# POLICY LAG IN GRPO: GROUP-ESTIMATOR BIAS, OPTIMIZER MEMORY, AND DELAYED FEEDBACK

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Policy lag makes a learner optimize older-policy responses while rollout reward evaluates that older policy. We diagnose this mismatch in group relative policy optimization (GRPO) through its group-gradient estimator, optimizer memory, and feedback channel. In a ten-arm study with 20 paired seed blocks and a 16-update refresh interval, token-level truncated importance sampling (TIS) reduces confirmed terminal rollout failures from 12/20 to 2/20 and raises mean final-learner accuracy from 22.7% to 57.9%. Selective resets identify the first moment as the strongest reset target, and local norm-matched controls favor its update direction. We derive the cross-response bias that survives response-level importance correction and use exact enumeration to separate directional alignment, estimator bias, and finite-batch error. In a disjoint monitoring cohort with 30 development and 50 held-out runs, current-learner evaluation gives 28/35 timely alerts versus 18/35 from rollout reward, with one false alert among 13 no-episode runs for each. These findings distinguish correcting stale-data updates, managing optimizer history, and observing current-policy quality.

## 1 INTRODUCTION

Asynchronous reinforcement learning separates response generation from policy updates to improve hardware utilization (Noukhovitch et al., 2025; Fu et al., 2025). The learner then consumes older-policy responses, and rollout reward evaluates that older policy. Policy lag therefore creates two problems: how to learn from stale samples, and how to observe current-learner quality. High rollout reward can remain accurate for the generating policy while missing deterioration of the learner.

Group relative policy optimization (GRPO; Shao et al., 2024) adds a coupled estimator to this problem. Responses share the reward statistics used to normalize their advantages, so sampling mismatch affects the whole group. Importance correction modifies these gradient contributions, which then pass through Adam's accumulated moments (Kingma and Ba, 2015). Refreshing the generating policy preserves those moments. Estimator correction, policy refresh, and optimizer reset are consequently distinct operations.

We connect prior work on off-policy correction and optimizer resets (Yao et al., 2025; Zheng et al., 2026; Asadi et al., 2023; Huang et al., 2025) through a controlled GRPO study. Figure 1 locates three intervention targets. Paired training, exact estimator analysis, and held-out monitoring establish three findings:

- **Correction and optimizer memory have distinct roles.** TIS and M2PO-style correction each reduce confirmed failures from 60% to 10%. First-moment reset is the strongest selective reset; local norm controls favor its direction, while combining it with TIS gives no further gain.
- **Response correction retains cross-response bias.** Our expansion identifies the normalization-induced terms that survive response IS. Exact calculations separate aligned direction from residual bias and show that truncation's MSE ranking depends on batch size.
- **Current-learner evaluation closes an observation gap.** On 50 held-out runs, it detects degradation ahead of sustained stale-rollout episodes more often than rollout reward, with the same observed false-alert count.



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.

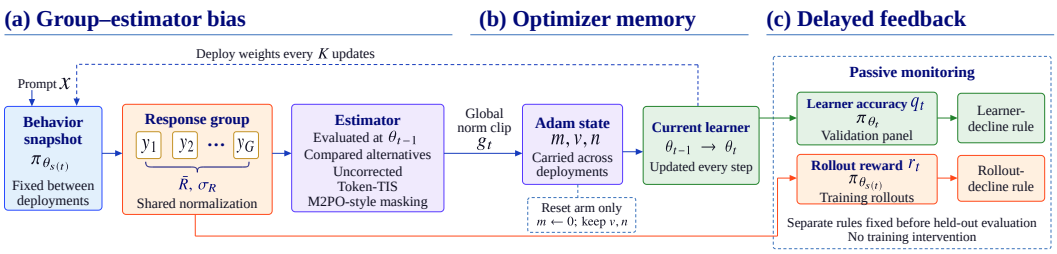



Figure 1: **Policy lag affects the update and the policy being observed.** (a) Behavior-sampled responses share group normalization before the uncorrected, token-TIS, or M2PO-style estimator is applied. (b) Adam carries optimizer state across weight deployments; the illustrated first-moment reset clears $m$ while retaining $v$ and $n$. (c) Current-learner accuracy and stale-rollout reward feed separate monitoring rules fixed before held-out evaluation. Monitoring is passive and does not change training.

## 2 Controlled Policy-Lag Protocol

**Policy versions.** Let $\theta_t$ be the learner after update $t$. With refresh interval $K$, update $t$ consumes responses from snapshot $\theta_{s(t)}$, where

$$s(t) = K \left\lfloor \frac{t-1}{K} \right\rfloor, \qquad a_t = t - 1 - s(t) \in \{0, \ldots, K-1\}. \tag{1}$$

The pre-update learner, or *proximal policy*, is $\pi_{\theta_{t-1}}$. Deployment follows updates $K, 2K, \ldots$. We vary policy age by gating weight refresh in a synchronous veRL training loop (Sheng et al., 2025), while holding the execution configuration fixed. This is a controlled study of stale-policy learning, not a throughput benchmark. Section 6 also considers batches containing multiple behavior versions.

**Objective and importance ratios.** Each prompt has $G = 8$ independently sampled responses with binary verifiable rewards. Advantages subtract the sampled group mean and divide by the sampled group standard deviation, with a numerical stabilizer. The policy term uses a PPO surrogate (Schulman et al., 2017) with a detached proximal-policy denominator and a common reference-KL penalty. Each batch supplies one full-batch optimizer update, without response reuse. Thus the proximal ratio is one at gradient evaluation: its clipping is inactive, but its derivative supplies the policy score. The behavior ratio can differ from one at that same point. This distinction follows Hilton et al. (2022, Section 2) and specifies the single-update GRPO regime tested here.

Token-TIS adapts truncated importance weighting (Yao et al., 2025) to behavior-policy lag. It multiplies token $y_{i\ell}$'s surrogate contribution by the detached weight

$$w_{i\ell} = \min\left\{ \frac{\pi_{\theta_{t-1}}(y_{i\ell} \mid x, y_{i,<\ell})}{\pi_{\theta_{s(t)}}(y_{i\ell} \mid x, y_{i,<\ell})},\ 2 \right\}. \tag{2}$$

The reference regularizer is unchanged. The M2PO-style arm combines the adaptive mask of Zheng et al. (2026), at its published threshold $0.04$, with a differentiable behavior-policy ratio. It uses the common sequence-first reduction, binary rewards, and reference regularizer. This compares complete correction rules under a matched training protocol; it is not a mask-only ablation. Appendix F specifies the selected tokens, mask statistic, and objective.

**Paired interventions.** The primary configuration is Qwen2.5-1.5B-Instruct on GSM8K (Yang et al., 2024; Cobbe et al., 2021), with $K = 16$, 300 updates, and 20 seed blocks paired across ten conditions. A block shares initialization, prompt order, assigned seed, and evaluation settings; each condition samples from its own evolving policy. Four arms form a TIS-by-first-moment-reset comparison. The other arms use full, second-moment, or counter resets; two local direction–norm controls; and M2PO-style correction.

Every update uses 64 prompts, a 1,024-token response cap, learning rate $10^{-6}$, Adam coefficients $(0.9, 0.999)$, global gradient-norm clipping at one, zero weight decay, and reference-KL coefficient

Table 1: **Primary outcomes at $K = 16$.** Twenty seed blocks are paired across conditions. C and T count confirmed failures and unconfirmed terminal tails. Accuracy is the mean (sample SD), in percent, across all 20 runs; both checkpoints use identical test items.

| Condition | C | T | Learner $\theta_{300}$ | Snapshot $\theta_{288}$ |
|---|---|---|---|---|
| Uncorrected | 12 | 2 | 22.7 (37.3) | 23.6 (36.5) |
| First-moment reset | 3 | 3 | 50.8 (41.3) | 53.1 (39.2) |
| Token-TIS | **2** | 3 | **57.9** (38.8) | 59.8 (38.2) |
| TIS + first-moment reset | 3 | 3 | 55.6 (37.5) | 54.8 (39.1) |
| M2PO-style correction | **2** | 3 | 53.9 (40.3) | 57.3 (38.5) |
| Full optimizer reset | 4 | 4 | 40.7 (39.9) | 46.3 (40.4) |
| Second-moment reset | 10 | 2 | 26.1 (38.2) | 30.5 (39.2) |
| Counter reset only | 12 | 2 | 22.6 (37.1) | 24.7 (36.7) |
| Reset direction, carried norm | 3 | 4 | 50.9 (41.4) | 51.4 (40.9) |
| Carried direction, reset norm | 9 | 3 | 26.6 (38.9) | 31.1 (40.0) |

Table 2: **Paired intervention effects.** Differences are condition B minus A, in pp. Brackets give pointwise 95% percentile intervals from 20,000 whole-block resamples. Reset $m$ denotes first-moment reset. The M2PO-style comparison uses the same blocks and a separate analysis seed.

| Comparison A → B | Failure difference | Learner difference |
|---|---|---|
| Uncorrected → reset $m$ | $-45.0$ $[-65.0, -25.0]$ | $+28.1$ $[+11.7, +44.7]$ |
| Uncorrected → TIS | $-50.0$ $[-70.0, -30.0]$ | $+35.2$ $[+18.3, +52.5]$ |
| Reset $m$ → TIS | $-5.0$ $[-15.0, 0.0]$ | $+7.1$ $[+0.3, +18.1]$ |
| TIS → TIS + reset $m$ | $+5.0$ $[0.0, +15.0]$ | $-2.3$ $[-5.7, -0.5]$ |
| Uncorrected → M2PO-style | $-50.0$ $[-70.0, -30.0]$ | $+31.2$ $[+10.5, +51.5]$ |

$10^{-3}$. Resets occur immediately after deployment, before the next update. Appendix A gives the loss and state transitions. The outcome study contains 430 runs; a disjoint 80-run cohort supplies monitoring development and evaluation (Appendix E).

**Endpoints and replication.** A *confirmed terminal rollout failure* starts at the earliest $t \geq 20$ with rollout reward below $0.05$, all subsequent rewards through update 300 below $0.15$, and at least 20 observations from $t$ onward. Later terminal candidates, starting after update 281, are *unconfirmed tails*. A separate *sustained rollout episode* consists of 20 consecutive rewards below $0.25$ and remains an episode after recovery. The first endpoint measures terminal failure by a fixed horizon; the second is the monitoring target.

The learner $\theta_{300}$ and deployed snapshot $\theta_{288}$ are evaluated on the same 1,319 GSM8K test items with identical decoding. All assigned runs, including failures and tails, contribute their observed scores. The run is the replication unit. Paired percentile intervals resample whole seed blocks, and unconfirmed tails remain in the primary failure-rate denominator (Appendix D).

