WHAT OFF-POLICY CORRECTIONS KEEP: GRADIENT DIRECTION AND REALIZED GAINS FROM STALE LLM ROLLOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model reinforcement learning increasingly learns from stale rollouts, produced by asynchronous systems or by long agentic interactions that are expensive to repeat, and continued training on them can erode earlier gains. We ask what learning signal different off-policy corrections extract from the same stale responses and when that signal yields realized improvement. We develop a theoretical framework for learning from stale data. It represents each correction through per-token coefficients that factor into advantage, weight, and gate; separates gradient magnitude from off-policy gradient alignment (OGA); and derives a local condition linking alignment and step size to realized improvement. Across mathematical reasoning and instruction following, correction rules produce markedly different directions even at low policy drift. Rules with identical ratio bounds reverse their majority OGA ordering across tasks, and near-equal scalar coefficient-mass totals can accompany large directional differences. OGA generally deteriorates as the policy moves away from the sampling policy, motivating drift as an observable state variable for reuse control. In held-out retrospective evaluation, we show that outcome-calibrated, drift-aware control of reuse retains gains that continued training would otherwise erase, without requiring fresh-gradient measurements.

1 INTRODUCTION

Reinforcement learning (RL) for language models increasingly learns from responses that the current policy did not generate. Asynchronous RL systems overlap generation with optimization, so each update consumes responses sampled by earlier policies (Noukhovitch et al., 2024; Fu et al., 2025; Huang et al., 2026). Agentic training adds a second reason: a single rollout that calls tools and interacts with an environment can take tens of minutes to hours (Sheng et al., 2025; Zhang et al., 2026c; Tongyi DeepResearch Team, 2025; Sirovatka & Prime Intellect Team, 2026), which makes fresh samples expensive and makes it attractive to learn repeatedly from a stored pool, as in offline RL. In both settings the question is how to make good use of off-policy rollouts.

Off-policy corrections aim to recover the on-policy gradient from such rollouts, and they fall into two families. *Importance weighting* reweights tokens or sequences by the ratio between the current and the sampling policy, usually with a cap to control variance (MiniMax, 2025; Zheng et al., 2025a; Pang et al., 2025). *Clipping and masking* drops the gradient of tokens whose ratio has moved too far from one (Schulman et al., 2017; Yu et al., 2025; Zheng et al., 2025b). Both act on an advantage that the learner estimates from the same stale responses, by comparing rewards within a prompt group (Shao et al., 2024) or with a value model (Schulman et al., 2016).

We ask how these two ingredients, the correction rule and the advantage estimated from off-policy data, shape training at two linked levels: what direction do stale responses supply at the current policy, and which checkpoint along the resulting training path retains the observed task gains?

We answer with a theoretical framework and experiments organized by it (Figure 1). Differentiating each correction's surrogate down to the log probabilities of the stored tokens turns every tested rule into a weighted sum of the same score vectors, with per-token coefficients that factor into normalization, advantage, weight and gate. We split the resulting gradient into a relative magnitude,

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.

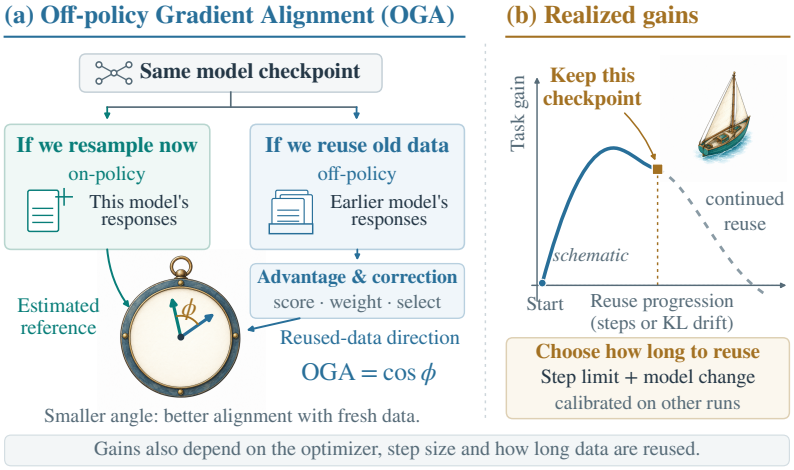

Figure 1: **Two questions for learning from stale rollouts. (a) Direction alignment.** At the same fixed checkpoint, we ask how learning would differ if we resampled responses now or reused responses from an earlier policy. Advantages and correction rules shape the gradient extracted from stored responses. Off-policy gradient alignment (OGA) measures its angular agreement with the estimated on-policy reference. **(b) Realized gains.** Retaining improvement also depends on the optimizer, step size, and how long data are reused: later updates can erase earlier gains.

which adaptive optimizers such as Adam largely remove, and a direction, which we measure as the cosine with a gradient from fresh samples at the same policy state. A pathwise gain identity and a local improvement proposition connect, but do not collapse, two questions: direction and step size condition the progress of one update, whereas delivered gain accumulates along the optimizer path and later reuse can cancel earlier progress. We therefore study two linked layers—the directions that corrections extract at common states, and outcome-calibrated control of how long to reuse the pool. Drift is an observable state signal for the latter, not a direct surrogate for either alignment or reward. We make three contributions.

1. **A theoretical framework for stale-data gradients.** A per-token coefficient decomposition places importance weighting, clipping, masking and sequence-level rules in one form at a common policy state and separates the two roles of the advantage: it scales each coefficient and, through its sign, selects clipping branches. A magnitude–direction decomposition identifies the direction as the quantity that adaptive optimization preserves, and a pathwise gain identity with a local improvement proposition states when a direction yields gain and what limits it.
2. **Findings about the directions stale data supply.** Across eight rules and two tasks, corrections extract markedly different directions from the same states, already at low drift, and near-equal scalar coefficient-mass totals can accompany large alignment shifts. Two rules with identical ratio bounds, which differ only in tails that the advantage sign selects, reverse their ordering across tasks. Alignment declines with drift for most rules.
3. **Gain retention through calibrated control of reuse.** We evaluate an outcome-calibrated step budget, a KL state signal and their combination by the models they deliver on held-out families. In instruction following, adding the KL override to the budget raises delivered gain from $+2.15$ to $+2.66$ points while continued training loses 17.8; in mathematics, conservative calibration retains the budget's $+2.25$ pp result rather than forcing an override.

Relation to prior work. PPO-style clipping (Schulman et al., 2017), group-relative objectives (Shao et al., 2024; Yu et al., 2025), detached ratio caps (MiniMax, 2025), sequence-level corrections (Zheng et al., 2025a; Pang et al., 2025) and second-moment masking (Zheng et al., 2025b) define different local operators on old samples; classical off-policy actor–critic, V-trace and Retrace analyze change of measure and truncated importance weights (Degris et al., 2012; Espeholt et al., 2018; Munos et al., 2016). We compare the tested operators through their per-token coefficients at common policy states, using gradient alignment, an established diagnostic (Ilyas et al., 2020; En-

gstrom et al., 2020), to connect advantage construction and correction branches to the direction of stale-data gradients. Asynchronous systems and correction studies motivate managing policy lag (Noukhovitch et al., 2024; Fu et al., 2025; Huang et al., 2026; Zheng et al., 2025b); our stopping study tests whether a rule calibrated on other trajectories preserves gains on a held-out method family. Appendix A expands these threads and distinguishes the reference and control protocols used by the closest related work.

