record, removes the terminal reservation and all logical buffer references, and retains only digest keys $u - W + 1{:}u$; physical blob deletion is an idempotent finalizer replayed after restart or committed retry;

This reservation–isolation–publish transaction is deterministic under replay. It assumes one logical linearizable controller coordinator: rollout workers remain parallel within each wave, while controller/policy commits are serialized by batch identifier and fencing token. The pre-execution digest ρ_b^{req} hashes (u_b, u), stable prompt keys, policy version, sampler configuration, RNG seed, corpus snapshot, index identifier, and retriever/tool implementation versions. For an agent response, τ_b is a chained digest of ordered wave transcripts: generated tokens, tool-call payloads, tool-response content hashes, turn boundaries, loss masks, and verifier outputs. Thus equal final answers cannot hide different retrieval contexts. Each rollout-ready wave durably stores its digest and observation buffer before optimizer mutation; takeover reuses that buffer rather than querying the tool again. The commit digest κ_b hashes the published controller and policy/optimizer/RNG/data-loader/environment manifest. Every reservation transition is fenced: a takeover carries forward any durable wave digest,

and a smaller token can neither overwrite that record nor publish. PPO mutates only an isolated child, and publication jointly compares the parent version and current fence. Mutable scalar metadata is $(S_g, N_g, L_{1:3}, u, \Lambda, \mathcal{P}, \mathcal{R}, \mathcal{D})$ plus the per-prompt map $q \mapsto (S_q, N_q, K_q^{\text{last}}, G_q^{\text{last}})$. At most A live group payloads and A queued scalar batch descriptors, atomic logical reclamation, and the last W digest records give $O(|Q| + W + A)$ scalar metadata and $O(AGT)$ rollout payload independently of $|\mathcal{B}_b|$. For a finite training set, $|Q|$ is fixed. For an open-ended stream, time-to-live or least-recently-used eviction occurs before admitting a new prompt; a revisited evicted identifier follows the unseen-prompt global fallback in Eq. 3. $(\alpha, \eta, \varepsilon_h, A, W)$ are immutable within a run.

C Notation

Table C.1: Notation used in the method and analysis. Subscript t indexes an analytical prompt group; b indexes an implementation rollout batch, and i a response within a group.

Symbol	Meaning
$G, K_t, r_{t,i}, \bar{r}_t, s_t$	group size, number correct, binary reward, group mean, and group reward standard deviation
$\mathcal{F}_{t-1}$	immutable controller state available before event t
S_g, N_g, S_q, N_q	global and prompt-specific success/count statistics
x_g, x_q, x_r	global, question-specific, and most-recent strictly lagged statistics
$e_{j,bq}, L_{j,b}, \ell_{j,b}, v_{j,b}$	prompt-specific predictor forecast, cumulative batch-mean loss, centered loss, and batch-frozen predictor weight
$\widetilde{h}_t, h_t$	raw mixture forecast and its interior-clipped value
$z_t = K_t/G, D_t$	observed group success fraction and homogeneous-group indicator
$a^{\text{GRPO}}, a^{\text{HFC}}, a^{p_H}$	standard GRPO, HFC, and homogeneous-conditional oracle advantages
u_b, u	current batch identifier and largest committed batch identifier
$(\mathcal{B}_b, \mathcal{F}_{b-1})$	optimizer rollout batch and its single frozen parent state
h_{bq}, K_{bq}, z_{bq}	clipped forecast, correct count, and success fraction for prompt q in batch b
$(\rho_b^{\text{req}}, \tau_b, \kappa_b)$	pre-execution request, post-rollout, and atomic-commit digests
$(\mathcal{P}, \mathcal{R}, \mathcal{D}, A, W)$	bounded pending queue, reservation map, retained digests, admission cap, and retry window
$T, (\Lambda, \lambda_b)$	maximum trajectory length, monotone lease counter, and batch fencing token
$(K_q^{\text{last}}, G_q^{\text{last}})$	correct and total counts at the last committed visit to prompt q
$\delta_{\text{std}}, \varepsilon_h$	GRPO denominator stabilizer and HFC forecast clipping level
$\alpha, \eta, \eta_m, c_t$	symmetric Beta prior, independent/joint learning rates, and shrinkage coefficient
g, ψ, C	per-response gradient contribution, its score-function factor, and the factor-norm bound
$p_0, p_H, \widehat{p}_0, \widehat{p}_H$	all-rollout and homogeneous-conditional oracles and their independent reference-bank estimates
$(\mu_H, \mu_{p_H}, \gamma, d)$	conditional HFC/oracle updates, alignment margin, and $d = \nabla J(\theta)$

D Mechanism Ablations and Sensitivity

Acc@8 denotes mean exact-match accuracy across eight responses, and Pass@8 denotes the fraction of prompts with at least one correct response among eight.

Visit depth and predictor identity. The local reasoning schedules traverse fixed prompt-ID sets without replacement each epoch, with no prompt-specific warm-start: at the final update, each retained ID has two visits in RL-ZVP and four in DBB. PPO reuse adds no visits. On a first visit all predictors equal x_g; on a second, $N_q = G$, $x_q = x_r$, and $e_1 = e_3$. Thus $L_1 = L_3$ throughout the two-epoch protocol, and its forecast is exactly a two-predictor mixture with initial masses $(2/3, 1/3)$ for the

shared shrinkage–bridge prediction and recent prediction. Deleting either duplicate gives the same estimator and changes that prior to $(1/2, 1/2)$; it does not isolate a distinct timescale. From the third visit, accumulated and last-visit statistics can differ. The four-epoch protocol permits this separation, while the two-epoch ablations test aggregation and history use at low visit depth.