## 3 Separating Estimator Correction from Optimizer Memory

### 3.1 Correction reduces failures and improves final accuracy

Token-TIS reduces confirmed failures from 12/20 to 2/20 and raises mean final-learner accuracy by 35.2 percentage points (pp), from 22.7% to 57.9% (Table 1). The paired intervals favor TIS on both endpoints (Table 2). Including unresolved terminal tails as candidates preserves the reduction: 14/20 for uncorrected training versus 5/20 for TIS. Failure incidence and all-run accuracy capture complementary aspects of performance without discarding unsuccessful runs.

M2PO-style correction also produces 2/20 confirmed failures and increases mean learner accuracy to 53.9%. Its failure indicators match TIS in every block. The paired accuracy difference, M2PO-style minus TIS, is $-4.0$ pp with interval $[-12.2, +0.2]$. Both rules improve on uncorrected training; their paired accuracy interval includes zero.

### 3.2 SELECTIVE-RESET GAINS CONCENTRATE IN THE FIRST MOMENT

First-moment reset yields 3/20 confirmed failures and 50.8% learner accuracy, compared with 10/20 and 26.1% for second-moment reset, and 12/20 and 22.6% for counter reset. It is the most effective selective reset in this comparison. Full reset gives 4/20 failures and 40.7% accuracy: clearing additional state does not improve the result.

Adding first-moment reset to TIS gives 3/20 failures and 55.6% accuracy. Relative to TIS alone, the changes are $+5.0$ pp $[0.0, +15.0]$ in failure incidence and $-2.3$ pp $[-5.7, -0.5]$ in accuracy. The reset's failure contrast changes from $-45$ pp without TIS to $+5$ pp with TIS, a risk-scale interaction of $+50$ pp $[+30, +70]$. This is less-than-additive risk reduction on a bounded probability scale, not evidence by itself of mechanistic antagonism (Appendix D).

Deployment and reset act on different state variables. Let $g_t$ be the globally clipped ascent-sign gradient. After deployment at $\tau$, a carried first moment satisfies

$$m_{\tau+j} = \beta_1^j m_\tau + (1-\beta_1)\sum_{\ell=1}^{j} \beta_1^{j-\ell} g_{\tau+\ell}. \tag{3}$$

Refreshing the generating policy leaves the inherited term $\beta_1^j m_\tau$ intact; resetting $m$ removes it. The parameter update also depends on the second moment and bias correction. The next comparison separates the reset's instantaneous direction from its norm.

### 3.3 LOCAL NORM CONTROLS FAVOR THE RESET DIRECTION

Each control trajectory maintains a carried first moment $m^C$ and a reset first moment $m^R$. Both consume the same clipped gradient and share that trajectory's current parameters, second moment, and counter. For their candidate parameter updates $u^C$ and $u^R$, the applied rules are

$$u^D = \frac{\|u^C\|_2}{\|u^R\|_2} u^R, \qquad u^N = \frac{\|u^R\|_2}{\|u^C\|_2} u^C. \tag{4}$$

The first uses the reset direction at the carried norm; the second uses the carried direction at the reset norm. With zero weight decay, these norms cover the entire instantaneous parameter update. Appendix B specifies cross-shard aggregation and zero-norm handling.

The reset-direction control gives 3/20 failures and 50.9% accuracy, whereas the reset-norm control gives 9/20 and 26.6%. The larger benefit follows the reset direction under local norm matching. This comparison matches candidate updates at a shared instantaneous state within each trajectory; subsequent gradients, optimizer histories, and responses evolve under the applied rule. It therefore tests local update rules with downstream training feedback retained.

## 4 WHAT IMPORTANCE CORRECTION CHANGES IN A GROUP ESTIMATOR

The training results motivate three distinct questions: does correction align the expected gradient direction, recover an on-policy estimator, or reduce finite-batch error? We answer them at fixed learner and behavior policies, before global clipping and Adam. The sampled-prefix reference regularizer remains a separate term.

### 4.1 SHARED NORMALIZATION CHANGES THE CORRECTION TARGET

For a prompt $x$, let $Y = (y_1, \ldots, y_G)$ contain independent responses from $\pi_\phi$. Fix the learner $\theta$ and define its sequence-summed group contribution

$$h_\theta(x, Y) = \frac{1}{G}\sum_{i=1}^{G} A_i(Y) s_\theta(x, y_i), \qquad s_\theta(x, y) = \nabla_\theta \log \pi_\theta(y \mid x). \tag{5}$$

The training reduction instead averages token scores within each response:

$$h_\theta^{\text{train}}(x, Y) = \frac{1}{G}\sum_{i=1}^{G} \frac{A_i(Y)}{L_i} \sum_{\ell=1}^{L_i} \nabla_\theta \log \pi_\theta(y_{i\ell} \mid x, y_{i,<\ell}). \tag{6}$$

The following expansion applies to either fixed contribution. In both cases, advantages retain their sampled group mean and standard deviation. Define

$$u_G(\theta) = \mathbb{E}_{x,Y\sim\pi_\theta^G}[h_\theta(x,Y)], \qquad b_G(\theta,\phi) = \mathbb{E}_{x,Y\sim\pi_\phi^G}[h_\theta(x,Y)] - u_G(\theta). \tag{7}$$

This reference keeps the estimator's normalization and reduction unchanged. It is generally not the expected-reward gradient.

**Proposition 1** (Local group-distribution bias)**.** *Assume common support, twice differentiable sampling probabilities, valid differentiation under the expectation, and an integrable second-derivative bound. For $\Delta = \theta - \phi$ and $S_\theta = \sum_i s_\theta(x, y_i)$,*

$$b_G(\theta, \theta-\Delta) = -M_G(\theta)\Delta + O(\|\Delta\|_2^2), \qquad M_G = \mathbb{E}_{\pi_\theta^G}[h_\theta S_\theta^\top]. \tag{8}$$

The score of the product sampling distribution gives the first-order term. For Equation (5), $M_G = G^{-1}\sum_{i,j}\mathbb{E}[A_i s_i s_j^\top]$: shared normalization retains cross-response terms with $i \neq j$. Appendix H gives the proof and an explicit remainder bound.

**What response-level IS removes.** The distinction can be made exact at first order. Write $h_\theta = G^{-1}\sum_i f_{\theta,i}$, with $f_{\theta,i} = A_i s_i$ for Equation (5), or $f_{\theta,i} = A_i s_i / L_i$ for Equation (6). Define

$$D_G = \frac{1}{G}\sum_i \mathbb{E}_{\pi_\theta^G}[f_{\theta,i}s_i^\top], \qquad C_G = \frac{1}{G}\sum_{i\neq j}\mathbb{E}_{\pi_\theta^G}[f_{\theta,i}s_j^\top], \tag{9}$$

so $M_G = D_G + C_G$.

**Corollary 1** (Residual bias after response correction)**.** *Under the same local regularity conditions for the response-weighted expectation, let $h^{\mathrm{IS}}_{\theta,\phi} = G^{-1}\sum_i[\pi_\theta(y_i \mid x)/\pi_\phi(y_i \mid x)]f_{\theta,i}$. Then*

$$\mathbb{E}_{\pi^G_{\theta-\Delta}}[h^{\mathrm{IS}}_{\theta,\theta-\Delta}] - u_G(\theta) = -C_G(\theta)\Delta + O(\|\Delta\|_2^2). \tag{10}$$

Response IS removes the self-response term $D_G$, but the group partners still contribute $C_G$. Differentiating a weight with respect to the behavior policy is essential to this calculation; detaching it during learner optimization does not remove that dependence.

Exact correction instead reweights the joint response group:

$$W_G(x,Y) = \prod_{i=1}^G \frac{\pi_\theta(y_i \mid x)}{\pi_\phi(y_i \mid x)}, \qquad \mathbb{E}_{\pi_\phi^G}[W_G h_\theta] = u_G(\theta). \tag{11}$$

By contrast, weighting only response $i$ changes its summand's sampling law to

$$\mathbb{E}_{\pi_\phi^G}\left[\frac{\pi_\theta(y_i \mid x)}{\pi_\phi(y_i \mid x)} f_{\theta,i}(x,Y)\right] = \mathbb{E}_{y_i\sim\pi_\theta,\, Y_{-i}\sim\pi_\phi^{G-1}}[f_{\theta,i}(x,Y)]. \tag{12}$$

A token ratio additionally leaves the prefix and continuation behavior-conditioned. These are different correction targets: response or token weighting does not generally recover the estimator-matched on-policy mean.

### 4.2 Direction, bias, and finite-batch error are distinct

We enumerate all 256 ordered groups for a four-action softmax learner with probabilities $p = (0.1, 0.2, 0.3, 0.4)$, rewards $(0, 0, 0, 1)$, and $G = 4$. The behavior grid contains $p$ and four mixtures toward each of $q_{\mathrm{rev}} = (0.4, 0.3, 0.2, 0.1)$ and $q_{\mathrm{rare}} = (0.4, 0.35, 0.24, 0.01)$. Responses have one token, so response and token weighting coincide. Homogeneous groups contribute zero; the finite-state calculation uses no numerical stabilizer. Independent aggregation over 35 multinomial count states checks every moment (Appendix I).

For response-TIS with cap $c$, let $\mu_c$ and $\Sigma_c$ be its mean and covariance, $\mu_\infty$ the untruncated response-IS mean, and $u$ the on-policy group-estimator mean. Then

$$\mu_c - u = (\mu_\infty - u) + (\mu_c - \mu_\infty), \qquad \mathbb{E}\|\bar{g}_c - u\|_2^2 = \|\mu_c - u\|_2^2 + \frac{\operatorname{tr}\Sigma_c}{B}, \tag{13}$$

where $\bar{g}_c$ averages $B$ independent groups. The first identity separates residual group error from the truncation shift; the second separates squared bias from variance.

Table 3: **Exact estimator error at the behavior endpoints.** Bias is relative to the same group estimator's on-policy mean. The variance column is the covariance trace divided by $B = 64$. Joint-group IS is unbiased. All entries are rounded independently.

| Endpoint | Estimator | Squared bias | Variance / 64 | MSE |
|---|---|---|---|---|
| Reverse | Unweighted | 0.08815 | 0.00119 | 0.08934 |
| | Response IS | 0.00039 | 0.00701 | 0.00740 |
| | TIS, cap 2 | 0.03820 | 0.00223 | 0.04043 |
| | Joint-group IS | 0 | 0.10236 | 0.10236 |
| Rare reward | Unweighted | 0.19314 | 0.00018 | 0.19331 |
| | Response IS | 0.00091 | 0.09098 | 0.09189 |
| | TIS, cap 2 | 0.18432 | 0.00033 | 0.18465 |
| | Joint-group IS | 0 | 52.25541 | 52.25541 |

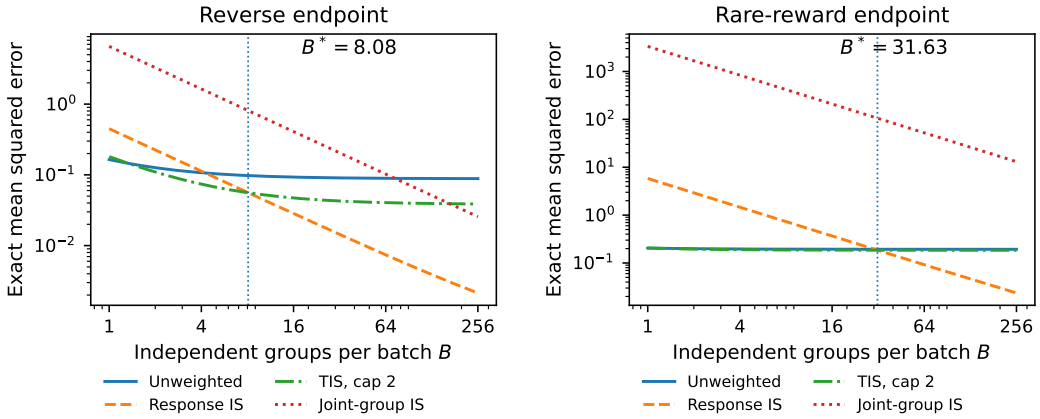



Figure 2: **Batch size changes the error ranking of truncated and untruncated correction.** Curves show exact MSE for independent groups at fixed learner and behavior distributions. Vertical lines mark the TIS/response-IS crossovers. Joint-group IS removes bias at a potentially large variance cost.