2 A Framework for Learning Signals from Stale Data

2.1 Policy gradients and off-policy corrections

A policy π_θ answers a prompt x with a response $y = (y_1, \dots, y_T)$ and receives a reward $R(x, y)$. Reinforcement learning maximizes $J(\theta) = \mathbb{E}_x \mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[R(x, y)]$. Its standard gradient is $\nabla J(\theta) = \mathbb{E}[\sum_t A_t \nabla_\theta \log \pi_\theta(y_t \mid x, y_{<t})]$ (Williams, 1992; Sutton et al., 2000). The update raises the probability of tokens in responses that did better than expected and lowers it for responses that did worse; the advantage A_t measures "better than expected". Group-relative methods sample several responses per prompt and give every token of response i the same advantage $A_{it} = R_i - \bar{R}_{g(i)}$, its reward minus the prompt-group mean (Shao et al., 2024). A value model V instead supplies token-dependent advantages by generalized advantage estimation (GAE), $A_t = \sum_{l\geq 0}(\gamma\lambda)^l \delta_{t+l}$ with $\delta_t = r_t + \gamma V_{t+1} - V_t$ (Schulman et al., 2016).

This gradient requires responses from the *current* policy, whereas our setting generates a fixed pool $\mathcal{D}_\mu$ once with $\mu = \pi_{\theta_0}$ and consumes it in ordered minibatches while θ_k moves away from θ_0. Importance sampling corrects the mismatch in principle, $\mathbb{E}_{\pi_\theta}[g(y)] = \mathbb{E}_\mu[g(y)\,\pi_\theta(y)/\mu(y)]$, but the sequence ratio is a product of token ratios whose variance grows quickly with length and drift. Practical corrections therefore work with token ratios $r_z = \pi_\theta(y_z \mid x, y_{<z})/\mu(y_z \mid x, y_{<z})$ or length-normalized sequence ratios, in one of two ways: they *reweight* tokens by a possibly capped ratio, or they *clip or mask* tokens whose ratio leaves a trust band. Each rule defines a surrogate loss $L(\theta)$ on the old responses, and training follows $-\nabla L$.

2.2 From a surrogate to per-token coefficients

Every tested surrogate depends on θ only through the log probabilities $\ell_z = \log \pi_\theta(y_z \mid x, y_{<z})$ of stored tokens z: rewards, advantages and μ are fixed (advantages are detached), and $r_z = \exp(\ell_z - \ell_z^\mu)$. The chain rule therefore writes the gradient of any surrogate on a batch $\mathcal{B}$ as

$$b \equiv -\nabla_\theta L = \sum_{z\in\mathcal{B}} \frac{\partial(-L)}{\partial \ell_z} \nabla_\theta \ell_z = \sum_{z\in\mathcal{B}} c_z\, q_z, \qquad q_z = \nabla_\theta \log \pi_\theta(y_z \mid x, y_{<z}). \tag{1}$$

The score vector q_z is the parameter direction that makes token z more likely, and the scalar c_z states how strongly, and with which sign, the rule pushes along it. Two examples show how c_z arises. For the importance-weighted token objective $-L = M^{-1}\sum_z A_z r_z$, with M valid tokens, $\partial r_z/\partial \ell_z = r_z$ gives $c_z = A_z r_z/M$. PPO-style clipping uses $\min\{A_z r_z,\ A_z \operatorname{clip}(r_z, l, u)\}$ instead; the clipped branch is constant in θ, so $c_z = A_z r_z/M$ where the unclipped branch is active and $c_z = 0$ elsewhere. The same calculation for every tested rule (Appendix C.1) yields four factors,

$$c_z = \nu_z\, A_z\, w_z\, m_z. \tag{2}$$

Read from right to left: the gate $m_z \in \{0, 1\}$ decides whether token z contributes at all, the weight w_z (one, a ratio, or a capped detached ratio) rescales what remains, the detached advantage A_z supplies sign and size, and ν_z collects length and batch normalization. For sequence-level rules we differentiate the full sequence objective before reading off the factors.

Two consequences organize the analysis. First, rules applied to the same old batch share every q_z and differ only in their coefficients, so the difference between two rules is a signed sum over the tokens on which their factors disagree. Second, the advantage enters twice: as a factor, and through gates that depend on its sign. If the advantages change from A_f to A_o and the gates from m_f to m_o while ν, w and q stay fixed,

$$c_o - c_f = \nu w\big[(A_o - A_f)\, m_o + A_f\, (m_o - m_f)\big]. \tag{3}$$

The first term changes the signal on an unchanged gate; the second changes which tokens the gates select. A baseline change always moves the first term and, if it flips an advantage sign, can also move the second. We therefore analyze advantage estimation and correction together (Appendix C).

2.3 Relative magnitude and direction

To judge a stale gradient we compare it with a reference at the same policy state. We sample new responses from π_{θ_k}, compute group-relative advantages, and form the gradient of the same objective with the declared reduction; this fresh-sample estimate $\hat{f}$ approximates $f = \nabla J(\theta_k)$. Projecting b on $\hat{f}$ separates two quantities:

$$P(b) = \frac{\langle b, \hat{f}\rangle}{\|\hat{f}\|^2} = \underbrace{\frac{\|b\|}{\|\hat{f}\|}}_{\rho(b)} \cdot \underbrace{\frac{\langle b, \hat{f}\rangle}{\|b\|\,\|\hat{f}\|}}_{\mathrm{OGA}(b)}, \qquad b \neq 0,\ \hat{f} \neq 0. \tag{4}$$

$P(b)$ is the component of the stale gradient along the fresh one, in units of the fresh gradient. The relative magnitude ρ measures size, and the off-policy gradient alignment $\mathrm{OGA} \in [-1, 1]$ measures direction: 1 is the fresh direction, 0 is orthogonal, and negative values oppose it.

Which of the two reaches the parameters depends on the optimizer. Under plain gradient descent both do, but our runs, like most LLM RL, use AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) with global gradient-norm clipping. Adam divides each coordinate of a running gradient mean by a running root-mean-square estimate, attenuating common rescalings, while global clipping also preserves a gradient's direction. Coordinatewise histories can still rotate or shrink different rule gradients, so neither raw magnitude nor OGA is an executed-step statistic. We use OGA as the primary raw-direction diagnostic and ρ as a scale diagnostic; Appendix B.2 states the conditions and remaining optimizer effects.