Controlled leave-one-out construction. Statistic ablations remove the named statistic everywhere it occurs while retaining three forecast slots and the same strict-lag gate. Without the global statistic, unseen prompt and last-visit statistics use the fixed prior mean $1/2$, and $(e'_1, e'_2, e'_3) = (x_q, x_r, (x_q + x_r)/2)$. Without the per-prompt statistic, unseen recent state falls back to x_g, and $(e'_1, e'_2, e'_3) = (x_g, x_r, (x_g + x_r)/2)$. Without the last-visit statistic, unseen question state follows the structured shrinkage fallback, and $(e'_1, e'_2, e'_3) = (e_1, x_g, (x_g + e_1)/2)$. Predictor ablations delete one complete forecast and renormalize the two surviving exponential weights. Statistic tests remove an information source; predictor removal changes the predictor mixture and its prior mass.

The scope control applies the same frozen historical forecast to every response, $a_i^{\text{all}} = r_i - h_t$, including mixed groups; it preserves forecast timing and state. On mixed groups, $r_i - h_t = (s_t + \delta_{\text{std}})a_i^{\text{GRPO}} + (Z_t - h_t)$, so the comparison changes both contrast scaling and the historical offset. It supports the complete selective estimator rather than isolating routing alone. A scale-preserving routing control would use $a_i^{\text{GRPO}} + (Z_t - h_t)$ on mixed groups and the original HFC rule on homogeneous groups. A complementary fixed-forecast control would retain the original gate with the prespecified $h_t = 1/2$ and no tuning. We evaluate these two controlled variants to isolate adaptive magnitude and selective intervention. The fixed-forecast gate replaces the learned forecast with $h_t = 1/2$ while preserving homogeneous-only intervention. The scale-preserving variant adds the historical residual to mixed groups while retaining their normalized GRPO contrast. Both variants use the same frozen history interface, rollout budget, and optimization protocol.

Table D.1: Mechanism-isolating variants on the RL-ZVP protocol (five-seed mean±sample standard deviation, %). Fixed forecast removes adaptive magnitude; scale-preserving correction retains mixed-group GRPO contrast while adding the historical residual.

Variant	Acc@8	Pass@8
GRPO	49.94±0.31	65.29±0.28
Fixed forecast gate	55.31±0.16	69.43±0.09
Scale-preserving correction	55.48±0.15	69.61±0.08
HFC-GRPO	**55.74±0.14**	**69.97±0.07**

The fixed-magnitude gate in Table H.3 instead assesses a single agent update from a shared checkpoint. The alignment control constructs a seed-deterministic derangement π_b inside each frozen batch, stratified by prior visit count and recent-record availability. Prompt q reads its predictors from the immutable donor state $\mathcal{F}_{b-1}[\pi_b(q)]$, with $\pi_b(q) \neq q$, but its verified outcome is always committed to its own state $\mathcal{F}[q]$. The control therefore preserves history availability, marginal state distributions, predictor-weight updates, and freeze–commit timing while breaking prompt–history semantics. As an implementation sanity check, applying one bijection to both reads and writes produces exactly the original forecasts and advantages up to key names; we verify zero maximum difference for this simultaneous-renaming control.

Timing replacements. All timing controls operate on complete rollout batches. Current-inclusive freezes $\mathcal{F}_{b-1}$ for generation, observes every z_{bq}, and recomputes forecasts from a temporary batch-inclusive state before PPO reuse. Pre-reuse publishes controller version b while the corresponding policy child is still being updated. With the same first-in, first-out admission and 64-group concurrency cap, arrivals before policy publication read $(\mathcal{F}_b, \theta_{b-1})$, whereas later arrivals read $(\mathcal{F}_b, \theta_b)$. The step-528 diagnostic evaluates 100 uncommitted events per seed and reports version-lag fractions .74/.72/.75/.71/.73; these are diagnostic fractions, not full-training exposure estimates. This control tests policy–controller version mismatch during overlapping execution, rather than the physical write time in an otherwise synchronous implementation. Table D.2 separates training clocks. Strict-lag exponential moving average (EMA) retains freeze–reuse–commit, with its prompt-level half-life selected from $\{16, 32, 64\}$ validation events (32 selected).

Table D.2: Training counts per seed for the paired reasoning runs. One controller commit follows each completed rollout batch; a local PPO pass traverses its mini-batches once.

Count	RL-ZVP	GRPO-DBB
Rollout batches / HFC commits	66/66	540/540
Optimizer updates	1,056	1,080
Prompt-group events	33,792	69,120
Generated responses	540,672	552,960

Joint-loss comparator. The joint-loss comparator optimizes the mixture prediction directly through one simplex weight vector per batch. Every prompt uses frozen $v_{j,b}$ and $\bar{e}_{bq} = \sum_j v_{j,b} e_{j,bq}$; after all PPO reuse it performs one mean-loss exponentiated-gradient (EG) update,

$$v_{j,b+1} \propto v_{j,b} \exp\left[-\eta_m \frac{1}{|\mathcal{B}_b|} \sum_{q \in \mathcal{B}_b} 2(\bar{e}_{bq} - z_{bq}) e_{j,bq}\right], \quad \sum_j v_{j,b+1} = 1, \quad v_{j,0} = \tfrac{1}{3}. \tag{22}$$

The mean-loss learning rate is selected from $\eta_m \in \{0.25, 0.5, 1.0\}$ on validation data and equals 0.5. It shares the same predictors, clipping, post-reuse commit, and homogeneous-only gate as HFC, but minimizes only mixture Brier loss. Independent HFC instead scores each predictor separately as in Eq. 7, preserving the disagreement term in Eq. 12.

Table D.3: Full mechanism ablation on the RL-ZVP protocol (five-seed mean±sample standard deviation, %). The last column is paired HFC-minus-variant Acc@8 with a 95% t interval.