**Direction can align while bias remains.** At the reverse and rare-reward endpoints, the unweighted means have cosines $0.9439$ and $0.9551$ with $u$. Response IS and cap-two TIS have cosine one at both endpoints, despite nonzero bias (Table 3). With one rewarding action and untruncated nonreward-action ratios, both weighted means are positive multiples of $e_4 - p$ in softmax-logit coordinates (Proposition 2). The remaining error is in magnitude. A positive scalar rescaling of the unweighted mean preserves its angle and cannot reproduce this directional correction.

**Truncation trades variance for bias.** At $B = 64$, cap-two TIS has lower MSE than the unweighted estimator at all eight off-policy grid points. Against response IS, it reduces variance but increases squared bias at both endpoints. At the rare-reward endpoint and $B = 16$, its MSE is $0.18565$, versus $0.36483$ for response IS; at $B = 64$, response IS has lower MSE. When the variance reduction and squared-bias increase are both positive, the crossover is

$$B^* = \frac{\operatorname{tr} \Sigma_\infty - \operatorname{tr} \Sigma_2}{\|\mu_2 - u\|_2^2 - \|\mu_\infty - u\|_2^2}. \tag{14}$$

Cap-two TIS has lower error for $B < B^*$. The crossovers are $8.08$ and $31.63$ (Figure 2). Joint-group IS is unbiased but has rare-reward MSE $52.26$ at $B = 64$. Exact distribution correction, directional alignment, and minimum finite-batch error are therefore different objectives. These fixed-state calculations explain the estimator distinctions; the LLM interventions evaluate their downstream effect after clipping and Adam.

Table 4: **Held-out monitoring in LLM training.** Timely (1–16 updates) and early ($> 16$ updates) first alerts use 35 eligible episode runs. False alerts use 13 no-episode runs; two unresolved tails are excluded. Median lead is in learner updates among timely alerts.

| Diagnostic | Timely | False alerts | Early | Median lead |
|---|---|---|---|---|
| Rollout reward | 18/35 | 1/13 | 1/35 | 5 |
| Log-ratio mismatch | 13/35 | 0/13 | 3/35 | 1 |
| Current-learner evaluation | **28/35** | 1/13 | 7/35 | 5 |
| Joint diagnostic | 27/35 | 1/13 | 6/35 | 5 |

## 5 Current-Learner Evaluation Exposes Delayed Feedback

The optimization results concern the learner's updates. Monitoring raises a separate question: *which policy does the observed reward evaluate?* An illustrative $K = 16$ run makes the distinction concrete: greedy learner accuracy falls from $0.61$ at update 112 to zero at update 128, while stale-rollout reward remains between $0.58$ and $0.74$. Low rollout reward appears at update 129, after deployment (Appendix G). Because these channels use different samples and decoding, the monitoring study evaluates their alert timing rather than interpreting their numerical difference as a matched policy-quality estimate.

**Frozen monitoring protocol.** We compare four diagnostics on 30 development and 50 independently assigned held-out LLM training runs covering periodic refresh and mixed-age queues. Every four updates, the current learner is greedily evaluated on a fixed 256-prompt validation panel. With verifier $V(x, y) \in \{0, 1\}$ and greedy response $\widehat{y}_\theta(x)$, its accuracy is

$$q_t = \frac{1}{256} \sum_{j=1}^{256} V(x_j, \widehat{y}_{\theta_t}(x_j)). \tag{15}$$

The other channels are batch rollout reward, token-pooled squared behavior/current log ratios, and the fraction of mixed-reward groups. The four rules use rollout decline, log-ratio mismatch, learner-accuracy decline, or a fixed combination of learner decline, mismatch, and group-diversity decline. Normalization and thresholds are chosen on development runs and frozen before held-out evaluation.

A timely first alert occurs 1–16 updates before the onset of 20 consecutive rollout rewards below $0.25$. Earlier, late, and missing alerts are separate categories. The held-out cohort has 35 eligible episode runs, 13 no-episode runs, and two unresolved tails. Appendix C specifies all features, thresholds, eligibility rules, and setting-level outcomes.

**Direct evaluation detects degradation before rollout feedback.** Current-learner evaluation gives 28/35 timely alerts, versus 18/35 for rollout reward: a paired gain of $28.6$ pp $[8.6, 48.6]$, with one false alert among 13 no-episode runs for each (Table 4). Thirteen runs have a timely learner-only alert and three have a timely rollout-only alert. The joint rule gives 27/35 timely alerts, so combining channels does not improve on direct evaluation here.

All 35 eligible episode runs receive a learner-evaluation alert before rollout-episode onset: 28 timely and seven early. In 22 of the 28 timely cases, learner accuracy is already below $0.25$ at the alert. The benefit is *earlier detection of realized learner degradation*, rather than prediction before learner quality declines. Monitoring is passive and evaluates detection, not intervention.

The fixed timing window explains variation across settings. At $K = 32$, four learner alerts lead onset by 21, 21, 21, and 29 updates and are classified as early. In the long-age queue, rollout reward gives 6/7 timely alerts, versus 5/7 for learner evaluation. The pooled comparison summarizes the specified mixture of lag schedules. Separately, all nine non-M2PO arms' final learner-minus-snapshot accuracy intervals include zero (Appendix D); a terminal checkpoint difference and a within-run observation delay are different quantities.

## 6 Coverage Across Models and Lag Schedules

Three additional model/task cells use ten paired runs per arm: Qwen2.5-7B/GSM8K, Llama-3.2-3B/GSM8K, and Qwen2.5-1.5B/MATH (Meta, 2024; Hendrycks et al., 2021). TIS reduces confirmed failures from 6/10 to 3/10, 7/10 to 4/10, and 7/10 to 3/10, respectively; first-moment reset gives 5/10, 5/10, and 6/10. TIS improves the observed failure count in all three cells, with smaller reductions than in the primary configuration (Appendix E).

At $K = 1$, uncorrected training and first-moment reset each give 1/20 confirmed failures. The reset interval remains 16 updates, separating reset frequency from policy age. At $K = 8$ and $K = 32$, TIS changes failures from 5/10 to 1/10 and from 8/10 to 3/10.

Mixed-age queues select one behavior version per prompt group after 24 synchronized startup updates. Narrow, uniform, and long-age profiles all have post-startup mean age 7.5 updates, but differ in age concentration, maximum age, and zero-age mass. TIS changes failures from 5/10 to 1/10, 7/10 to 2/10, and 2/10 to 0/10, respectively. Interventions are paired within each profile; profiles use separate blocks. These results compare complete age distributions, rather than isolating the effect of age variance.

## 7 Related Work

**Asynchronous and off-policy learning.** Noukhovitch et al. (2025) study the efficiency–staleness trade-off in asynchronous RLHF. Cleanba (Huang et al., 2024) shows that actor–learner interaction affects reproducibility, and AReaL (Fu et al., 2025) develops a scalable asynchronous reasoning system. Leftover Lunch (Baheti et al., 2024) studies advantage-based offline policy gradients with response-level importance weights and a per-token approximation. Our controlled-refresh design uses the behavior/proximal separation of Hilton et al. (2022) to examine GRPO's coupled group estimator. The local expansion complements staleness–learning-rate analysis (Song et al., 2026) by exposing self-response and cross-response terms.

**Correction and optimizer state.** Yao et al. (2025) use token-level TIS to address implementation-induced rollout–training mismatch; here we apply capped token weights to stale behavior policies. M2PO (Zheng et al., 2026) adaptively masks selected extreme log-ratio contributions. VCPO (Huang et al., 2026) combines sequence-level TIS, effective-sample-size-based learning-rate control, and an off-policy baseline. We compare TIS and a matched M2PO-style correction, then separate the estimator's direction, residual bias, and finite-batch error. Optimizer resets address changing deep-RL objectives (Asadi et al., 2023), while SPAM (Huang et al., 2025) resets both moments and uses spike-aware clipping for LLM training. Our selective resets and direction–norm exchanges distinguish the first moment's role across behavior-policy refreshes.

**Optimization signals and model quality.** Reward-model ensembles mitigate reward overoptimization (Coste et al., 2024), and likelihood-displacement analysis explains how preference optimization can lower the likelihood of desirable responses (Razin et al., 2025). We study a distinct measurement problem with verifiable rewards: rollout feedback can correctly evaluate an older policy while failing to describe the current learner. The held-out comparison measures the detection-time gain from evaluating that learner directly.

## 8 Conclusion

Policy lag changes both the applied update and the observed policy. Paired GRPO interventions distinguish importance correction from optimizer memory: TIS and M2PO-style correction reduce failures, while local norm controls favor the first-moment-reset direction. The analysis identifies the cross-response bias left by response correction and separates direction, bias, and finite-batch error. Held-out monitoring shows that current-learner evaluation detects degradation ahead of stale-rollout feedback. Reliable lagged training therefore calls for separate attention to estimator correction, optimizer memory, and policy-version-aware evaluation.

## Reproducibility Statement

Appendices A–F specify the objective, policy versions, interventions, endpoint definitions, monitoring rules, run allocation, and paired analyses. Appendix G documents an illustrative checkpoint record. Appendices H–I provide the estimator identities, proofs, and finite-state specification. The supplementary verification programs enumerate the finite-state models in two independent ways, check all displayed exact-error values, regenerate Figure 2, and validate aggregate accounting. These deterministic checks are separate from the LLM training runs and their run-level statistical analyses.

## AI Use Statement

Generative AI tools assisted with literature discovery, manuscript organization and language editing, refinement of the conceptual framing, formulation and checking of mathematical claims, proof exposition, interpretation of reported results, and preparation of deterministic verification scripts and figures. The authors are responsible for the experimental records, analyses, and final content of this paper.