Because P is linear in b, a rule difference decomposes over tokens:

$$P(b_2) - P(b_1) = \sum_z \left(c_z^{(2)} - c_z^{(1)}\right) \frac{\langle q_z, \hat{f}\rangle}{\|\hat{f}\|^2}. \tag{5}$$

Each changed coefficient counts in proportion to how its score vector aligns with the fresh gradient: removing contributions that oppose $\hat{f}$ raises the projection, and removing contributions along $\hat{f}$ lowers it. Scalar summaries such as the gated fraction or the retained coefficient mass see only the c_z, so small coefficient changes can move the direction substantially when they hit tokens with large $|\langle q_z, \hat{f}\rangle|$ (Section 3.2). Cosines are not additive, so we report OGA and use (5) to locate its changes.

2.4 From direction to realized gain

Training applies an optimizer to minibatch gradients along a path. Let $\mathcal{B}_k$ be the kth training minibatch, $\mathcal{U}_k$ the optimizer map with its state, clipping and weight decay, $u_k = \mathcal{U}_k(\sum_{z\in\mathcal{B}_k} c_z q_z)$ the executed update, and $f_k = \nabla J(\theta_k)$. Expanding each step to second order and summing gives

$$\Delta J_K = J(\theta_K) - J(\theta_0) = \sum_{k<K} \Big[\langle f_k,\, u_k\rangle + \mathcal{R}_k\Big], \qquad |\mathcal{R}_k| \le \tfrac{1}{2} L_k \|u_k\|^2, \tag{6}$$

where L_k bounds the Hessian norm on the kth segment. The realized gain is the accumulated first-order progress of the executed steps plus curvature corrections. The identity names three places where stale data can fail to deliver gain: the coefficients can point away from the objective gradient, the optimizer map transforms them, and continuing to step K adds later terms that can cancel earlier progress. This structure separates two empirical questions that the rest of the paper treats in turn: a *direction* question, which correction extracts a well-aligned gradient from a shared state (Section 3), and a *delivery* question, how long to reuse the pool before earlier gains are cancelled (Section 4). The optimizer map links the two but we do not measure the executed AdamW update directly, so we treat direction and delivery as related but distinct rather than derive one from the other. For a single step, the fresh reference gives an operational condition.

Proposition 1 (Reliable direction with a controlled step)**.** *Let $f = \nabla J(\theta)$ with $\|f - \hat{f}\| \le \epsilon$ and $\hat{f} \neq 0$, and let ∇J be L-Lipschitz on the segment from θ to $\theta + u$. For any update $u \neq 0$,*

$$J(\theta + u) - J(\theta) \;\ge\; \|u\|\big(\|\hat{f}\|\cos(u, \hat{f}) - \epsilon\big) - \tfrac{L}{2}\|u\|^2, \tag{7}$$

which is positive if $\|\hat{f}\|\cos(u, \hat{f}) > \epsilon$ and $\|u\| < 2(\|\hat{f}\|\cos(u, \hat{f}) - \epsilon)/L$. For a step along the stale gradient, $u = \alpha b$ with $\alpha > 0$, $\cos(u, \hat{f}) = \mathrm{OGA}(b)$.

The first term is progress: the step length times the part of the fresh gradient the step points along, less what the reference error can hide; the second is a curvature cost that grows quadratically with the step. A step therefore helps when its direction clears the reference error and the step is short enough (proof in Appendix B). In this one-step bound the update enters through its direction and length. Across a trajectory, reuse length determines how many such conditional steps accumulate, but the proposition neither assigns utility to a checkpoint nor defines an alignment threshold for stopping. We use it as a guide without estimating L or ϵ; for AdamW, $\cos(u, \hat{f})$ refers to the preconditioned update.

3 The Directions Stale Rollouts Supply

Tasks, data and training. Mathematical reasoning uses Qwen3-1.7B-Base (Qwen Team, 2025) with a binary correctness reward from a math verifier. Its response pool comes from DeepScaleR (Luo et al., 2025): the initial model answers each problem eight times, and we keep the 17,085 problems whose answers are neither all correct nor all wrong. Instruction following uses the non-Base Qwen3-1.7B on 8,192 writing prompts that we generate with rule templates following IFEval (Zhou et al., 2023), each combining a word-count range with 4–10 hard constraints of ten IFEval types and answered 16 times; the reward multiplies the fraction of satisfied constraints by a length factor. Training consumes each pool in order without new samples, using AdamW at two constant learning rates, 5×10^{-6} and 2.5×10^{-6}, and seed 42. Scores use held-out prompts disjoint from the training pools: 1,560 problems from the DeepScaleR evaluation split, MATH-500, OlympiadBench, Minerva Math, AMC 2023 and AIME 2024/2025, and 800 instruction prompts from the same generator (Appendix F).

Correction rules. We compare eight rules, ordered by the factor of (2) they act on. *No-IS* ($w = m = 1$) is the uncorrected baseline. CISPO multiplies each token by a detached ratio capped at 1.2 and never closes a gate. Two signed clips follow: *Asym-Clip* uses DAPO-style bounds $[0.8, 1.28]$ and *Sym-Clip* uses $[0.8, 1.2]$; their gates depend on the advantage sign. Two masks close gates regardless of sign: *TSM* (two-sided masking) removes tokens with ratios outside $[0.8, 1.2]$, and M2PO removes tokens selected by a second-moment criterion. GSPO and TIC-GRPO act on sequence ratios, a geometric mean with a narrow clipping band and an unnormalized product with a live cap. Figures and tables use this order; Appendix C.1 lists the factors.

3.1 Correction rules at shared policy states

The advantage marks which tokens a step should make more or less likely, and the rule decides which of these contributions it keeps, reweights or closes; for the signed clips the two interact, because their gates depend on the advantage sign (3). Figure 2 compares rules on shared states: each comparison fixes the checkpoint, old batch, group-relative advantages and fresh reference and changes only the rule. Instruction following provides 100 such states for seven rules and mathematics 154 states for all eight; we evaluate instruction-following Sym-Clip on 39 of these states. Appendix E repeats the stale-side comparison with a critic's GAE on three critic-trained mathematical hosts, against the same group-relative fresh-reference contract.

Two patterns hold in both tasks. No-IS and CISPO align best, with medians of 0.66 and 0.68 in instruction following and 0.53 and 0.52 in mathematics. The signed clips align worst: in mathematics their medians, like TSM's, are 0.05–0.06, and in instruction following Asym-Clip is negative at 65 of the 100 states (median -0.24), while Sym-Clip matches its sign on all 39 states where we evaluate both (medians -0.27 and -0.28). The other rules fall between and differ by task; TIC-GRPO is close to No-IS in mathematics (0.49), whereas M2PO has 36 negative instruction-following states