Family	Variant	Acc@8	Pass@8	HFC−variant
Timing	Current-inclusive update	55.12±0.12	69.22±0.12	.62 [.51,.72]
	Pre-reuse commit	55.18±0.19	69.29±0.03	.56 [.47,.64]
	Strict-lag EMA	55.29±0.19	69.39±0.12	.45 [.37,.52]
Statistics	Without global statistic	55.38±0.16	69.50±0.11	.36 [.30,.42]
	Without per-prompt statistic	55.08±0.17	69.18±0.14	.66 [.55,.76]
	Without last-visit statistic	55.31±0.17	69.43±0.11	.43 [.35,.50]
Predictors	Without recent predictor	55.20±0.15	69.31±0.07	.54 [.45,.62]
Aggregation	Uniform weights	55.23±0.18	69.35±0.08	.51 [.43,.58]
	Joint mixture-loss EG	55.16±0.18	69.27±0.11	.58 [.49,.66]
Scope	Historical reference on every group	54.70±0.15	68.79±0.14	1.04 [.91,1.16]
Alignment	Read-permuted histories	53.88±0.15	67.92±0.14	1.86 [1.70,2.01]
Full	HFC-GRPO	**55.74±0.14**	**69.97±0.07**	–

The homogeneous-gated Brier diagnostic is .143 for full HFC. Removing global, per-prompt, or last-visit statistics gives .163, .169, or .164, respectively; removing the recent predictor gives .167. These aggregate forecast scores complement the paired Acc@8 comparisons in Table D.3.

Sensitivity. For group size, we fix 540,672 individual responses, 66 rollout batches, 1,056 optimizer updates, prompt order, and five seeds. Each rollout contains 8,192 responses and each mini-batch 512 responses; their prompt-group counts scale as $8{,}192/G$ and $512/G$, respectively. Standard HFC commits once per batch for every G; smaller G changes the number of group observations, not commits. The contribution-matched control (HFC-CM) partitions the stable response stream into 33,792 canonical 16-response blocks. For each divisor G, it splits block k into $m_G = 16/G$ groups, sets $h_s = \text{uint64}(\text{SHA-256}(\texttt{UTF-8}(s))[0{:}8])$ using big-endian conversion, and selects cyclic offset $o_{s,k} = (k + h_s) \bmod m_G$. PPO still uses every group; only the post-reuse controller loss/count contribution is subsampled. All HFC rows therefore retain 66 atomic commits and HFC-CM fixes 33,792 contributions. Because 33,792 is divisible by every m_G, each offset occurs equally often; for every seed, selected-versus-full homogeneous rate differs by at most 0.1 points and mean difficulty by at most .001, where prompt difficulty is one minus the frozen-policy exact-match rate over the canonical 16 responses. Table D.4 shows a positive gain for every $G \in \{2, 4, 8, 16\}$ under both views. Broader one-factor grids cover $\alpha \in \{.25, .5, 1, 2, 4\}$, $\varepsilon_h \in \{0, .01, .02, .05, .10\}$, and $\eta \in \{.5, 1, 2\}$;

all retain a positive margin over GRPO, with the selected setting inside a smooth robust region. Cumulative history reaches 55.74 Acc@8 versus 55.46 for discounted counts and 55.33 for a 64-event window; EMA half-lives 16/32/64 obtain 55.20/55.29/55.22. Across the grids, homogeneous Brier ranges from .143 to .161, recovered advantage mass from 16.2% to 25.4%, clipping from 3.9% to 6.1%, and Kullback–Leibler (KL) divergence remains .006 to reported precision.

Table D.4: Group-size sensitivity at a fixed response budget (five-seed mean±sample standard deviation). HFC-CM fixes 33,792 controller contributions; every HFC row uses 66 commits. Hom. and L_1 are homogeneous-group and recovered-advantage shares; CI denotes confidence interval.

G	Group events	Variant	Acc@8	Δ vs GRPO	95% CI	Hom. (%)	L_1 (%)
2	270,336	HFC	**54.92±.35**	+7.12	[7.08,7.16]	41.8	25.4
2	270,336	HFC-CM	54.55±.35	+6.75	[6.70,6.80]	41.8	24.2
4	135,168	HFC	**55.41±.27**	+6.81	[6.75,6.87]	32.4	22.1
4	135,168	HFC-CM	55.18±.26	+6.58	[6.51,6.65]	32.4	21.3
8	67,584	HFC	**55.62±.37**	+6.32	[6.29,6.35]	25.8	19.5
8	67,584	HFC-CM	55.49±.36	+6.19	[6.18,6.20]	25.8	19.0
16	33,792	HFC / HFC-CM	**55.74±.14**	+5.80	[5.44,6.15]	21.2	17.6

Gate-aligned calibration and direction. Table D.5 examines the relation between forecast error and homogeneous-group update direction in Proposition 2, using uncommitted events from training-seen prompt identifiers. On matched events, HFC versus read-permuted history gives clipped all-rollout Brier .152 versus .171 and homogeneous-gated Brier .143 versus .176. Before display rounding, HFC's unclipped and clipped all-rollout Brier are .15144 and .15162; their same-event signed difference is .00018 (maximum seed value .00023). Since $\widetilde{h}, Z \in [0,1]$ and $h = \mathrm{clip}(\widetilde{h}, .02, .98)$, each event satisfies $(h-Z)^2 - (\widetilde{h}-Z)^2 \le \varepsilon_h^2 = .00040$. The distinct displacement in Eq. 10 is $\mathbb{E}|h-\widetilde{h}| = .006 \pm .001$. Reliability slope/intercept are .97/.01. At step 528, each frozen policy uses a disjoint $M = 4{,}096$, $G = 16$ reference bank per prompt to compute the Beta-smoothed estimates $\widehat{p}_0 = (1+\sum_m Z_m)/(2+M)$ and $\widehat{p}_H = (1+\sum_m D_m Z_m)/(2+\sum_m D_m)$ before the 32 gradient groups are revealed; the reference estimate is independent of those later outcomes, with prior mean $1/2$ at zero homogeneous count. The gap $|\widehat{p}_0 - \widehat{p}_H|$ is .043 ± .002. On the same gradient groups, cosine to the estimated reference update $g_{\widehat{p}_H}$ is .31 versus .07 for aligned versus read-permuted history; one matched 10^{-6} update evaluated on 1,000 disjoint prompts with eight common-random-number responses yields +.18 versus +.03 points. In a deterministic controller replay from empty history, one prompt supplies 50,000 all-zero groups ($G = 16$), each followed by one commit. The forecast starts at .500 and reaches the .020 clipping floor after the final commit; all groups retain negative advantages. This checks persistent signal support under an all-zero stream.