## References

Kavosh Asadi, Rasool Fakoor, and Shoham Sabach. Resetting the optimizer in deep RL: An empirical study. In *Advances in Neural Information Processing Systems*, volume 36, 2023. https://proceedings.neurips.cc/paper_files/paper/2023/hash/e4bf5c3245fd92a4554a16af9803b757-Abstract-Conference.html.

Ashutosh Baheti, Ximing Lu, Faeze Brahman, Ronan Le Bras, Maarten Sap, and Mark Riedl. Leftover Lunch: Advantage-based offline reinforcement learning for language models. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/3b18d368150474ac6fc9bb665d3eb3da-Abstract-Conference.html.

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. https://arxiv.org/abs/2110.14168.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/dda7f9378a210c25e470e19304cce85d-Abstract-Conference.html.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, Tongkai Yang, Binhang Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. https://proceedings.neurips.cc/paper_files/paper/2025/hash/33c00862bfa29ac72ecf630a41e19352-Abstract-Conference.html.

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. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17086–17098, 2022. https://proceedings.neurips.cc/paper_files/paper/2022/hash/6ceb6c2150bbf46fd75528a6cd6be793-Abstract-Conference.html.

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

Shengyi Huang, Jiayi Weng, Rujikorn Charakorn, Min Lin, Zhongwen Xu, and Santiago Ontañón. Cleanba: A reproducible and efficient distributed reinforcement learning platform. In *International Conference on Learning Representations*, 2024. https://proceedings.iclr.cc/paper_files/paper/2024/hash/1bd8a2967508c5cf068ac204151b6ecd-Abstract-Conference.html.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. In *International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/7a70ad3d9c704fb9b81b5c69eda722dc-Abstract-Conference.html.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. https://arxiv.org/abs/1412.6980.

Meta. Llama-3.2-3B-Instruct model card, 2024. https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. In *International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/0b99315234cc95e6ef281f9155b68832-Abstract-Conference.html.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations*, 2025. https://proceedings.iclr.cc/paper_files/paper/2025/hash/3df38ca67befaed9c03b95ffee07d9f8-Abstract-Conference.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 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. https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, 2025. doi: 10.1145/3689031.3696075. https://arxiv.org/abs/2409.19256.

Jingwei Song, Haofeng Xu, Jie Xiao, Chengke Bao, Pengbin Feng, Jingwei Shi, Weixun Wang, Yuhang Han, Chuan Wu, Linfeng Zhang, and Bill Shi. Staleness–learning rate scaling laws for asynchronous RLHF. *arXiv preprint arXiv:2607.01083v2*, 2026. https://arxiv.org/abs/2607.01083v2.

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*, 2024. https://arxiv.org/abs/2412.15115.

Feng Yao, Liyuan Liu, Dinghuai Zhang, Chengyu Dong, Jingbo Shang, and Jianfeng Gao. Your efficient RL framework secretly brings you off-policy RL training. Technical blog, August 5, 2025. Updated October 13, 2025. https://fengyao.notion.site/off-policy-rl.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? In *International Conference on Learning Representations*, 2026. https://proceedings.iclr.cc/paper_files/paper/2026/hash/8590b946e4f0c2b1262cb5b047924653-Abstract-Conference.html.

# A Training Protocol and Endpoint Definitions

## A.1 Sampling, pairing, and evaluation

The primary configuration is Qwen2.5-1.5B-Instruct on GSM8K. Each update uses $B = 64$ training prompts and $G = 8$ independently sampled responses per prompt, temperature one, top-$p = 1$, no top-$k$ truncation, and a 1,024-token response cap including EOS. The MATH configuration uses a 4,096-token cap. Each batch supplies one full-batch optimizer update, without response reuse.

A paired block shares initialization, prompt order, assigned seed, evaluation items, and schedule across conditions. Each arm samples from its own evolving policy. A fixed panel of 256 validation prompts is reserved; the learner and deployed checkpoint are greedily evaluated every four updates. Final GSM8K accuracy uses the same 1,319 test items for both checkpoints, separate from training and validation. Accuracy is verifier acceptance. The monitoring study uses the validation panel and logged rollout, behavior/current log-ratio, and reward-diversity channels.

## A.2 Loss and normalization

For binary reward $R_{bi}$, define

$$\overline{R}_b = \frac{1}{G}\sum_{j=1}^{G} R_{bj}, \qquad A_{bi} = \frac{R_{bi} - \overline{R}_b}{\sqrt{G^{-1}\sum_{j=1}^{G}(R_{bj} - \overline{R}_b)^2 + 10^{-6}}}. \tag{A.1}$$

The empirical variance uses divisor $G$. Homogeneous-reward groups have zero advantages. Rewards and advantages are detached. Let $L_{bi}$ count valid generated tokens, including generated EOS but excluding prompt and padding tokens. Both the policy and reference-KL terms average first over tokens within each response, then over responses and prompts.

Let $\pi_{\mathrm{prox}} = \pi_{\theta_{t-1}}$ be detached immediately before the update, and let $\pi_{\mathrm{beh}}$ be the generating policy. At token $\ell$ with prefix $h_{bi\ell}$, write

$$\rho_{bi\ell}(\theta) = \frac{\pi_\theta(y_{bi\ell} \mid h_{bi\ell})}{\pi_{\mathrm{prox}}(y_{bi\ell} \mid h_{bi\ell})}, \qquad w_{bi\ell} = \min\left\{\frac{\pi_{\mathrm{prox}}(y_{bi\ell} \mid h_{bi\ell})}{\pi_{\mathrm{beh}}(y_{bi\ell} \mid h_{bi\ell})}, 2\right\}. \tag{A.2}$$

The uncorrected policy loss is

$$\mathcal{L}_{\mathrm{pol}}(\theta) = -\frac{1}{B}\sum_{b=1}^{B}\frac{1}{G}\sum_{i=1}^{G}\frac{1}{L_{bi}}\sum_{\ell=1}^{L_{bi}} \min\{\rho_{bi\ell}(\theta)A_{bi},\ \mathrm{clip}(\rho_{bi\ell}(\theta), 0.8, 1.2)A_{bi}\}. \tag{A.3}$$

Token-TIS multiplies each surrogate summand by detached $w_{bi\ell}$. At the single gradient-evaluation point, $\rho = 1$, so proximal clipping is inactive and the derivative supplies the policy score. The behavior ratio can differ from one.

The shared regularizer is categorical KL divergence to the frozen initial checkpoint, evaluated at sampled prefixes:

$$d_{\mathrm{ref}}(h;\theta) = \sum_v \pi_\theta(v \mid h)\log\frac{\pi_\theta(v \mid h)}{\pi_{\mathrm{init}}(v \mid h)}, \tag{A.4}$$

$$\mathcal{R}_{\mathrm{ref}}(\theta) = \frac{1}{B}\sum_b \frac{1}{G}\sum_i \frac{1}{L_{bi}}\sum_{\ell=1}^{L_{bi}} d_{\mathrm{ref}}(h_{bi\ell};\theta). \tag{A.5}$$

The minimized loss is $\mathcal{L}_{\mathrm{pol}} + 10^{-3}\mathcal{R}_{\mathrm{ref}}$, with the appropriate policy term for each arm. TIS weights only the policy surrogate. The M2PO-style arm uses the same unmasked regularizer (Appendix F). Fixing the regularizer across arms does not change the estimator reference in Section 4, which concerns the policy-gradient contribution alone.

## A.3 Optimizer and reset timing

We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_A = 10^{-8}$, constant learning rate $\eta = 10^{-6}$, and zero weight decay. Global gradient-norm clipping at one precedes the optimizer step. With ascent-sign

clipped gradients $g_t$, the update is

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t \odot g_t, \tag{A.6}$$

$$n_t = n_{t-1} + 1, \qquad \widehat{m}_t = \frac{m_t}{1-\beta_1^{n_t}}, \quad \widehat{v}_t = \frac{v_t}{1-\beta_2^{n_t}}, \tag{A.7}$$

$$\theta_t = \theta_{t-1} + \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon_A}. \tag{A.8}$$

The counter increments before bias correction; vector divisions and square roots are coordinatewise. With $K = 16$, resets occur after the boundary update and weight deployment, before the next update. First-moment reset clears $m$; second-moment reset clears $v$; counter reset clears $n$; full reset clears all three. At $K = 1$, behavior weights refresh every update while resets remain every 16 updates. Reset frequency is therefore fixed when comparing these two policy-age settings.

### A.4 Failures, episodes, and terminal checkpoints

A confirmed terminal rollout failure begins at the earliest $t \geq 20$ with rollout reward below $0.05$, all subsequent rewards below $0.15$ through $T = 300$, and at least 20 observations from $t$ onward. The latest confirmable start is 281. A later terminal candidate is an unconfirmed tail. A sustained rollout episode begins at the first of 20 consecutive rewards below $0.25$; recovery does not erase it.

Numerical failures and incomplete horizons are separate statuses. The reported outcome runs have complete horizons and no numerical failures. The primary binary endpoint is confirmation by update 300: tails remain in the denominator and are also reported separately. Table 7 retains the three-state accounting.

The final learner is $\theta_{300}$. The deployed checkpoint after update $T$ has index

$$d(T) = K\lfloor T/K \rfloor, \qquad d(300) = 288 \quad (K = 16). \tag{A.9}$$

Both checkpoints use identical test items and decoding, without additional final synchronization. The last rollout reward is sampled from the behavior policy before the final learner update and is a separate training endpoint.

The run is the analysis unit. Means and sample SDs include all assigned runs' observed scores, including failures and tails. No replacement score is assigned to failed runs. Test items, rollout windows, and repeated probes are not independent run replications.

## B Shadow-State Direction–Norm Controls

Each local control maintains carried and reset first moments along its own trajectory. Initialize $m_0^C = m_0^R = v_0 = 0$ and $n_0 = 0$. After deployment at $\tau = jK$, set $m_\tau^R = 0$ while retaining $m_\tau^C$, $v_\tau$, and $n_\tau$. At the next update, both moments consume the same trajectory-specific clipped gradient:

$$m_t^a = \beta_1 m_{t-1}^a + (1-\beta_1) g_t, \qquad u_t^a = \eta \frac{m_t^a/(1-\beta_1^{n_t})}{\sqrt{v_t/(1-\beta_2^{n_t})} + \epsilon_A}, \quad a \in \{C, R\}. \tag{B.1}$$

Candidates share the current parameters, second moment, and unreset optimizer counter. The carried-state and first-moment-reset reference rules apply $u_t^C$ and $u_t^R$, respectively. The controls apply

$$u_t^D = \frac{\|u_t^C\|_2}{\|u_t^R\|_2} u_t^R, \qquad u_t^N = \frac{\|u_t^R\|_2}{\|u_t^C\|_2} u_t^C. \tag{B.2}$$

Norms aggregate all trainable coordinates across optimizer shards. With zero weight decay, they are norms of the complete instantaneous update. If either candidate norm is zero, the fallback applies $u_t^C$ and records the event. A nonfinite candidate is classified as a numerical failure.