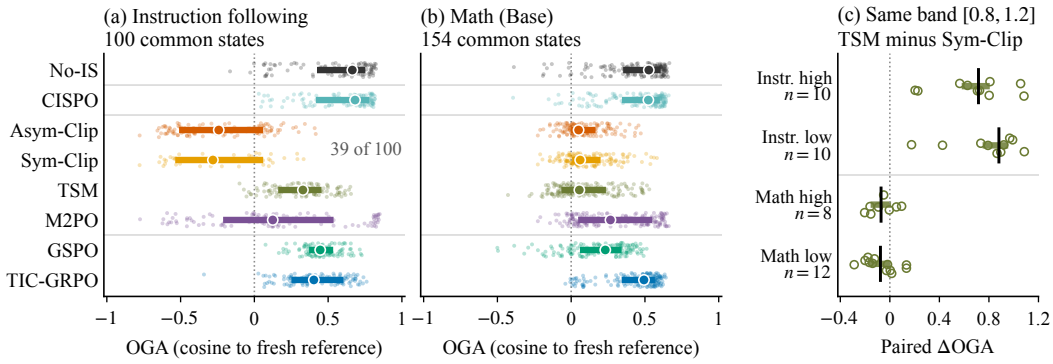

Figure 2: **Correction rules extract different directions from the same states.** (a,b) Faint points are common states; the large dot and bar are the median and interquartile range. Rows group no correction, a weight cap, signed clips and masks, and sequence ratios; instruction-following Sym-Clip uses the 39 states where we evaluate it. (c) Paired TSM minus Sym-Clip alignment at identical bounds on Sym-Clip checkpoints (high/low learning rate); the tick is the median.

despite a positive median. Alignment also varies widely across states within each rule, so no rule is uniformly best or worst.

The differences already appear at low drift. On the 96 shared mathematical states with sampled forward KL of at most 0.02 per token, median alignment is 0.590 for No-IS, 0.589 for CISPO, 0.109 for Asym-Clip and 0.154 for Sym-Clip; on 40 such instruction-following states, Asym-Clip is negative at 34 while No-IS and CISPO are positive at all (Table D.1).

3.2 What drives the directional differences?

A same-band comparison isolates branch selection from the numerical bounds. For token ratio r and bounds $l < 1 < u$, the clipped ascent surrogate $\min\{Ar, A\,\mathrm{clip}(r, l, u)\}$ has score coefficient $\nu ArG_{l,u}(A, r)$ away from ties, where

$$G_{l,u}(A, r) = \mathbf{1}\{A > 0, r < u\} + \mathbf{1}\{A < 0, r > l\}. \tag{8}$$

Sym-Clip thus retains the positive-advantage lower tail and the negative-advantage upper tail, whereas TSM deletes both regardless of the advantage sign. With $[l, u] = [0.8, 1.2]$ and identical remaining factors,

$$b_{\mathrm{Sym}} = b_{\mathrm{TSM}} + \underbrace{\sum_{A>0, r<l} \nu Arq}_{u_+} + \underbrace{\sum_{A<0, r>u} \nu Arq}_{u_-}. \tag{9}$$

The two operators agree inside the band, so their difference measures exactly the contribution of the retained tails.

At the 20 instruction-following Sym-Clip checkpoints, TSM has higher alignment in every pair; the mean difference $\mathrm{OGA}_{\mathrm{TSM}} - \mathrm{OGA}_{\mathrm{Sym}}$ is $+0.733$, and in 16 pairs a negative Sym-Clip direction becomes positive under TSM; on the 39 shared states of other runs, TSM is higher at 31. The mathematical majority reverses: Sym-Clip is higher in 14 of 20 pairs, with mean difference -0.069. Adding the retained tails therefore lowers alignment in instruction following and mostly raises it in mathematics: the same ratio bounds give opposite orderings, depending on what the tails carry in each task (whose model, data and reward all differ).

Near-equal coefficient totals, different directions. On low-drift states ($D_k \leq 0.02$; Section 3.3), Asym-Clip closes the derivative gate on a median 0.83% of valid mathematical tokens and keeps 98.7% of No-IS's absolute coefficient mass, yet No-IS is more aligned by a median 0.43 (47 states). In instruction following the corresponding figures are 1.73%, 98.6% and 1.26 (31 states). By (8), the closed gates are exactly the tokens whose probabilities have already moved past the bound in

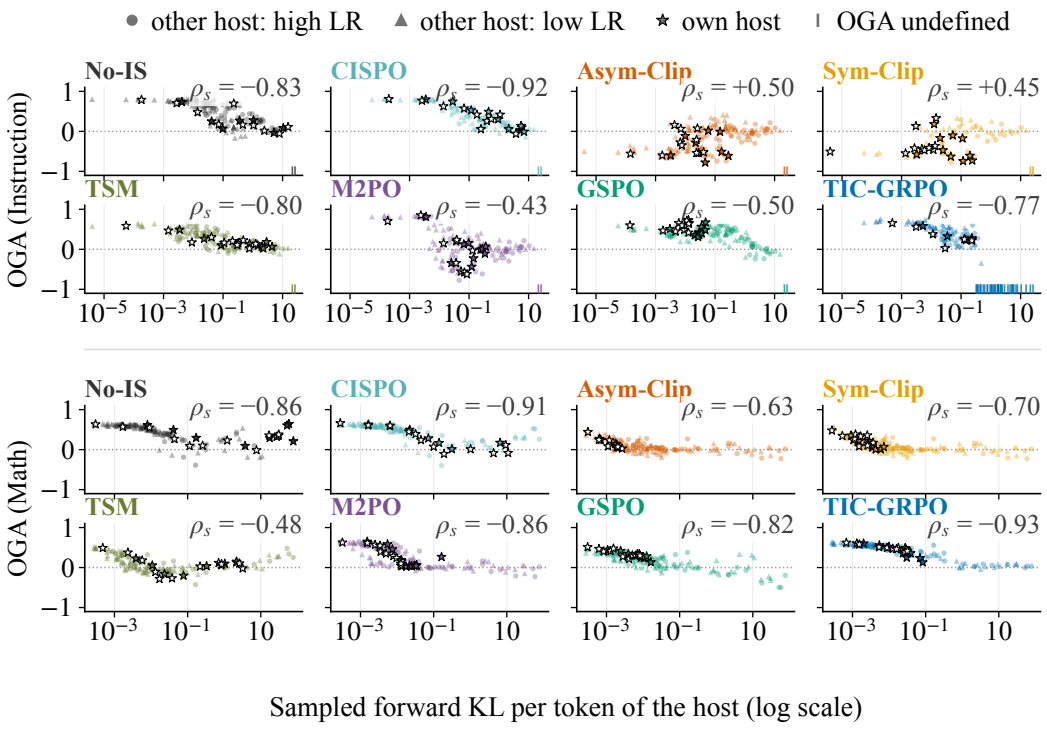

Figure 3: **OGA changes with host-policy drift.** The top two rows show instruction following and the bottom two Base math. Circles/triangles mark other-host high/low-LR states, stars mark own-host states, and bottom ticks mark undefined OGA. ρ_s is the descriptive pooled Spearman correlation over correlated checkpoints (Appendix D.3).

the direction their advantage favors ($A > 0, r > 1.28$ or $A < 0, r < 0.8$); CISPO and M2PO keep most such tokens, whose informativeness motivated both designs (MiniMax, 2025; Zheng et al., 2025b). Asym-Clip also reweights every retained token by its near-unit ratio, so this comparison combines deletion with reweighting. A nested control isolates deletion: on the same mathematical states, tightening Asym-Clip to Sym-Clip deletes only the positive-advantage strip $1.2 < r < 1.28$, a median 0.14% of tokens and 0.57% of coefficient mass. Most such deletions change alignment modestly (median $+0.02$), and the largest change is $+0.12$. The directional effect of a coefficient change therefore depends on where it acts: by (5), on how the score vectors of the changed tokens align with the fresh gradient (Appendix D.2).