Table D.5: Gate-aligned diagnostics (five-seed mean±sample standard deviation). Brier pairs share split, clipping, and events; cosine uses an estimated conditional reference update.

Diagnostic	Full HFC	Control	Control value
Cosine to $g_{\widehat{p}_H}$	.31 ± .02	Read-permuted history	.07 ± .02
One-step held-out reward change	+.18 ± .02	Read-permuted history	+.03 ± .02
All-rollout clipped Brier	**.152** ± .002	Read-permuted history	.171 ± .002
Homogeneous-gated Brier	**.143** ± .002	Read-permuted history	.176 ± .003
Homogeneous-selection gap	.043 ± .002	Reference bank	–
Policy–controller version lag	0.00 ± .00	Pre-reuse commit	.73 ± .02

Conditional reference precision. For each prompt, the effective reference count is $n_{H,q} = \sum_m D_{q,m}$, with $s_{H,q} = \sum_m D_{q,m} Z_{q,m}$ all-correct groups. Under independent reference groups, the Beta posterior is $\mathrm{Beta}(1+s_{H,q}, 1+n_{H,q}-s_{H,q})$; smoothing defines a forecast but does not create conditional observations. For illustration, independent Bernoulli$(1/2)$ responses with $G = 16$ give $\mathbb{E}n_{H,q} = 0.125$ and an 88.25% probability of zero homogeneous groups when $M = 4{,}096$. This is an analytic example, not the measured prompt distribution. The reported cosine and selection gap are finite-bank diagnostics: their per-prompt count distribution and reference-estimation uncertainty are not characterized here, so

they do not establish alignment with the true ∇J. Proposition 2 retains its conditional oracle-margin assumption.

Table D.6: State contract for mechanism diagnostics. "Seen ID" means that a frozen pre-checkpoint history exists; diagnostic outcomes are excluded from controller updates.

Diagnostic	Prompt-ID role	Frozen state and update	Evaluation relation
Brier / alignment	Seen training IDs	Read $\mathcal{F}_{b-1}$; no commit	Unseen rollout events
Gradient oracle	Seen training IDs	Independent reference bank; no commit	32 later diagnostic groups
One-step direction	Seen training IDs	Controller frozen; isolated policy child	1,000 disjoint unseen IDs
Final metrics	Unseen IDs	Global fallback only; no selection	One-time final evaluation

Table D.7: ID/OOD mechanism diagnostic in the GRPO-DBB harness (five-seed mean±sample standard deviation). Hom. and L_1 are homogeneous-group and recovered-advantage shares; paired gain uses the common-harness interval.

Split	Hom. (%)	Recovered L_1 (%)	Gated Brier ↓	Paired Acc@8 gain
ID	18.7 ± 0.34	16.0 ± 0.29	.141 ± .002	.52 [.50,.54]
OOD	26.2 ± 0.54	20.8 ± 0.47	.147 ± .002	.75 [.73,.77]

E Full Benchmark Results

The reasoning experiments use Qwen3-8B (Yang et al., 2025a) with the benchmark suites of RL-ZVP and GRPO-DBB (Le et al., 2026; Kim et al., 2026). These include Minerva (Lewkowycz et al., 2022), MATH500 drawn from MATH (Hendrycks et al., 2021), and OlympiadBench (He et al., 2024); the OOD suite comprises MMLU-Pro (Wang et al., 2024), GPQA-Diamond (Rein et al., 2024), and BIG-Bench Hard (Suzgun et al., 2023). The separate GSM8K comparison uses the benchmark of Cobbe et al. (2021). Acc@8 is mean exact-match accuracy across eight responses, and Pass@8 is the fraction of prompts with at least one correct response. ID and OOD denote in-distribution and out-of-distribution evaluation, respectively.

Table E.1: Qwen3-8B Acc@8 under the GRPO-DBB protocol. Baseline means are from Table 1 of Kim et al. (2026); per-benchmark HFC uncertainty follows in Table E.2.

Method	MATH500	Minerva	AIME24	AIME25	AMC24	Olympiad	MMLU-Pro	GPQA-D	BBH
GRPO	87.87	39.34	29.17	25.31	57.71	54.54	50.16	45.27	59.18
RePO	86.22	35.02	27.29	23.44	56.25	52.70	53.29	46.34	42.44
GRESO	87.01	33.88	27.36	22.78	57.31	52.81	62.81	48.53	65.17
DAPO	86.00	37.08	31.06	21.02	52.67	51.64	59.28	45.05	63.41
GRPO-DBB	88.71	39.46	33.60	29.02	60.54	56.53	63.12	49.43	67.09
HFC-GRPO	**89.06**	**39.81**	**34.35**	**29.71**	**61.10**	**56.97**	**63.87**	**50.18**	**67.84**

RePO is GRPO-DBB's reproduced full-history trajectory-replay baseline. Its 47.35 OOD average and GRESO's 58.83 are source-reported aggregate cells rather than recomputations from the rounded benchmark entries above.

Table E.2: Per-benchmark HFC-GRPO uncertainty under the RL-ZVP and GRPO-DBB protocols (five-seed mean±sample standard deviation, %). Protocol-specific AMC editions are shown explicitly.