Each condition evolves its own second moment and shadow histories. Pairing fixes initialization and seed assignments, while subsequent gradients and responses follow the applied rule. The controls isolate direction and norm at each shared instantaneous state within a trajectory, retaining downstream feedback from that rule.

# C Empirical Monitoring in LLM Training

The monitoring cohort contains 30 development and 50 independently assigned held-out runs, each with 300 learner updates, covering periodic refresh and mixed-version queues. Development runs determine signal scaling and thresholds. The held-out cohort is used once for first-alert evaluation.

## C.1 Observation channels and features

At each four-update probe, $q_t$ is greedy current-learner accuracy on the fixed 256-prompt validation panel, $r_t$ is update-level rollout reward, $\kappa_t$ is token-pooled behavior/current log-ratio mismatch, and $d_t$ is the fraction of 64 prompt groups containing both binary reward values. Thus

$$q_t = \frac{1}{256}\sum_{j=1}^{256} V(x_j, \widehat{y}_{\theta_t}(x_j)). \tag{C.1}$$

Validation items and greedy decoding remain fixed. For the $N_t$ valid generated tokens at update $t$, define

$$\kappa_t = \frac{1}{N_t}\sum_{n=1}^{N_t}\left[\log \pi_{\theta_{t-1}}(y_n \mid h_n) - \log \pi_{\mathrm{beh},n}(y_n \mid h_n)\right]^2. \tag{C.2}$$

The stored checkpoint identity selects $\pi_{\mathrm{beh},n}$ for each response, so one update can contain several behavior versions. This squared log-ratio statistic is distinct from the reference KL in Equation (A.4). Group reward heterogeneity is

$$d_t = \frac{1}{64}\sum_{g=1}^{64}\mathbf{1}\left\{0 < \sum_{i=1}^{8} R_{tgi} < 8\right\}. \tag{C.3}$$

At $t \in \{4, 8, \ldots, 300\}$, set $\overline{v}_t = \frac{1}{4}\sum_{j=0}^{3} v_{t-j}$ for $v \in \{r, \kappa, d\}$, and let $H_t = \{4, 8, \ldots, t\}$. The feature vector is

$$x_t = \left(\max_{s\in H_t}\overline{r}_s - \overline{r}_t,\ \overline{\kappa}_t,\ \max_{s\in H_t} q_s - q_t,\ \max_{s\in H_t}\overline{d}_s - \overline{d}_t\right). \tag{C.4}$$

Historical maxima include the current probe. Alerts are eligible from update 24. Channels use saved learner checkpoints, generated responses, and training logs, with observations available by the probe time.

## C.2 Development and fixed diagnostic rules

Pool eligible development probes to estimate feature means $\mu_k$ and population standard deviations with scale floor $0.02$:

$$\sigma_k = \max\left\{\sqrt{\frac{1}{N_P}\sum_{i,t}(x_{i,t,k} - \mu_k)^2},\ 0.02\right\}, \qquad z_{t,k} = \frac{x_{t,k} - \mu_k}{\sigma_k}. \tag{C.5}$$

Here $N_P$ counts pooled development probes. The rollout, mismatch, and learner scores are $z_{t,1}$, $z_{t,2}$, and $z_{t,3}$, respectively. The joint score is $0.7z_{t,3} + 0.2z_{t,2} + 0.1z_{t,4}$. Each rule reports its first strict threshold crossing:

$$t_A = \min\{t \in \{24, 28, \ldots, 264\} : S_t > c\}. \tag{C.6}$$

An empty set means no alert; later crossings do not replace the first.

Development runs $i = 0, \ldots, 29$ use whole-run folds $f(i) = \lfloor i/5 \rfloor \bmod 5$, with sizes $10, 5, 5, 5, 5$. For each $c \in \{0, 0.1, \ldots, 4\}$, scaling is fitted on the other folds and alerts are evaluated in the held-out fold. The pooled selection criterion is

$$U(c) = \mathrm{TimelyRecall}(c) - \mathrm{FPR}(c) - 0.25\,\mathrm{EarlyFraction}(c). \tag{C.7}$$

Recall and early fraction use eligible episodes; the false-positive rate uses no-episode runs. A zero denominator contributes zero. Ties favor lower false-positive rate, then a higher threshold. Scaling is refitted on all development runs, and the entire rule is frozen before held-out evaluation.

Table 5: **Fixed development parameters for monitoring.** Values are rounded. The mixed-reward-group feature contributes only to the joint score.

| Feature or diagnostic | Mean $\mu_k$ | Scale $\sigma_k$ | Threshold $c$ |
|---|---|---|---|
| Rollout decline | 0.207077 | 0.276064 | 0.0 |
| Log-ratio mismatch | 0.055687 | 0.059835 | 4.0 |
| Learner-accuracy decline | 0.217567 | 0.286340 | 0.8 |
| Mixed-reward-group decline | 0.219499 | 0.262028 | — |
| Joint diagnostic | — | — | 1.0 |

Table 6: **First-alert accounting on held-out runs.** $N_E$ and $N_0$ are episode and no-episode denominators. Timely, early, late, and missing partition $N_E$. Tails are counts out of ten assigned runs per setting. Learner denotes current-learner evaluation.

| Setting | Diagnostic | $N_E$ | Timely | Early | Late | Miss | False/$N_0$ | Tail |
|---|---|---|---|---|---|---|---|---|
| $K = 8$ | Rollout | 6 | 5 | 0 | 1 | 0 | 1/3 | 1/10 |
| | Mismatch | 6 | 0 | 0 | 0 | 6 | 0/3 | 1/10 |
| | Learner | 6 | 6 | 0 | 0 | 0 | 1/3 | 1/10 |
| | Joint | 6 | 6 | 0 | 0 | 0 | 1/3 | 1/10 |
| $K = 16$ | Rollout | 8 | 2 | 0 | 6 | 0 | 0/2 | 0/10 |
| | Mismatch | 8 | 6 | 1 | 0 | 1 | 0/2 | 0/10 |
| | Learner | 8 | 7 | 1 | 0 | 0 | 0/2 | 0/10 |
| | Joint | 8 | 7 | 1 | 0 | 0 | 0/2 | 0/10 |
| $K = 32$ | Rollout | 7 | 0 | 1 | 6 | 0 | 0/2 | 1/10 |
| | Mismatch | 7 | 4 | 2 | 0 | 1 | 0/2 | 1/10 |
| | Learner | 7 | 3 | 4 | 0 | 0 | 0/2 | 1/10 |
| | Joint | 7 | 3 | 4 | 0 | 0 | 0/2 | 1/10 |
| Narrow queue | Rollout | 7 | 5 | 0 | 2 | 0 | 0/3 | 0/10 |
| | Mismatch | 7 | 2 | 0 | 0 | 5 | 0/3 | 0/10 |
| | Learner | 7 | 7 | 0 | 0 | 0 | 0/3 | 0/10 |
| | Joint | 7 | 6 | 0 | 1 | 0 | 0/3 | 0/10 |
| Long-age queue | Rollout | 7 | 6 | 0 | 1 | 0 | 0/3 | 0/10 |
| | Mismatch | 7 | 1 | 0 | 0 | 6 | 0/3 | 0/10 |
| | Learner | 7 | 5 | 2 | 0 | 0 | 0/3 | 0/10 |
| | Joint | 7 | 5 | 1 | 1 | 0 | 0/3 | 0/10 |

### C.3 Timing target and held-out accounting

Let $t_R$ be the earliest $t \geq 20$ starting 20 consecutive rewards below $0.25$. Eligible onsets satisfy $40 \leq t_R \leq 265$. With lead $L_R = t_R - t_A$, 1–16 updates is timely, more than 16 is early, and a nonpositive lead is late. No alert is the fourth category. An alert on a no-episode run is false. Unresolved tails are excluded from both episode and no-episode denominators.

The held-out cohort has ten runs per setting: $K = 8, 16, 32$, narrow queue, and long-age queue. It contains 35 eligible episodes, 13 no-episode runs, and two unresolved tails. Three episode runs recover, defined by mean rollout reward at least $0.25$ over the final 20 updates. Table 6 partitions all eligible episodes into the four first-alert categories.

The learner-evaluation rule gives 28 timely and seven early alerts. At $K = 32$, four early alerts lead onset by 21, 21, 21, and 29 updates. In the long-age queue, rollout reward gives 6/7 timely alerts versus 5/7 for learner evaluation. Pooled rates summarize this specified mixture of settings, with passive alerts that do not change training. Using the rounded parameters in Table 5, the learner rule corresponds to a decline of approximately $0.446639$ from its historical validation-accuracy maximum. It is a detector of substantial realized decline, consistent with 22/28 timely alerts occurring after measured learner accuracy falls below $0.25$.

# D Paired Analyses and Endpoint Sensitivity

## D.1 Resampling and the analysis unit

The four core contrasts use the same 20 complete paired blocks. The percentile bootstrap uses 20,000 whole-block resamples, NumPy PCG64 seed 2026092305, and linear quantiles at $0.025$ and $0.975$. After sorting block identifiers, a $20{,}000 \times 20$ index matrix is sampled with replacement. Each draw retains all core arms and checkpoints from the selected block. The same matrix serves the four core contrasts, factorial interaction, and within-run checkpoint comparisons. M2PO-style comparisons use the same block identities and seed 2026092307.

The monitoring comparison continues the core random-number stream with $20{,}000 \times 35$ eligible-run and $20{,}000 \times 13$ no-episode-run index matrices, retaining paired diagnostic outcomes. Its intervals condition on those eligibility groups. All intervals are pointwise summaries of run-level variation under the fixed evaluation items and protocol.

## D.2 Paired terminal states

Table 7 retains confirmed, unconfirmed-tail, and neither states. The primary binary contrast is confirmation by update 300, with tails retained in the denominator.

Table 7: **Paired endpoint states.** Each comparison contains 20 blocks. Neither means neither a confirmed failure nor an unconfirmed terminal candidate within the horizon. Columns give B's state. Reset denotes first-moment reset.

| Comparison A → B | A state | B: confirmed | B: tail | B: neither |
|---|---|---|---|---|
| Uncorrected → reset | Confirmed | 3 | 3 | 6 |
| | Tail | 0 | 0 | 2 |
| | Neither | 0 | 0 | 6 |
| Uncorrected → TIS | Confirmed | 2 | 3 | 7 |
| | Tail | 0 | 0 | 2 |
| | Neither | 0 | 0 | 6 |
| Reset → TIS | Confirmed | 2 | 0 | 1 |
| | Tail | 0 | 3 | 0 |
| | Neither | 0 | 0 | 14 |
| TIS → TIS + reset | Confirmed | 2 | 0 | 0 |
| | Tail | 0 | 3 | 0 |
| | Neither | 1 | 0 | 14 |

## D.3 Factorial interaction

Let $Y_{ij,b}$ indicate confirmed failure in block $b$, where $i$ indexes TIS and $j$ indexes first-moment reset. Define

$$I_b = Y_{11,b} - Y_{10,b} - Y_{01,b} + Y_{00,b}, \qquad I = \frac{1}{20}\sum_{b=1}^{20} I_b. \tag{D.1}$$

The estimate is $+50.0$ pp, paired SD $51.3$, and interval $[+30.0, +70.0]$. Reset changes failure incidence by $-45.0$ pp without TIS and $+5.0$ pp with TIS. At these margins, the additive-risk prediction is $p_{11} = p_{10} + p_{01} - p_{00} = -0.35$, outside the probability range. The interaction therefore quantifies less-than-additive risk reduction; it does not alone identify an antagonistic optimization mechanism.