3.3 Directional alignment across hosts and policy drift

We next evaluate every rule on the policy states reached by *all* training runs. We call the run that produced a checkpoint its *host*; drift is a property of the host policy. We measure it by the sampled forward KL per token on old responses, $D_k = \frac{1}{N_B}\sum_{i\in B}\frac{1}{T_i}\sum_t \log\big(\mu(y_{it} \mid x_i, y_{i,<t})/\pi_{\theta_k}(y_{it} \mid x_i, y_{i,<t})\big)$, which needs only the current and stored log probabilities of the old responses.

Across all host states (Figure 3), OGA is generally lower at greater drift: all eight pooled correlations are negative in mathematics (-0.48 to -0.93), six of eight are negative in instruction following (-0.43 to -0.92), and 27 of 32 own-path correlations are negative (Appendix D.4). The instruction-following signed clips are the exception: their initially negative directions move toward zero, producing positive pooled correlations ($+0.50$ and $+0.45$).

These associations motivate drift as an observable monitor of a reused pool without a fresh rollout. Drift does not measure checkpoint utility, so Section 4 calibrates delivery on task gain rather than an alignment threshold.

	Instruction following (reward $\times 100$)		Math (Base) (accuracy pp)	
Rule	LR 5×10^{-6}	LR 2.5×10^{-6}	LR 5×10^{-6}	LR 2.5×10^{-6}
No-IS	$+1.49 \to -47.12$	$+13.26 \to -46.48$	$+11.18 \to -16.19$	$+12.54 \to -20.16$
CISPO	$+2.82 \to -46.50$	$+2.72 \to -34.41$	$+6.65 \to -15.17$	$+11.59 \to +5.61$
Asym-Clip	$+4.98 \to -16.08$	$+4.11 \to +3.89$	$+2.25 \to +2.17$	$+1.80 \to +1.80$
Sym-Clip	$+3.56 \to -8.20$	$+3.69 \to +3.26$	$+2.61 \to +2.11$	$+1.74 \to +1.74$
TSM	$+1.92 \to -39.12$	$+7.36 \to +2.31$	$+2.84 \to +2.04$	$+5.21 \to -5.26$
M2PO	$+7.02 \to -3.18$	$+6.50 \to +6.50$	$+3.96 \to +2.60$	$+4.85 \to +4.85$
GSPO	$+3.87 \to +2.79$	$+2.39 \to +1.86$	$+1.75 \to +1.47$	$+1.79 \to +1.63$
TIC-GRPO	$+2.99 \to -45.83$	$+2.12 \to -18.68$	$+5.22 \to +1.40$	$+3.41 \to +3.28$

Table 1: **Obtaining gains and retaining them are different problems.** Each cell gives the best observed gain up to step 256 (a hindsight maximum over the initial model and saved checkpoints) $\to$ the gain at step 256, the common reporting cutoff. Curves appear in Appendix G.

4 From Learning Signals to Realized Gains

4.1 Useful direction and a finite reuse window

Section 3 answers the local direction question; here we ask a distinct pathwise question: after useful updates have been obtained, which checkpoint should be delivered? Equation (6) shows that later terms can cancel earlier progress, while Proposition 1 identifies direction and step size as conditional ingredients of one update. Neither result determines where an AdamW trajectory reaches its best score. We therefore first measure whether continued reuse loses observed gains, then calibrate delivery decisions directly on held-out task outcomes. Task gain is the change from the initial model, in continuous reward $\times 100$ for instruction following and accuracy points (pp) for mathematics.

Table 1 reports both quantities for all 32 trajectories. Every run improves on its initial model at some checkpoint; the difficulty is retaining the gain. At the higher rate in instruction following, seven of eight rules end below the initial model, four of them by more than 39 points, although all eight peaked within the first eight updates. Halving the rate rescues Asym-Clip, Sym-Clip, TSM and M2PO in that task but not No-IS, CISPO or TIC-GRPO, and in mathematics it turns TSM's retained gain into a loss. The largest mathematical peaks, those of No-IS and CISPO, also produce the largest losses: three of these four runs end below the initial model. The clipping rules and GSPO obtain smaller gains and keep most of them. This contrast matches the prosperity-before-collapse pattern that Zheng et al. (2025b) report for stale-data training without a trust region. The learning rate thus changes where a run ends up, and a useful reuse window is specific to a rule and a protocol. This motivates a calibrated delivery rule. The alignment–drift relation motivates KL as a state signal, but the calibration target below is delivered task gain, not an assumed alignment threshold.

4.2 Stopping rules calibrated on other trajectories

Drift is observable on old responses without a fresh rollout. Within each task we hold out one method family at both learning rates (the two clips form one family), calibrate on the remaining trajectories $\mathcal{H}_{-h}$, and never use target scores for selection. The three rules isolate different roles. *Fixed step* selects a gain-maximizing budget K_{-h}, an elapsed-update coordinate that bounds reuse. *KL only* selects a forward-KL threshold τ_{-h}; the first $D_k > \tau_{-h}$ delivers the preceding checkpoint, or step 256 without an alarm. *KL+cap* freezes K_{-h} and calibrates a residual threshold τ^{res}_{-h}, including $+\infty$ as a no-override candidate. An alarm at or before the cap rolls delivery back one checkpoint; otherwise the cap is delivered (Appendix F.4).

4.3 Delivered gain on held-out trajectories

Each task contributes 16 trajectories (Figure 4). In instruction following, continuing to step 256 gives -17.8 points; Fixed step delivers $+2.15$ after 8 updates and KL only $+2.31$ after 86.5. KL+cap changes four higher-rate deliveries at the step-8 cap: two improve and two worsen, yielding a net $+0.51$ over Fixed step ($+2.66$ total) and reducing negative deliveries from two to zero. All four