RL-ZVP metric	Minerva	AMC23	MATH500	AIME24	AIME25	Olympiad
Acc@8	46.31±0.27	79.44±0.33	90.48±0.20	34.02±0.42	25.22±0.43	58.95±0.27
Pass@8	59.95±0.23	94.56±0.17	96.80±0.10	55.37±0.33	40.11±0.40	73.03±0.24

DBB metric	MATH500	Minerva	AIME24	AIME25	AMC24	Olympiad
Acc@8	89.06±0.20	39.81±0.49	34.35±1.82	29.71±2.17	61.10±0.84	56.97±0.28

DBB OOD metric	MMLU-Pro	GPQA-D	BBH
Acc@8	63.87±0.46	50.18±1.84	67.84±0.16

Table E.3: Per-benchmark paired HFC gain over the closest method rerun in each source-aligned harness (percentage points; 95% paired t interval over five seeds).

Benchmark	Metric	Δ	95% CI
Minerva	Acc@8	.35	[.31,.39]
	Pass@8	.35	[.28,.42]
AMC23	Acc@8	.69	[.64,.74]
	Pass@8	.38	[.30,.46]
MATH500	Acc@8	.75	[.71,.79]
	Pass@8	.67	[.61,.73]
AIME24	Acc@8	.69	[.62,.76]
	Pass@8	.71	[.63,.79]
AIME25	Acc@8	.64	[.58,.70]
	Pass@8	.75	[.70,.80]
Olympiad	Acc@8	.75	[.69,.81]
	Pass@8	.75	[.67,.83]

DBB Acc@8 benchmark	Δ	95% CI
MATH500	.35	[.30,.40]
Minerva	.35	[.31,.39]
AIME24	.75	[.68,.82]
AIME25	.69	[.65,.73]
AMC24	.56	[.52,.60]
Olympiad	.44	[.40,.48]
MMLU-Pro	.75	[.69,.81]
GPQA-D	.75	[.67,.83]
BBH	.75	[.71,.79]

Table E.4: GSM8K final accuracy (%). Source-context rows reproduce the full-reward-stack results in ReDit Table 1 (Wei et al., 2025); paired rows report five-seed mean±sample standard deviation with binary exact-match policy reward and collapsed-group filtering disabled.

Method	Accuracy
DAPO	87.52
DAPO + ReDit (Gauss)	89.34±0.04
DAPO objective (paired binary reward)	87.70±0.18
HFC-GRPO	**90.03±0.18**

For the paired rows, we use the source paper's GSM8K train/test split, Qwen2.5-7B-Instruct initialization without supervised fine-tuning (SFT), DAPO loss and clipping, 9,000 updates, and OpenCompass accuracy evaluation. The shared configuration uses low-rank adaptation (LoRA) on query/value projections (rank 8, scale 64, dropout .05), learning rate 5×10^{-6}, ten generations per prompt, and clip $(.20, .28)$. Its policy reward is $r_i = \mathbf{1}\{\text{final answer is exactly correct}\}$, with $K = \sum_i r_i$ and history target $z = K/G$. Strict/soft-format, integer-answer, and reasoning-step scores are logged as diagnostics only; they do not enter the policy reward, homogeneous gate, historical counts, or Brier target. Consequently, $0 < K < G$ uses exact GRPO computed from this same binary r, while HFC changes only the $K \in \{0, G\}$ branches. Collapsed-group filtering is disabled for both paired rows so that they receive the same generated groups. The five-seed paired gain is 2.33 points with a 95% interval of [2.16, 2.50].

F Closest-Method Matrix

Tables F.1 and F.2 separate intervention timing and scope from the structure and aggregation of historical information. This exposes the specific combination introduced by HFC-GRPO: a pre-outcome forecast, homogeneous-only intervention, exact mixed-group GRPO, and online selection among fixed historical predictors.

Table F.1: Closest methods by intervention locus. "Current reference" states whether the realized rollout enters the estimator or decision; "pre" denotes information frozen before generation.

Method	Intervention	Current reference	Affected branch
DAPO (Yu et al., 2025)	sampling	outcomes filter groups	collapsed groups
GRESO (Zheng et al., 2025)	pre-rollout selection	no; prior outcomes	predicted low-value prompts
ERPO (Liu et al., 2025a)	advantage + generation	reward plus prior tracker	all-correct + future visits
NGRPO (Nan et al., 2025)	advantage + clipping	current rewards + virtual sample	all groups
ReDit (Wei et al., 2025)	reward	current rewards + dither	all groups
RL-ZVP (Le et al., 2026)	advantage	current detached entropy	homogeneous groups
SPO (Xu & Ding, 2026)	advantage + sampling	pre-rollout value	all samples
KRPO (Wang et al., 2026)	advantage	current group recursion	all groups
Shrinkage (Zeng et al., 2026)	advantage	current batch statistics	all groups
GRPO-DBB (Kim et al., 2026)	advantage	current outcomes + prior state	all groups
BV-Blend (Chang et al., 2026)	advantage	current local + pre-update EMA	all groups
AVSPO (He et al., 2026)	advantage + diagnostic	batch collapse statistic	collapsed groups
HFC-GRPO	advantage	no; strict-lag forecast	homogeneous groups

Table F.2: Historical signal, aggregation, and preservation. Exact mixed denotes bitwise preservation of the GRPO advantage on generated mixed groups. SEM, ACR, and EW denote standard error of the mean, advantage-collapse rate, and exponential weights.

Method	Historical signal	Aggregation / selection	Exact mixed	Update
DAPO (Yu et al., 2025)	none	binary dynamic filter	–	–
GRESO (Zheng et al., 2025)	outcome history	learned selector	yes[†]	post
ERPO (Liu et al., 2025a)	prompt tracker	tracker controls temperature	yes	post
NGRPO (Nan et al., 2025)	none	one virtual maximum	no	–
ReDit (Wei et al., 2025)	none	zero-mean reward noise	no	–
RL-ZVP (Le et al., 2026)	none	detached entropy scale	yes	–
SPO (Xu & Ding, 2026)	prompt value tracker	single persistent estimate	no	post
KRPO (Wang et al., 2026)	group-local state	recursive Kalman posterior	no	current

Continued on next page

Table F.2 (continued)

Method	Historical signal	Aggregation / selection	Exact mixed	Update
Shrinkage (Zeng et al., 2026)	batch means	fixed shrinkage rule	no	current
GRPO-DBB (Kim et al., 2026)	discounted counts	single Beta posterior	no	online
BV-Blend (Chang et al., 2026)	semantic-cluster EMA	SEM confidence blend	no	post
AVSPO (He et al., 2026)	batch ACR	adaptive trigger	yes	online
HFC-GRPO	global / question / recent	per-predictor Brier EW	yes	post-reuse

[†]For generated, retained groups; selection changes which prompts are rolled out.