## D.4 Current learner versus deployed snapshot

For each arm, $\delta_b = a_{300,b} - a_{288,b}$ compares a run's checkpoints on identical test items. Table 8 retains all 20 pairs, including failures and tails, for the nine non-M2PO arms. Every pointwise interval

includes zero. These terminal comparisons address a different quantity from within-run detection delay.

Table 8: **Current-minus-deployed test accuracy for the nine non-M2PO arms.** Differences are in pp. Each row has 20 pairs; SD is the sample SD of paired differences.

| Condition | Mean difference | SD | Pointwise 95% interval |
|---|---|---|---|
| Uncorrected | $-0.85$ | 4.81 | $[-3.19, +0.56]$ |
| First-moment reset | $-2.27$ | 8.05 | $[-6.22, +0.28]$ |
| Token-TIS | $-1.91$ | 6.75 | $[-5.25, +0.31]$ |
| TIS + first-moment reset | $+0.78$ | 14.46 | $[-4.48, +7.93]$ |
| Full optimizer reset | $-5.62$ | 17.36 | $[-14.17, +0.13]$ |
| Second-moment reset | $-4.42$ | 16.55 | $[-12.44, +0.27]$ |
| Counter reset only | $-2.13$ | 10.67 | $[-7.12, +0.60]$ |
| Reset direction, carried norm | $-0.50$ | 2.76 | $[-1.83, +0.51]$ |
| Carried direction, reset norm | $-4.49$ | 19.62 | $[-13.74, +0.62]$ |

### D.5 Threshold and terminal-tail sensitivity

The initial/persistence threshold pairs $(0.03, 0.10)$, $(0.05, 0.15)$, and $(0.10, 0.20)$ give identical confirmed-failure counts for every non-M2PO primary arm under the 20-update confirmation requirement. In the order uncorrected, first-moment reset, TIS, TIS plus reset, full reset, second-moment reset, counter reset, reset-direction control, and reset-norm control, the counts are $12, 3, 2, 3, 4, 10, 12, 3, 9$ out of 20. This concerns terminal failure; monitoring uses the separate $0.25$/20-update episode definition.

Counting confirmed failures and unresolved terminal candidates together gives 14/20 for uncorrected training, 6/20 for first-moment reset, 5/20 for TIS, 6/20 for TIS plus reset, and 5/20 for M2PO-style correction. These are sums of the disjoint C and T columns in Table 1, not additional observations beyond update 300. They show that the primary correction contrast is not explained solely by how terminal tails are classified.

### D.6 Paired monitoring comparison

Among 35 eligible held-out runs, learner evaluation and rollout reward give timely first alerts on 28 and 18 runs. Thirteen detections are learner-only and three are rollout-only, giving $+28.6$ pp with interval $[+8.6, +48.6]$. Both rules give their false alert on the same one of 13 no-episode runs. Their empirical paired-bootstrap false-alert difference is therefore $[0, 0]$, as there are no observed discordant pairs. TIS and M2PO-style correction likewise have identical confirmed-failure indicators in all 20 blocks. Such degenerate empirical intervals describe the observed pairs, not population equivalence.

## E Allocation and Coverage Across Settings

Table 9 separates 430 outcome runs (129,000 learner updates) from the disjoint 80-run monitoring cohort (24,000 updates). Thirty monitoring runs develop thresholds; 50 evaluate the frozen rules. The M2PO-style extension shares the primary controls, so its 20 additional runs are counted once. Appendix G reports a separate illustrative checkpoint record.

### E.1 Model and task coverage

Table 10 reports terminal failures and tails for a larger Qwen model, a second model family, and MATH. Each configuration has ten paired blocks per arm. MATH uses disjoint training and evaluation splits and a 4,096-token response cap. Comparisons are within configurations; accuracies from different tasks are not pooled.

Table 9: **Study allocation.** Every outcome and monitoring run has 300 learner updates.

| Cohort | Count | Allocation |
|---|---|---|
| Core interventions | 180 | Nine arms, 20 paired blocks, $K = 16$ |
| M2PO-style extension | 20 | One additional arm on the same 20 blocks |
| Synchronous reference | 40 | Two arms, 20 blocks; refresh every update, reset every 16 |
| Model/task coverage | 90 | Three configurations, three arms, ten blocks |
| Lag coverage | 40 | $K = 8, 32$; uncorrected/TIS; ten blocks |
| Mixed-age queues | 60 | Three profiles; uncorrected/TIS; ten blocks |
| Outcome subtotal | 430 | 129,000 learner updates |
| Monitor development | 30 | LLM training runs; threshold selection |
| Monitor evaluation | 50 | Held-out LLM runs; four fixed diagnostics |
| Monitoring subtotal | 80 | 24,000 learner updates |

Table 10: **Model and task coverage.** Entries are confirmed failures out of ten, with unconfirmed-tail counts in parentheses. Reset $m$ denotes first-moment reset.

| Configuration | Uncorrected | Token-TIS | Reset $m$ |
|---|---|---|---|
| Qwen2.5-7B / GSM8K | 6/10 (1) | 3/10 (1) | 5/10 (1) |
| Llama-3.2-3B / GSM8K | 7/10 (0) | 4/10 (0) | 5/10 (0) |
| Qwen2.5-1.5B / MATH | 7/10 (0) | 3/10 (0) | 6/10 (0) |

### E.2 Queue construction and lag coverage

Each queue run begins with 24 synchronized startup updates within the 300-update budget. From update 25, one behavior age is sampled per prompt group, and all eight responses use that snapshot. Age $A_{t,b}$ selects $\theta_{t-1-A_{t,b}}$.

The narrow profile puts equal mass at ages seven and eight. The uniform profile is uniform on ages zero through 15. The long-age profile puts mass $0.75$ at age two and $0.25$ at age 24. Their post-startup target means are all $7.5$, and their variances are $0.25$, $21.25$, and $90.75$. Zero-age mass is zero, $1/16$, and zero, respectively. Profiles use distinct blocks, with interventions paired within each profile.

The long-age queue has fewer observed confirmed failures than the other profiles (Table 11). Because the profiles also differ in age concentration and maximum age, this comparison concerns complete age distributions, not a single variance parameter. The synchronous reference refreshes behavior weights every update and retains a 16-update reset interval.

Table 11: **Additional outcome summaries.** Performance is the mean (sample SD), in percent, over all assigned runs. Final learner test accuracy and last sampled training-rollout reward are distinct endpoints.

| Setting | Condition | Confirmed | Tail | Learner | Last rollout |
|---|---|---|---|---|---|
| $K = 1$ | Uncorrected | 1/20 | 4/20 | 59.8 (36.6) | 58.1 (35.5) |
| | First-moment reset | 1/20 | 4/20 | 59.0 (36.1) | 57.2 (34.9) |
| $K = 8$ | Uncorrected | 5/10 | 0/10 | 38.2 (41.0) | 36.1 (39.7) |
| | Token-TIS | 1/10 | 0/10 | 70.5 (29.5) | 67.8 (30.2) |
| $K = 32$ | Uncorrected | 8/10 | 0/10 | 9.9 (24.3) | 14.2 (26.9) |
| | Token-TIS | 3/10 | 1/10 | 33.0 (40.3) | 43.0 (37.1) |
| Narrow queue | Uncorrected | 5/10 | 2/10 | 26.9 (39.7) | 26.2 (37.3) |
| | Token-TIS | 1/10 | 1/10 | 60.2 (39.2) | 65.0 (32.3) |
| Uniform queue | Uncorrected | 7/10 | 1/10 | 19.3 (36.3) | 18.9 (34.8) |
| | Token-TIS | 2/10 | 1/10 | 57.8 (39.6) | 57.1 (38.6) |
| Long-age queue | Uncorrected | 2/10 | 0/10 | 49.2 (41.8) | 51.5 (39.5) |
| | Token-TIS | 0/10 | 1/10 | 49.6 (40.6) | 54.4 (35.9) |

# F M2PO-STYLE ADAPTIVE-MASKING BASELINE

This arm adapts Zheng et al. (2026) to the common sequence-first training objective. It shares initialization, prompts, seed blocks, response budget, evaluation endpoints, binary rewards, reference regularizer, global gradient clipping, and optimizer with the primary comparison. The mask statistic pools valid tokens across the optimizer batch and distributed shards, using the published threshold $\tau_{M2} = 0.04$. TIS uses cap two.

## F.1 RATIO AND SELECTED TOKENS

For valid token $(b, i, \ell)$, retain its behavior log-probability and define

$$r_{bi\ell}(\theta) = \exp\big(\log \pi_\theta(y_{bi\ell} \mid h_{bi\ell}) - \mathrm{sg}[\log \pi_{\mathrm{beh}}(y_{bi\ell} \mid h_{bi\ell})]\big), \qquad d_{bi\ell} = \mathrm{sg}[\log r_{bi\ell}(\theta^-)]. \quad \text{(F.1)}$$

Here $\theta^-$ is the learner immediately before the update and sg stops gradients. Set $z_{bi\ell} = d_{bi\ell}^2$. The selected set is

$$\mathcal{T}_0 = \{(b, i, \ell) : (A_{bi} > 0 \wedge d_{bi\ell} > 0) \vee (A_{bi} < 0 \wedge d_{bi\ell} < 0)\}. \quad \text{(F.2)}$$

Zero advantages and unit ratios are excluded from this set.

## F.2 MASK CONSTRUCTION

Initialize $M = 1$ for all valid tokens and $\mathcal{T} = \mathcal{T}_0$. Sort selected tokens by decreasing $z$, breaking ties lexicographically by $(b, i, \ell)$. While $\mathcal{T}$ is nonempty and

$$\sum_{j \in \mathcal{T}} z_j > 0.04|\mathcal{T}|, \quad \text{(F.3)}$$

set the largest remaining token's mask to zero and remove it from $\mathcal{T}$. Recompute the remaining sum and count after removal, stopping at equality or below. Tokens outside $\mathcal{T}_0$ retain mask one. The constraint is on the *mean squared log ratio of remaining selected tokens*, not a pointwise cutoff or a bound on every importance ratio.