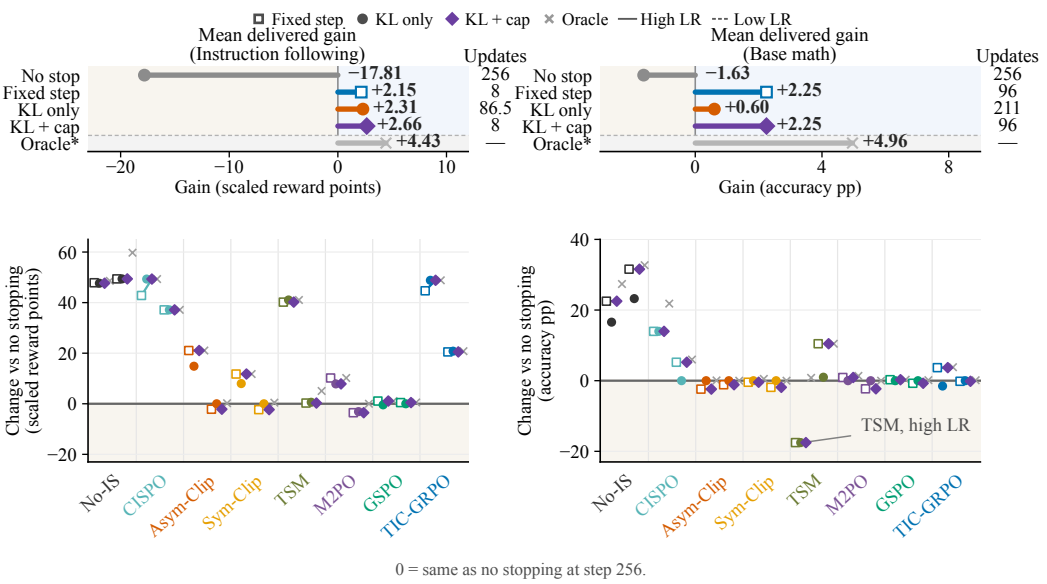

Figure 4: **Calibrated stopping retains gains that continued training loses.** Top: mean delivered gain and executed updates per policy; KL+cap adds a drift override to the fixed schedule. Bottom: per-run change vs. no stopping, by rule. Segments connect Fixed and KL+cap for the same run; × = oracle. 16 trajectories per task (one seed), calibrated on other families only.

alarms occur at the cap and roll back to step 4, so this is checkpoint selection, not saved updates. The LR-dependent shift in observed peak timing (median 6 versus 96) motivates state conditioning; the paired same-cap change is the evidence for KL's increment. In mathematics, no selected override fires by the cap, so all deliveries remain cap-determined and KL+cap retains the Fixed-step result of $+2.25$ pp.

5 Discussion and Conclusion

The direction that old responses supply depends jointly on the advantage and the correction rule. Corrections extract different directions from the same states even near the sampling policy; changes that keep almost all of the coefficient mass can still move the direction, because each changed coefficient acts through its score vector; and the same two ratio tails, selected by the advantage sign, help alignment in one task and hurt it in the other. As the policy drifts, alignment falls for most rules. This is a result about local learning-signal geometry: rule comparisons must state their advantage construction and policy state, and no alignment level by itself specifies a checkpoint's downstream utility.

The pathwise view supplies the complementary delivery question. In held-out replay, a calibrated step cap bounds reuse and KL provides a trajectory-conditioned override without reading target scores or fresh gradients. In instruction following, KL+cap adds $+0.51$ reward points over the same cap (two changed deliveries improve, two worsen, and negative deliveries vanish); in mathematics no selected override fires, so the rule preserves the calibrated budget. Drift is thus conditionally useful state information, not a universal surrogate for OGA, reward, or the optimal checkpoint, and we report direction diagnostics and delivered outcomes as separate, linked tests.

Scope. The evidence covers two tasks, one 1.7B model family (Base and non-Base) and one training seed. The fresh reference depends on the advantage and reduction choices, and we measure alignment before Adam's preconditioning. We evaluate stopping by retrospective replay over a fixed window and leave calibration and monitoring costs out of the comparison. The combined contract was chosen after examining alternatives on this recorded collection; independent seeds and prospective runs are the next test. Within this scope, the paper connects an account of the stale-data learning signal to a testable way of keeping the gains it produces.

AI USE STATEMENT

Generative AI assisted with research-note organization, manuscript drafting and critique, and code for analysis and visualization. Numerical claims come from recorded measurements or explicit algebra. The authors remain responsible for the scientific claims and final submission.

REPRODUCIBILITY STATEMENT

We will release the code and data upon publication: the training and diagnostic code, the training prompts with their stored responses, the evaluation sets, and the archived measurements, together with scripts that regenerate every figure, table and the retrospective stopping replay from those measurements, with input hashes and row-level provenance. Appendix C specifies the local derivative and factors of each correction, Appendix F the models, data, evaluation sets, measurement protocol and stopping replay, Appendix B the proof of Proposition 1, and Appendix E the critic comparisons.

ETHICS STATEMENT

This work studies optimization diagnostics for language-model reinforcement learning on mathematical and instruction-following tasks. It does not involve human subjects or private user data. The limited model, task, and seed coverage and the absence of cost-accounted prospective validation should be considered before operational use.

REFERENCES

Charles Arnal, Vivien Cabannes, Taco Cohen, Julia Kempe, et al. Efficient RL training for LLMs with experience replay, 2026.

Thomas Degris, Martha White, and Richard S. Sutton. Off-policy actor-critic. In *International Conference on Machine Learning (ICML)*, 2012.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep policy gradients: A case study on PPO and TRPO. In *International Conference on Learning Representations (ICLR)*, 2020.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning (ICML)*, 2018.

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, 2025.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research (JMLR)*, 5:1471–1530, 2004.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng 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*, 2024.

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, 2026. Project page: https://qingyonghu.github.io/AVSPO.

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 *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Luke J. Huang, Zhuoyang Zhang, Qinghao Hu, Shang Yang, and Song Han. Stable asynchrony: Variance-controlled off-policy RL for LLMs, 2026.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. A closer look at deep policy gradients. In *International Conference on Learning Representations (ICLR)*, 2020.

Amirhossein Kazemnejad, Milad Aghajohari, Eva Portelance, et al. Vineppo: Refining credit assignment in rl training of llms, 2024.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning, 2021.

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*, 2022.

Hunter Lightman, Vineet Kosaraju, Yura Burda, et al. Let's verify step by step, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, et al. Understanding R1-zero-like training: A critical perspective, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Jian Lu. Periodic asynchrony: An on-policy approach for accelerating llm rl, 2025.

Michael Luo, Sijun Tan, Roy Huang, et al. Deepscaler: Surpassing o1-preview with a 1.5b model by scaling rl. Notion/HuggingFace release, Agentica, 2025.

Yuchun Miao, Sen Zhang, Yuqi Zhang, Yaorui Shi, et al. When to stop reusing: Dynamic gradient gating for sample-efficient RLVR, 2026.

MiniMax. Minimax-m1: Scaling test-time compute efficiently with lightning attention, 2025.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Ashvin Nair, Abhishek Gupta, Murtaza Dalal, and Sergey Levine. Awac: Accelerating online reinforcement learning with offline datasets, 2020.

Michael Noukhovitch et al. Asynchronous rlhf: Faster and more efficient off-policy rl for language models, 2024.

Lucas Monteiro Paes, Natalie Mackraz, Barry-John Theobald, and Federico Danieli. Theoretical limits of language model alignment, 2026.

Lei Pang, Jun Luo, and Ruinan Jin. TIC-GRPO: Provable and efficient optimization for reinforcement learning from human feedback, 2025. URL https://arxiv.org/abs/2508.02833v3.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning, 2019.

YSS Prasanth, Ren, and others (Layer 6 AI). A gradient perspective on rlvr stability and winner advantage policy optimization (wapo), 2026.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Ritter, Oertell, Guo, Chang, Brantley, and Sun. Llms can learn to reason via off-policy rl (oapl), 2026.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024.