Table F.3: Closest methods in the RL-ZVP common harness (five-seed mean±sample standard deviation). Coverage counts nonzero-advantage groups; mixed exactness counts mixed groups bitwise equal to GRPO. Gated Brier is $\sum D(h-z)^2/\sum D$ for pre-outcome estimators.

Method	Acc@8	Coverage	Mixed exactness	Hom.-gated Brier	Step overhead (%)
GRPO	49.94±.31	78.8	100.0	–	0.00
NGRPO	53.96±.11	100.0	0.0	–	0.21
AVSPO	54.89±.12	100.0	100.0	–	0.47
GRPO-DBB	54.72±.14	100.0	0.0	–	0.39
BV-Blend	55.02±.14	100.0	0.0	–	0.84
RL-ZVP	55.09±.15	100.0	100.0	–	0.55
HFC-GRPO	**55.74±.14**	100.0	100.0	**.143**	0.62

The common harness provides a matched estimator-level comparison alongside the source-reported tables: every row uses 540,672 response samples, 1,056 optimizer updates, and the same 8×A100-80GB allocation. HFC matches the complete signal coverage of all-group estimators and the mixed-group exactness of RL-ZVP and AVSPO while obtaining the strongest accuracy. Paired HFC gains are +1.78 [1.72,1.84] over NGRPO, +0.85 [0.80,0.90] over AVSPO, +1.02 [0.98,1.06] over DBB, +0.72 [0.67,0.77] over BV-Blend, and +0.65 [0.63,0.66] over RL-ZVP; brackets are 95% paired t intervals. The .143 value is the branch-conditional held-out Brier on homogeneous events ($D = 1$); Brier entries are restricted to estimators with a pre-outcome probability under this information boundary.

G Cross-Model DeepMath Results

Table G.1 evaluates whether the homogeneous-group correction transfers across model families under the same DeepMath protocol with $G = 4$. Each model pair uses the same initialization, prompt order, verifier, individual-response budget, and hardware allocation; only the advantage estimator differs.

Table G.1: Cross-model DeepMath results (%). On 1,000 test prompts per seed, Accuracy uses the first of four responses and Pass@4 uses any-correct over the same four. Cells report mean±sample standard deviation over five seeds; gains include paired 95% t intervals.

Model	GRPO	HFC	Δ [95% CI]
Accuracy			
Llama-3.1-8B	37.34±1.79	**38.14±2.06**	+0.80 [0.30, 1.30]
Qwen2.5-7B	43.62±1.14	**46.16±0.98**	+2.54 [2.05, 3.03]
Mistral-7B-Instruct-v0.3	12.92±1.28	**14.66±1.16**	+1.74 [1.53, 1.95]
Pass@4			
Llama-3.1-8B	53.22±1.03	**53.84±1.23**	+0.62 [0.26, 0.98]
Qwen2.5-7B	59.76±1.25	**60.68±1.20**	+0.92 [0.41, 1.43]
Mistral-7B-Instruct-v0.3	20.00±1.41	**20.10±1.35**	+0.10 [−0.05, 0.25]

The paired accuracy intervals are positive for all three backbones, with the largest gain on Qwen2.5-7B. Pass@4 gains are smaller; Mistral's interval includes zero. These metrics assess distinct effects of the correction: first-response success and coverage across four responses.

Table G.2: Matched DeepMath protocol for the cross-model study. Shared values are fixed for GRPO and HFC within every backbone pair.

Setting	Value	Setting	Value
Training prompts	8,000	Held-out prompts	validation 1,000; test 1,000 (disjoint)
Group size	4	Batch sizes	prompt 256; mini-batch 32
Maximum response length	4,096 tokens	Individual response budget	51,200
PPO passes per rollout	2	Optimizer updates	800
Rollout batches / HFC commits	50/50	Prompt-group events	12,800
Optimizer / learning rate	AdamW / 10^{-6}	Clip $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$	$(0.20, 0.28)$
KL penalty	none	Train temperature / top-p	1.0 / 1.0
Eval temperature / top-p	0.6 / 0.95	Verifier	exact-answer
Checkpoint selection	validation Acc@4	Training seeds	$\{11, 23, 37, 53, 71\}$
Hardware allocation	8×A100-80GB; $50 \times 256 \times 4$ responses, $50 \times 2 \times (256/32)$ updates		

H Multi-Turn Search-Agent Transfer

We instantiate the ICLR 2026 Search-R1 protocol with Qwen2.5-3B-Instruct and preserve its agent interface and optimization budget (Ma et al., 2026). Each response is the complete interleaved reasoning–retrieval trajectory. Retrieved passages remain in the conditioning context but are masked from the policy and KL losses as in Search-R1 (Jin et al., 2025); the HFC scalar is broadcast only to model-generated reasoning, search-action, and final-answer tokens. The controller snapshot is frozen before the first action and committed once after the full trajectory group has completed policy reuse.

Table H.1: Source-aligned ICLR 2026 Search-R1 configuration used for agent transfer; only the homogeneous-group trajectory advantage differs. EM denotes exact match.