## F.3 OBJECTIVE AND INTERPRETATION

Minimize

$$\mathcal{L}_{M2}(\theta) = -\frac{1}{B}\sum_b \frac{1}{G}\sum_i \frac{1}{L_{bi}}\sum_{\ell=1}^{L_{bi}} \mathrm{sg}(M_{bi\ell}A_{bi}) r_{bi\ell}(\theta) + 10^{-3}\mathcal{R}_{\mathrm{ref}}(\theta). \quad \text{(F.4)}$$

The numerator of $r$ remains differentiable. At $\theta^-$, a retained contribution is proportional to $MAr\nabla \log \pi_{\theta^-}$. There is no additional TIS cap or proximal clipping. The regularizer is unmasked. Each $L_{bi}$ includes all valid tokens, including masked positions; masking changes the reduction's numerator only.

Published M2PO uses a token-mean objective. Our matched adaptation uses sequence-first reduction while retaining the token-pooled mask statistic and common reference regularizer. It also uses binary verifier rewards rather than the published reward shaping. These 20 runs share the primary blocks. Because the arm includes both behavior-ratio weighting and adaptive masking, its comparison with uncorrected training evaluates the combined correction rule, not masking in isolation.

# G ILLUSTRATIVE CHECKPOINT RECORD

A separate observed Qwen2.5-Instruct/GSM8K run with $K = 16$ illustrates delayed rollout feedback. It uses $G = 8$, $B = 64$, a 1,024-token response cap, learning rate $10^{-6}$, AdamW coefficients $(0.9, 0.999)$, global gradient clipping at one, and reference-KL coefficient $10^{-3}$.

Greedy learner accuracy falls from $0.61$ at update 112 to zero at update 128, while stale-policy rollout rewards remain between $0.58$ and $0.74$. Deployment at update 128 is followed by low rollout

reward at update 129. This record illustrates how learner deterioration can precede its appearance in generating-policy feedback.

The two channels use different evaluation samples and decoding protocols. The record is separate from the outcome and monitoring cohorts; it motivates the monitoring comparison rather than adding a replication to it. The 50-run held-out analysis instead uses the fixed validation panel and frozen rules in Appendix C.

# H Estimator Identities and Optimizer Memory

## H.1 Local expansion of the group sampling distribution

Let prompts follow a fixed distribution $\nu$. Conditional on $x$, draw independent responses $Y = (y_1, \dots, y_G)$ and write

$$P_\phi(Y \mid x) = \prod_{i=1}^{G} \pi_\phi(y_i \mid x), \qquad S_\phi(x, Y) = \sum_{i=1}^{G} \nabla_\phi \log \pi_\phi(y_i \mid x). \tag{H.1}$$

Fix the learner $\theta$ and all nonsampling inputs to $h_\theta(x, Y) \in \mathbb{R}^d$. Realized means and standard deviations remain functions of $Y$. Define

$$H_\theta(\phi) = \mathbb{E}_{x\sim\nu, Y\sim P_\phi}[h_\theta(x, Y)], \qquad u_G(\theta) = H_\theta(\theta), \qquad b_G(\theta, \phi) = H_\theta(\phi) - u_G(\theta). \tag{H.2}$$

The reference retains the normalization and gradient conventions of the off-policy estimator.

**Proposition 1, with explicit remainder.** Assume common support in a neighborhood $U$ of $\theta$, twice differentiable $P_\phi$, valid differentiation under the integral, and

$$\sup_{\phi\in U} \int \nu(dx) \int \|h_\theta(x, Y)\|_2 \, \|\nabla_\phi^2 P_\phi(Y \mid x)\|_{\mathrm{op}} \, dY \le C_\theta < \infty. \tag{H.3}$$

For $\Delta$ whose segment from $\theta$ to $\theta - \Delta$ lies in $U$,

$$b_G(\theta, \theta - \Delta) = -M_G(\theta)\Delta + R_\theta(\Delta), \qquad M_G = \mathbb{E}_{\nu, P_\theta}[h_\theta S_\theta^\top], \qquad \|R_\theta(\Delta)\|_2 \le \tfrac{1}{2} C_\theta \|\Delta\|_2^2. \tag{H.4}$$

*Proof.* The product rule gives $\nabla_\phi P_\phi = P_\phi S_\phi$. Therefore the Jacobian of $H_\theta$ at $\phi = \theta$ is $\mathbb{E}[h_\theta S_\theta^\top]$. Taylor expansion along $\theta - \Delta$ gives the negative first-order term. The integral remainder and Equation (H.3) give the stated norm bound. □

For Equation (5),

$$M_G = \frac{1}{G} \sum_{i,j=1}^{G} \mathbb{E}[A_i(Y) s_\theta(x, y_i) s_\theta(x, y_j)^\top]. \tag{H.5}$$

The $i \neq j$ terms retain the dependence introduced by shared group statistics. With finitely many prompts, a finite vocabulary and response cap, and smooth positive token probabilities, the local regularity conditions follow from finite sums. If normalization couples several groups, the sampling object and score must cover the entire coupled minibatch.

For a differentiable contribution that explicitly depends on the behavior policy, the Jacobian has both terms:

$$D_\phi \mathbb{E}_{\nu, P_\phi}[h_{\theta,\phi}] = \mathbb{E}_{\nu, P_\phi}[D_\phi h_{\theta,\phi} + h_{\theta,\phi} S_\phi^\top]. \tag{H.6}$$

Learner-side detachment does not remove the first term under behavior-policy differentiation.

### H.2 Proof of the response-IS residual-bias corollary

Write $h_\theta = G^{-1}\sum_i f_{\theta,i}$, and let $r_i(\theta,\phi) = \pi_\theta(y_i \mid x)/\pi_\phi(y_i \mid x)$. At $\phi = \theta$, $r_i = 1$ and $D_\phi r_i = -s_i^\top$. By Equation (H.6),

$$D_\phi \mathbb{E}_{\nu,P_\phi}[h^{\mathrm{IS}}_{\theta,\phi}]\big|_{\phi=\theta} = \frac{1}{G}\sum_i \mathbb{E}_{\nu,P_\theta}[f_{\theta,i}(S_\theta - s_i)^\top]$$
$$= \frac{1}{G}\sum_{i\neq j} \mathbb{E}_{\nu,P_\theta}[f_{\theta,i}s_j^\top] = C_G. \tag{H.7}$$

The response-weighted expectation equals $u_G$ at $\phi = \theta$. Taylor expansion therefore gives Equation (10), under a corresponding integrable second-derivative bound. Equivalently, first cancel the marginal density of response $i$ as in Equation (12); only its $G-1$ partners then change with $\phi$, producing the same cross term. Finite response spaces with positive smooth probabilities satisfy these conditions locally.

The result applies to both reductions. For sequence-first training, set $f_{\theta,i} = A_i s_i/L_i$, while the *sampling* score remains the full sequence score $s_i$. Response length is part of the sampled contribution, not a change to the sampling-law derivative. The verification script independently checks $M_G$ and $C_G$ against numerical Jacobians of exact finite sums and confirms quadratic remainders for both expansions.

**A response-independent baseline as a contrast.** Replace $A_i(Y)$ by $a_\theta(x,y_i) = R(x,y_i) - c_\theta(x)$, with fixed reward $R$ and a response-independent baseline held fixed in the expansion. For the sequence-summed contribution,

$$u_G = \nabla_\theta \mathbb{E}_{\nu,\pi_\theta} R, \qquad M_G = \mathbb{E}[a_\theta s_\theta s_\theta^\top], \qquad C_G = 0. \tag{H.8}$$

Independence and the zero conditional score mean eliminate cross-response terms. Group-standardized advantages use a different on-policy reference and retain those terms. The general expansion requires neither symmetry nor positive-semidefiniteness of $M_G$.

### H.3 Correction at token, response, and group levels

Assume the behavior policy has positive probability wherever the target contribution has support. The joint weight satisfies

$$W_G(x,Y) = \prod_{i=1}^{G} \frac{\pi_\theta(y_i \mid x)}{\pi_\phi(y_i \mid x)}, \qquad \mathbb{E}_{\nu,P_\phi}[W_G h_\theta] = u_G(\theta), \tag{H.9}$$

whenever the expectation is integrable. Cancelling the behavior density proves the identity. This reweights a fixed gradient contribution and does not require learner-side differentiation of $W_G$.

Response IS cancels only that response's marginal density, yielding Equation (12); its advantage partners remain behavior-sampled. A token ratio changes the action law at a sampled prefix, while the prefix, continuation, and group partners retain their behavior-conditioned distributions. Thus the correction levels have different sampling targets.

For vector contribution $f$ and nonnegative weight $w$, truncation at $c$ obeys

$$\|\mathbb{E}[(w - \min(w,c))f]\|_2 \le \mathbb{E}[(w-c)_+\|f\|_2]. \tag{H.10}$$

The triangle inequality proves the bound. If $\|f\|_2 \le H$, it is at most $H\mathbb{E}[(w-c)_+]$. It bounds the truncation-induced mean shift; the finite-state calculations measure the accompanying covariance change separately.

### H.4 Carried first moments

Unrolling Equation (A.6) after deployment at $\tau$ gives

$$m_{\tau+j} = \beta_1^j m_\tau + (1-\beta_1)\sum_{\ell=1}^{j} \beta_1^{j-\ell} g_{\tau+\ell}. \tag{H.11}$$

A first-moment reset removes the inherited term while retaining the second moment and counter. Shadow-state controls compare candidate updates at common instantaneous parameters using the same clipped gradient. With zero weight decay, Euclidean norm matching covers the full applied update.

# I Reproducible Exact Estimator Calculations

All calculations in this appendix enumerate response groups at fixed learner and behavior distributions. Means, covariances, bias vectors, weight diagnostics, and batch errors are exact finite sums. These mathematical checks are separate from the LLM training experiments.

## I.1 Binary check of local bias and group correction

Each response is $y_i \in \{0, 1\}$ with reward $R(y_i) = y_i$. Set $G = 8$, behavior success probability $p_\phi = \sigma(\phi)$, and target logit $\theta = -0.7$. The score is $s_\theta(y_i) = y_i - p_\theta$. Advantages subtract the realized group mean and divide by the standard deviation with variance divisor $G$. This calculation uses no stabilizer and assigns zero advantages to homogeneous groups.

For $h_\theta = G^{-1} \sum_i A_i s_\theta(y_i)$, enumeration of all $2^8 = 256$ groups yields

$$u_8(\theta) = 0.428044782926, \qquad M_8(\theta) = 0.099111558331. \tag{I.1}$$

An independent expression, with $k = \sum_i y_i$, is

$$h_\theta(Y) = \frac{\sqrt{k(8-k)}}{8}, \qquad k \sim \text{Binomial}(8, p_\phi). \tag{I.2}$$

Ordered-group and count-based expectations agree to floating-point precision. Table 12 checks the quadratic remainder, and Table 13 compares response and group correction. At $\theta = 0$, the group-estimator mean is $0.463913987402$, whereas the expected-reward gradient is $0.25$, directly distinguishing the two references.