Shen, Zhao, Cheng, Huang, and Yu. VESPO: Variational sequence-level soft policy optimization for stable off-policy LLM training, 2026.

Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less, 2025.

Guangming Sheng, Yuxuan Tong, Borui Wan, Wang Zhang, Chaobo Jia, Xibin Wu, et al. Laminar: A scalable asynchronous RL post-training framework, 2025.

Matej Sirovatka and Prime Intellect Team. RL at 1T scale: prime-rl performance deep dive. Prime Intellect engineering report, June 21, 2026. URL `https://www.primeintellect.ai/blog/rl-at-1t-scale`.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

Tian, Xie, and Wei. How off-policy can grpo be? (μ-grpo), 2026.

Tongyi DeepResearch Team. Tongyi DeepResearch technical report, 2025.

Wan, Wang, Huang, and Sun. Buffer matters: Unleashing the power of off-policy rl in llm reasoning (bapo), 2026.

Wang et al. When importance sampling misallocates credit: Asymmetric ratios for outcome-supervised rl (aspo), 2025.

Wang et al. What are key factors for updates in RL for LLM reasoning?, 2026.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Uncertainty in Artificial Intelligence (UAI)*, 2001.

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

Zhiheng Xi, Xin Guo, Yang Nan, Enyu Zhou, Junrui Shen, Wenxiang Chen, Jiaqi Liu, Jixuan Huang, Zhihao Zhang, Honglin Guo, Xun Deng, Zhikai Lei, Miao Zheng, Guoteng Wang, Shuo Zhang, Peng Sun, Rui Zheng, Hang Yan, Tao Gui, Qi Zhang, and Xuanjing Huang. Bapo: Stabilizing off-policy reinforcement learning for llms via balanced policy optimization with adaptive clipping, 2025.

Haofeng Xu, Junwei Su, Yukun Tian, Lansong Diao, Zhengping Qian, and Chuan Wu. GAC: Stabilizing asynchronous rl training for LLMs via gradient alignment control, 2026.

Fengkai Yang, Zherui Chen, Xiaohan Wang, Xiaodong Lu, et al. Your group-relative advantage is biased, 2026.

Qiying Yu et al. Dapo: An open-source llm reinforcement learning system at scale, 2025.

Zhang, Ye, Jin, Yu, Xiong, Sahu, and Jiang. Rethinking importance sampling in llm policy optimization: A cumulative token perspective (ctpo), 2026a.

Guilin Zhang, Chuanyi Sun, Kai Zhao, Xu Chu, et al. EvalStop: Using world feedback to detect and correct reward overoptimization in multi-tenant RLHF platforms, 2026b.

Zili Zhang, Yinmin Zhong, Chengxu Yang, Chao Jin, Bingyang Wu, Xinming Wei, et al. Heddle: A distributed orchestration system for agentic RL rollout, 2026c.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. Group sequence policy optimization, 2025a.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy rl reach with stale data on llms?, 2025b.

Zhou et al. Demystifying grpo: Its policy gradient is a u-statistic, 2026.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

A More Related Work

We organize related work around the components of Figure 1: advantage estimation, weighting and gating, directional diagnosis, and control of the training path. The comparisons below emphasize protocol and reference choices rather than claim priority.

Off-policy reuse and policy lag. Asynchronous systems train on responses sampled by earlier policies, while the cost of long agentic rollouts makes reuse attractive (Noukhovitch et al., 2024; Fu et al., 2025; Huang et al., 2026). Prior work measures staleness in model updates, policy versions, reuse counts or buffer age (Zheng et al., 2025b; Arnal et al., 2026; Wan et al., 2026). Responses include synchronizing weights (Lu, 2025), bounding lag (Fu et al., 2025; Wan et al., 2026), correcting stale data (Zheng et al., 2025b; Huang et al., 2026), and stopping reuse from gradient statistics (Miao et al., 2026). We instead fix one pool, compare corrections at common policy states, and separately calibrate which checkpoint to deliver.

Corrections as weights and gates. Correction rules differ in ratio granularity, behavior outside a trust band, and whether gates depend on the advantage sign. Token clipping and capped weighting (Schulman et al., 2017; Yu et al., 2025; MiniMax, 2025; Xi et al., 2025), sequence ratios (Zheng et al., 2025a; Pang et al., 2025; Zhang et al., 2026a; Shen et al., 2026), batch-statistic masking (Zheng et al., 2025b), and sign-dependent gates (Prasanth et al., 2026; Tian et al., 2026) instantiate these choices. Their closest analyses emphasize unbiasedness, variance, stability or reward; our coefficient decomposition makes their local gradient directions comparable against one fresh-sample reference.

Advantage construction under reuse. Related analyses study finite-group and normalization bias, within-group reward collapse, critic error and variance-reducing baselines (Yang et al., 2026; Zhou et al., 2026; Liu et al., 2025; He et al., 2026; Kazemnejad et al., 2024; Weaver & Tao, 2001; Greensmith et al., 2004; Schulman et al., 2016). VinePPO recomputes Monte Carlo advantages at fixed on-policy states (Kazemnejad et al., 2024), whereas OAPL replaces the advantage with a sampling-policy-independent regression target (Ritter et al., 2026). Our setting isolates how the stale advantage both rescales coefficients and, through its sign, selects clipping branches.

Directional diagnostics. Gradient cosine against a large-sample reference diagnoses estimator fidelity on-policy (Ilyas et al., 2020); other references include adjacent-step self-correlation (Xu et al., 2026), on-policy-configuration subspaces (Wang et al., 2026), and sign-conditioned token geometry (Prasanth et al., 2026; Wang et al., 2025). We use a fresh-sample gradient at the same policy state, shared across the correction rules being compared. This reference diagnoses raw-gradient geometry, not the executed AdamW update.

From diagnostics to delivered gains. Prior controls scale learning rates, cap policy lag, gate reuse from gradient statistics, or stop on evaluation scores (Huang et al., 2026; Xu et al., 2026; Miao et al., 2026; Zhang et al., 2026b). KL-budget studies locate overoptimization peaks retrospectively (Gao et al., 2023; Rafailov et al., 2024; Shenfeld et al., 2025; Paes et al., 2026). Our leave-family-out replay separates a fixed update budget, a KL-only state signal and a KL override on the frozen budget, then evaluates the checkpoint each policy delivers.

Boundary: offline policy improvement. Offline RL uses support constraints and advantage-weighted regression on fixed datasets (Peng et al., 2019; Nair et al., 2020; Kostrikov et al., 2021). Its coverage assumptions and objectives differ from our question of how long one stale pool continues to supply useful directions at an evolving policy.

B Local Derivatives, Optimizer Scope and Improvement Condition

B.1 Local derivatives

For detached A and ratio $r = \pi_\theta/\mu$, differentiating $\min\{Ar, A\,\mathrm{clip}(r, l, u)\}$ with respect to token log probability gives, away from ties,

	$r < l$	$l < r < u$	$r > u$
$A > 0$	Ar	Ar	0
$A < 0$	0	Ar	Ar

This proves (8). For any disjoint token partition $\{S_j\}$ and a shared nonzero reference, linearity gives $P(\sum_z c_z q_z) = \sum_j \langle \hat{f}, \sum_{z\in S_j} c_z q_z\rangle / \|\hat{f}\|^2$, which proves (5); cosines are not additive.