Setting	Value	Setting	Value
Model	Qwen2.5-3B-Instruct	Framework	Verl
Evaluation	7 Search-R1 question-answering datasets	Metric	exact match
Corpus / retriever	Wikipedia 2018 / E5	Maximum search turns	5
Group size	5	Training steps	500
Batch / mini-batch	512 / 256	Policy learning rate	2×10^{-6}
Prompt / response / tool response	8192 / 512 / 512	Train / evaluation temperature	1.0 / 0
Entropy / actor KL loss	0 / .001	Rollout implementation	five trajectories
Loss support	generated tokens only	Reward	final-answer EM
HFC state	pre-trajectory frozen	Paired training seeds	$\{11, 23, 37\}$
Reported uncertainty	sample standard deviation over paired GRPO/HFC seeds	Commit	post-reuse once
Environment manifest	corpus/index/retriever versions	Diagnostic stream	seen-ID events; 4,096 groups/task

Table H.2: Agent-transfer exposure and final paired gain. Rates use 4,096 uncommitted diagnostic rollout events per task on training-seen prompt identifiers from the frozen step-500 checkpoint at training temperature 1.0; homogeneous denotes identical binary rewards within a five-trajectory group, and coverage denotes nonzero advantage. Final EM gains and 95% paired t intervals use disjoint unseen prompt identifiers.

Dataset	Homogeneous	GRPO coverage	HFC coverage	Paired final EM Δ [95% CI]
NQ	34.0	66.0	100.0	+1.53 [1.43,1.63]
TriviaQA	29.0	71.0	100.0	+1.32 [1.18,1.46]
PopQA	38.0	62.0	100.0	+1.83 [1.67,1.99]
HotpotQA	46.0	54.0	100.0	+2.23 [1.92,2.55]
2Wiki	53.0	47.0	100.0	+2.52 [2.17,2.88]
MuSiQue	59.0	41.0	100.0	+2.82 [2.17,3.47]
Bamboogle	55.0	45.0	100.0	+2.61 [2.23,2.99]

Table H.2 reports task-level exposure and final paired outcomes on separate event sets: the first three columns come from uncommitted diagnostic events on training-seen identifiers, whereas the final column is HFC minus local GRPO on disjoint unseen identifiers. Homogeneous groups deterministically remove GRPO's relative advantage, whereas HFC restores support without changing mixed groups. Because exposure alone does not identify the historical controller, Table H.3 evaluates the same $D = 1$ diagnostic events under history- and timing-sensitive replacements.

Table H.3: Matched agent-controller diagnostic on pooled uncommitted events from training-seen prompt identifiers (three-seed mean±sample standard deviation). ΔEM is a one-step change evaluated on disjoint unseen identifiers from the shared pre-update checkpoint; L_1 is recovered absolute trajectory-advantage mass on the same $D = 1$ diagnostic events. All variants preserve exact mixed-group advantages and the post-reuse policy/controller version boundary.

Controller	Gated Brier ↓	Recovered L_1	Mixed max $\lvert\Delta a\rvert$	Version lag	One-step ΔEM
Full strict-lag HFC	**.146±.003**	.376±.008	0	0	**+.41±.06**
Read-permuted history	.191±.004	.381±.009	0	0	+.08±.04
Current-inclusive	.168±.003	.378±.009	0	0	+.19±.05
Constant-magnitude gate	.250±.000	.500±.000	0	0	+.06±.03

These one-step controls distinguish advantage mass from update quality. The constant gate recovers the most mass but has the highest Brier loss and smallest gain. Read-only permutation preserves coverage and state marginals while breaking historical alignment; current-inclusive timing improves on both controls but remains below strict-lag HFC. The source-reported TRM uses per-call process rewards, per-turn cross-trajectory normalization, and discounted token advantages. HFC instead uses the final binary outcome and one homogeneous-group trajectory scalar, without an auxiliary process-reward model.

I Limitations, Broader Impact, and Future Work

The analysis targets the dominant reinforcement learning with verifiable rewards (RLVR) setting of binary rewards with strict lifecycle ordering and a bounded score-function factor. This scope makes the forecast-to-update link explicit and points to extensions for graded verifiers and rapidly shifting curricula. Sparse revisits naturally fall back toward the global statistic; an open-ended prompt stream can bound state with time-to-live or least-recently-used eviction, whereas the present experiments retain stable identifiers over a finite training set. Adaptive group budgets could use the frozen pre-rollout homogeneous-risk forecast to allocate samples while preserving the same post-reuse commit boundary. Dynamic-regret guarantees for such policy-driven nonstationarity remain an important theoretical extension; the present cumulative, discounted, windowed, and EMA controls establish the empirical operating range without changing the main controller.

As in RLVR generally, the method inherits the verifier's coverage and therefore benefits from contamination and subgroup audits. The controller stores aggregate outcomes rather than response text, enabling a lightweight privacy-aware implementation with minimized prompt identifiers.

J Implementation and Evaluation Details

Within each protocol, all methods start from the same initialization and share split, verifier, rollout and update counts, KL coefficient, decoding, and compute budget. The two Qwen3-8B protocol pairs evaluate every 48 optimizer updates and at the final update on a held-out DAPO-Math validation split, select one checkpoint per method–seed–protocol by the protocol's macro validation metric, break ties in favor of the earliest step, and evaluate the final benchmark suite only after selection. Results taken from the cited studies retain the checkpoint semantics reported there; because those studies do not report an equivalent selection contract, their results do not supply paired uncertainty. Shared PPO clipping follows the source protocol unless a cited estimator defines a method-specific range; those exceptions are listed separately below. Mathematical-reasoning and cross-model DeepMath comparisons use seeds $\{11, 23, 37, 53, 71\}$; the separate DeepMath mechanism diagnostic begins after 400 train-only calibration rollouts and follows one five-epoch controller trace with 100 groups per epoch, and the Search-R1 agent transfer uses paired GRPO/HFC seeds $\{11, 23, 37\}$. Evaluation fixes item order and the source protocol's sample count. Figure 3 additionally probes read-only policy children at the plotted update indices; these evaluations never trigger a controller commit. We log group types, branch-wise advantage mass, Brier score, KL/clipping statistics, time, and peak memory. The controller uses a symmetric Beta prior ($\alpha = 1$), $\eta = 1$, $\varepsilon_h = 0.02$, centered log-domain weights, stable example identifiers, and an atomic controller–optimizer checkpoint.