Table 12: **Exact binary-group check of the local expansion.** The signed remainder is $r(\delta) = b_8(\theta, \theta - \delta) + M_8(\theta)\delta$. Halving order is $\log_2 |r(2\delta)/r(\delta)|$.

| $\delta$ | $r(\delta)$ | $r(\delta)/\delta^2$ | Halving order |
|---|---|---|---|
| 0.10000 | $-5.948006 \times 10^{-4}$ | $-0.059480$ | — |
| 0.05000 | $-1.506064 \times 10^{-4}$ | $-0.060243$ | 1.981624 |
| 0.02500 | $-3.788344 \times 10^{-5}$ | $-0.060614$ | 1.991143 |
| 0.01250 | $-9.499439 \times 10^{-6}$ | $-0.060796$ | 1.995653 |
| 0.00625 | $-2.378407 \times 10^{-6}$ | $-0.060887$ | 1.997847 |

Table 13: **Exact expectations at target logit $\theta = -0.7$.** The on-policy group-estimator mean is $u_8 = 0.428044783$.

| Behavior $\phi$ | Unweighted | Response IS | Joint-group IS |
|---|---|---|---|
| $-1.3$ | 0.350403675 | 0.461614744 | 0.428044783 |
| $-0.2$ | 0.460899292 | 0.412850361 | 0.428044783 |
| $+0.4$ | 0.451945907 | 0.417056167 | 0.428044783 |

## I.2 Four-action design and estimators

The learner is a four-action softmax with logits $\log p$, $p = (.1, .2, .3, .4)$, rewards $(0, 0, 0, 1)$, and $G = 4$. One-token responses make token and response weighting identical. We enumerate all $4^4 = 256$ ordered groups using population-variance normalization and zero homogeneous-group

contributions. With $s(y) = e_y - p$,

$$h(Y) = \frac{1}{4}\sum_{i=1}^{4} A_i(Y)s(y_i), \qquad u = \mathbb{E}_{p^4}h = (-.0648266568, -.1296533136, -.1944799704, .3889599408). \tag{I.3}$$

The nine behavior distributions are $p$ and $q = (1-\lambda)p + \lambda q^*$ for $\lambda \in \{.25, .50, .75, 1\}$, with

$$q^*_{\text{reverse}} = (.4, .3, .2, .1), \qquad q^*_{\text{rare}} = (.4, .35, .24, .01). \tag{I.4}$$

Writing $r(y) = p_y/q_y$, define

$$h_w(Y) = \frac{1}{4}\sum_i A_i(Y)w(y_i)s(y_i), \quad w \in \{1, r, \min(r, c), r\mathbf{1}\{r \le 2\}\}, \quad c \in \{1, 2, 4, 8\}. \tag{I.5}$$

The fixed-threshold masked IS (MIS) rule is distinct from adaptive M2PO-style training. Joint-group IS uses $X = W_G h$. An oracle control-variate estimator is

$$X_{\text{CV}} = X - A^*C, \qquad C = (n_1 - 4q_1, n_2 - 4q_2, n_3 - 4q_3)^\top, \qquad A^* = \text{Cov}(X, C)\,\text{Cov}(C)^{-1}, \tag{I.6}$$

where $n_j$ counts action $j$. Since $\mathbb{E}_q C = 0$, the control preserves $\mathbb{E}X = u$. Positive category masses make

$$\text{Cov}(C) = 4\{\text{diag}(q_{1:3}) - q_{1:3}q_{1:3}^\top\} \tag{I.7}$$

invertible. Completing the square in $\mathbb{E}\|X - \mathbb{E}X - AC\|_2^2$ gives the covariance-trace-minimizing coefficient among linear controls using $C$. The oracle uses exact population moments.

### I.3 Why response weighting aligns the expected direction

**Proposition 2** (One rewarding category)**.** *Consider strictly positive categorical distributions $p, q$, softmax-logit score $s_p(y) = e_y - p$, group size $G \ge 2$, and one rewarding action $a$. Use group-standardized binary advantages with zero homogeneous-group contributions. Define*

$$\alpha_+(q_a) = \mathbb{E}_{q^G}[A_1 \mid y_1 = a] > 0, \qquad \alpha_-(q_a) = \mathbb{E}_{q^G}[A_1 \mid y_1 \ne a] < 0. \tag{I.8}$$

*Then*

$$\mathbb{E}_q h = q_a(\alpha_+ - \alpha_-)(e_a - q), \qquad \mathbb{E}_q h_r = p_a(\alpha_+ - \alpha_-)(e_a - p). \tag{I.9}$$

*For cap $c > 0$, suppose all nonreward-action ratios are untruncated. With $b_a = \min(p_a, cq_a)$,*

$$\mathbb{E}_q h_{\min(r,c)} = (b_a\alpha_+ - p_a\alpha_-)(e_a - p). \tag{I.10}$$

*Both weighted means are positive multiples of the on-policy group-estimator direction in logit coordinates; their magnitudes can differ from its on-policy magnitude.*

*Proof.* Conditional on an action's reward class, the remaining reward counts determine its expected advantage. Nonreward-action identities do not change that expectation. Exchangeability and $\sum_i A_i = 0$ imply $q_a\alpha_+ + (1-q_a)\alpha_- = 0$. Taking the expectation of one summand gives the unweighted expression. Response weighting replaces $q_j$ by $p_j$ in that summand. The identity $\sum_{j\ne a} p_j(e_j - p) = -p_a(e_a - p)$ yields the response-IS expression. When only the rewarding action can be capped, replace its coefficient by $b_a$ to obtain Equation (I.10). The scalar is positive because $\alpha_+ > 0$ and $\alpha_- < 0$. □

The cap-two condition holds throughout the behavior grid. At the reverse and rare-reward endpoints, unweighted cosines with $u$ are $0.9438798$ and $0.9551176$; response IS and cap-two TIS have cosine one. Positive scalar rescaling preserves these angles. The result concerns this categorical logit model and separates direction from magnitude and covariance.

Table 14: **Exact vector MSE for $B = 64$ independent groups.** Nine estimators are evaluated at nine distinct behavior distributions. The on-policy column is repeated for readability. The control variate uses exact population coefficients.

| Reverse endpoint family | | | | | |
|---|---|---|---|---|---|
| Estimator | $q = p$ | $\lambda = .25$ | $\lambda = .50$ | $\lambda = .75$ | $\lambda = 1$ |
| Unweighted | 0.00152 | 0.00582 | 0.01864 | 0.04354 | 0.08934 |
| Response IS | 0.00152 | 0.00156 | 0.00209 | 0.00347 | 0.00740 |
| TIS, cap 1 | 0.00152 | 0.00531 | 0.01968 | 0.04886 | 0.09855 |
| TIS, cap 2 | 0.00152 | 0.00156 | 0.00209 | 0.00402 | 0.04043 |
| TIS, cap 4 | 0.00152 | 0.00156 | 0.00209 | 0.00347 | 0.00740 |
| TIS, cap 8 | 0.00152 | 0.00156 | 0.00209 | 0.00347 | 0.00740 |
| Response MIS, mask 2 | 0.00152 | 0.00156 | 0.00209 | 0.14174 | 0.16881 |
| Joint-group IS | 0.00152 | 0.00306 | 0.00762 | 0.02206 | 0.10236 |
| Joint IS + score CV | 0.00080 | 0.00123 | 0.00337 | 0.01310 | 0.08175 |
| **Rare-reward endpoint family** | | | | | |
| Estimator | $q = p$ | $\lambda = .25$ | $\lambda = .50$ | $\lambda = .75$ | $\lambda = 1$ |
| Unweighted | 0.00152 | 0.00702 | 0.02801 | 0.08011 | 0.19331 |
| Response IS | 0.00152 | 0.00165 | 0.00273 | 0.00671 | 0.09189 |
| TIS, cap 1 | 0.00152 | 0.00751 | 0.03256 | 0.08919 | 0.19548 |
| TIS, cap 2 | 0.00152 | 0.00165 | 0.00273 | 0.03401 | 0.18465 |
| TIS, cap 4 | 0.00152 | 0.00165 | 0.00273 | 0.00671 | 0.16523 |
| TIS, cap 8 | 0.00152 | 0.00165 | 0.00273 | 0.00671 | 0.13123 |
| Response MIS, mask 2 | 0.00152 | 0.00165 | 0.00273 | 0.16593 | 0.20569 |
| Joint-group IS | 0.00152 | 0.00351 | 0.01134 | 0.06430 | 52.25541 |
| Joint IS + score CV | 0.00080 | 0.00134 | 0.00530 | 0.04674 | 52.00027 |

### I.4 Residual bias, truncation, and finite-batch error

For mean $\mu_c$, covariance $\Sigma_c$, and average $\overline{X}_B$ of $B$ independent groups,

$$\mathbb{E}\|\overline{X}_B - u\|_2^2 = \|\mu_c - u\|_2^2 + \frac{\operatorname{tr}\Sigma_c}{B}. \tag{I.11}$$

The cross term vanishes because $\mathbb{E}[\overline{X}_B - \mu_c] = 0$, and independence gives $\operatorname{Cov}(\overline{X}_B) = \Sigma_c/B$. The mean error decomposes as

$$\mu_c - u = (\mu_\infty - u) + (\mu_c - \mu_\infty). \tag{I.12}$$

The first term is residual error after response-only weighting; the second is the truncation shift. Their inner product determines whether they reinforce or offset each other.

Let $\Delta V = \operatorname{tr}\Sigma_\infty - \operatorname{tr}\Sigma_2$ and $\Delta b^2 = \|\mu_2 - u\|_2^2 - \|\mu_\infty - u\|_2^2$. When both are positive, the crossover is $B^* = \Delta V/\Delta b^2$, and cap-two TIS has lower MSE exactly for $B < B^*$. The reverse and rare-reward endpoints give $B^* = 8.08145807$ and $31.63088687$.

At the rare-reward endpoint with $B = 16$, cap-two TIS has MSE $0.18565488$ versus $0.36482893$ for response IS. At $B = 64$, their values are $0.18465181$ and $0.09189148$. Table 14 gives the complete grid. Cap two improves on the unweighted estimator at all eight off-policy points for $B = 64$, whereas exact joint-group correction can incur a large variance cost.

**Deterministic verification.** The accompanying `verification/verify_exact.py` checks ordered-group enumeration against an independent implementation over the 35 multinomial count states. Means, covariance matrices, and oracle coefficients agree to floating-point precision. It checks the binary identities, local and response-IS residual-bias expansions, joint-group correction, zero-mean control variates, direction alignment, and all 90 displayed MSE entries (81 distinct estimator/distribution pairs). It writes full-precision numerical results and regenerates Figure 2. The companion aggregate checker verifies reported counts, margins, and arithmetic without substituting for run-level experimental logs.