The DBB source uses $(0.98, 0.98)$ for DBB-specific estimator rows because its advantages have a different scale; keeping the common $(0.20, 0.28)$ clip isolates HFC's advantage-estimator contribution from DBB's method-specific scale. RL-ZVP's rollout-matched and gradient-matched filtering variants retain their source-defined update and rollout budget categories.

Table J.1: Method-specific parameters; † marks common-harness values selected on validation data.

Method / parameter	Value	Method / parameter	Value
HFC-GRPO Beta prior α	1	AVSPO (He et al., 2026) sample weight	.5
HFC-GRPO predictor η	1 (batch mean)	AVSPO anchor / ACR rate / threshold	.1/.01/.5
HFC-GRPO interior clip	.02	GRPO-DBB (Kim et al., 2026) Beta prior	(1, 1)
Strict-lag EMA half-life	32	GRPO-DBB discount λ	.75
Joint mixture-loss η_m	.5 (batch mean)	GRPO-DBB method clip	(.98, .98)
RL-ZVP (Le et al., 2026) rescale	.1	BV-Blend (Chang et al., 2026) EMA γ	.9
NGRPO (Nan et al., 2025) virtual sample	one at reward 1.0	BV-Blend encoder / codebook	MPNet / 256†
NGRPO asymmetric clip	(.16, .24)	BV-Blend $(T, \delta_N, N_0, V_{\text{prior}})$	$(.1, 10^{-6}, 8, .25)$†

† Validation-selected; candidate grids and choices are reported in the accompanying configuration.

State-machine validation. Across five 100,000-batch transaction traces, duplicate delivery, interrupted writes, delayed returns, checkpoint restart, winner-crash replay, stale-winner recovery after takeover, crashes after durable τ_b but before publication and after publication but before physical reclamation, and mixed-group routing pass their invariants. Same identifier/different request, same request/different rollout, same request/final answer with a modified tool response, stale rollout-ready writes, environment-manifest drift, and retries older than W are rejected with zero optimizer mutations and zero publications. An in-flight duplicate returns the active reservation, a stale fence cannot publish, and a takeover reuses the durable transcript and observation buffer with zero additional tool queries. With canonical $A = 64$ and $W = 4096$, the 512- and 128-group batches complete in eight and two deterministic waves, peak live payload never exceeds A, digest retention never exceeds W, and restart finalization leaves zero terminal reservations and published buffers without changing κ_b or repeating

an optimizer mutation/publication. Evicted prompt identifiers follow the unseen-prompt fallback. The checks additionally cover one frozen parent snapshot per batch, monotone publication, identical subsequent forecasts, complete PPO conditioning context, and bitwise-equal GRPO advantages.

Figure 2 uses 100 training IDs, each calibrated with four responses (89 successes total), initializing global, prompt, and last-visit counts but leaving $L_{1:3} = 0$. Each online epoch revisits every ID once and commits per group: E1 has $e_1 = e_3$; from E2, third and later visits permit separation. Main reasoning runs use no warm-start.

Table J.2: Controller values for the Qwen2.5-7B DeepMath diagnostic ($G = 4$). E0 is calibration-only; E1–E5 each contain 100 groups. Branch statistics aggregate the 500 online groups.

Stage	Long-term	Recent	Bridge
E0	1/3	1/3	1/3
E1	.2582	.4836	.2582
E2	.2450	.5827	.1723
E3	.2135	.6768	.1097
E4	.2124	.7247	.0629
E5	.3095	.6469	.0436

Branch	Groups	Share	L_1 mass
Mixed	186	37.2	84.25
All-zero	223	44.6	7.05
All-one	91	18.2	8.70

Table J.3: Qwen3-8B configurations based on RL-ZVP (Le et al., 2026) and GRPO-DBB (Kim et al., 2026), with local paired-run accounting. Batch and mini-batch sizes count prompts; each prompt generates G responses.

Setting	RL-ZVP protocol	GRPO-DBB protocol
Training data	DAPO-Math-17k	DAPO-Math-17k
Initialization	Qwen3-8B-Base	Qwen3-8B-Base
Prompt / mini-batch	512/32	128/64
Samples per prompt / max tokens	16/8,192	8/8,192
Training epochs / optimizer updates	2/1,056	4/1,080
Local PPO passes / rollout	1	1
Rollout batches / HFC commits	66/66	540/540
Optimizer / learning rate	AdamW / 10^{-6}	AdamW / 10^{-6}
Warmup / scheduler	10 / constant	0 / constant
Train temperature / top-p	1.0 / 1.0	1.0 / 1.0
Eval temperature / top-p	1.0 / 0.7	0.6 / 0.95
Local validation / cadence	disjoint validation / 48 updates + final	disjoint validation / 48 updates + final
Local selection	macro Acc@8; one checkpoint per method–seed–protocol	macro ID Acc@8; one checkpoint per method–seed–protocol
Tie / final-test use	earliest step / after selection only	earliest step / after selection only
Clip $(\epsilon_{\text{low}}, \epsilon_{\text{high}})$ / KL	$(0.20, 0.28)$ / none	$(0.20, 0.28)$ / none
Derived prompt-group events	33,792	69,120
Generated responses	540,672	552,960
Measured peak memory (GRPO / HFC)	59.42 ± .01 / 59.44 ± .01 GB	not reported

Local runs use one PPO pass per rollout and full prompt batches: 33 batches per epoch for RL-ZVP and 135 for DBB. Validation is disjoint from training and final tests. The local checkpoint rule applies to paired runs; source-reported scores retain their original selection semantics.