For the sequence mean $s_i = \exp(T_i^{-1} \sum_t (\ell_{it} - \ell_{it}^0))$, differentiation contributes s_i/T_i per valid token. The unnormalized trajectory product contributes its full ratio. A live cap contributes its Jacobian, whereas a detached bounded weight contributes only its value. Appendix C.1 records the resulting factors for every retained rule.

B.2 AdamW and raw-gradient diagnostics

AdamW maintains coordinatewise moments $m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2$, then updates with $-\eta[\hat{m}_t/(\sqrt{\hat{v}_t} + \epsilon_{\mathrm{A}}) + \lambda\theta_t]$ (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). If an entire gradient history is multiplied by a common positive scalar, the moment ratio is unchanged when ϵ_{A} is negligible; global norm clipping also preserves the direction of a single gradient. These invariances do not equate updates from two non-collinear rule gradients. Coordinatewise history, weight decay and ϵ_{A} can rotate or shrink the executed update. Thus OGA and ρ diagnose raw gradients; $\cos(u, \hat{f})$ in Proposition 1 concerns the preconditioned update.

B.3 Proof of Proposition 1

Because ∇J is L-Lipschitz on the update segment,

$$J(\theta + u) - J(\theta) \geq \langle f, u\rangle - \tfrac{1}{2} L \|u\|^2.$$

Writing $f = \hat{f} + (f - \hat{f})$ and applying Cauchy–Schwarz gives

$$\langle f, u\rangle \geq \langle \hat{f}, u\rangle - \epsilon\|u\| = \|u\|\big(\|\hat{f}\|\cos(u, \hat{f}) - \epsilon\big).$$

Substitution proves (7). The lower bound is positive when the stated directional margin is positive and the update norm is below the stated threshold. For $u = \alpha b$, $\alpha > 0$, it becomes $\alpha\|b\|(\|\hat{f}\|\mathrm{OGA}(b) - \epsilon) - \frac{L}{2}\alpha^2\|b\|^2$. These conditions are sufficient for one smooth population-objective update; drift estimates neither L nor ϵ.

C Operator Factors and Advantage Boundaries

C.1 Local factors of the retained methods

The following table records the local operators used by the analysis in the order used throughout the paper. Parameters are taken from the archived experimental definitions; the displayed reductions define the common-state diagnostic gradients, not the executed AdamW updates.

$$\ell_{it} = \log \pi_\theta(y_{it} \mid x_i, y_{i,<t}), \quad q_{it} = \nabla_\theta \ell_{it}, \quad c_{it} = \frac{\partial(-L)}{\partial \ell_{it}}, \quad b = \sum_{i,t} c_{it} q_{it}, \qquad c_{it} = \nu_{it} A_{it} w_{it} m_{it}.$$

The partial derivative includes all log-probability paths before factorization; sg denotes stop-gradient. Advantages and discrete selection decisions are detached. Here v_{it} is valid-token membership, $T_i = \sum_t v_{it}$, $M = \sum_i T_i$, and N is the number of nonempty responses (not the prompt-group size).

$$d_{it} = \ell_{it} - \ell^0_{it}, \quad r_{it} = e^{d_{it}}, \quad s_i = \exp\left(T_i^{-1} \sum_t v_{it} d_{it}\right), \quad \rho_i = \exp\left(\sum_t v_{it} d_{it}\right),$$
$$G_{l,u}(A, r) = \mathbf{1}\{A > 0, r < u\} + \mathbf{1}\{A < 0, r > l\}.$$

Rule	ν_{it}	A	w	m / recorded parameter
No-IS	v_{it}/M	A_{it}	1	1
CISPO	v_{it}/M	A_{it}	$\mathrm{sg}\min(r_{it}, 1.2)$	1; detached upper cap, $\epsilon_{\mathrm{hi}} = .2$
Asym-Clip	v_{it}/M	A_{it}	r_{it}	$G_{.8,1.28}(A_{it}, r_{it})$; $\epsilon = (.2, .28)$
Sym-Clip	v_{it}/M	A_{it}	r_{it}	$G_{.8,1.2}(A_{it}, r_{it})$; $\epsilon = (.2, .2)$
TSM	v_{it}/M	A_{it}	r_{it}	$\mathbf{1}\{.8 \le r_{it} \le 1.2\}$; detached band test
M2PO	v_{it}/M	A_{it}	r_{it}	$H^{\mathcal{U}}_{.04}(i,t)$; no PPO clip; original-token denominator
GSPO	$v_{it}/(NT_i)$	A_i	s_i	$G_{.9997,1.0004}(A_i, s_i)$; $\epsilon = (3,4)10^{-4}$
TIC-GRPO	v_{it}/N	A_i	ρ_i	$\mathbf{1}\{\rho_i < 1.28\}$; $\epsilon_{\mathrm{hi}} = .28$

ν above records the diagnostic reduction. For token-reduced training microbatch j, the documented contract instead gives $\nu^{\mathrm{train}}_{it} = \alpha_j v_{it}/M_j$; uniform accumulation with one response per microbatch is sequence-weighted, not a global-token mean. GSPO and TIC-GRPO use sequence means. TIC's common theoretical $1/T_\star$ is omitted in the recorded implementation, not replaced by $1/T_i$. All displayed derivatives are away from clipping and sorting ties and assume unsaturated log ratios.

$$P_{l,u}(A, r) = \min\{Ar, A\,\mathrm{clip}(r, l, u)\},$$
$$-L_{\mathrm{GSPO}} = \frac{1}{N}\sum_i P_{.9997,1.0004}(A_i, s_i), \qquad -L_{\mathrm{TIC}} = \frac{1}{N}\sum_i A_i \min(\rho_i, 1.28).$$
$$\frac{\partial[\mathrm{sg}(A\min(r,u))\ell]}{\partial \ell} = A\min(r,u), \qquad \frac{\partial[A\min(r,u)]}{\partial \ell} = Ar\mathbf{1}\{r < u\}.$$

Thus CISPO retains above-cap score gradients, whereas the live TIC cap does not, for either advantage sign. GSPO uses a geometric-mean ratio; TIC uses a trajectory product. A token stop-gradient GSPO construction has the same local coefficient for response-constant advantages, not for arbitrary token GAE.

$$\mathcal{T} = \{z \in \mathcal{U} : v_z = 1,\ A_z d_z > 0\}, \quad D_{(1)} \ge \cdots \ge D_{(n)}, \quad D_z = d_z^2, \quad n = |\mathcal{T}|,$$
$$k_* = \min\left\{k \in \{0, \ldots, n\} : \frac{\sum_{j=k+1}^n D_{(j)}}{\max(n-k, 1)} \le .04\right\},$$
$$H^{\mathcal{U}}_{.04}(z) = 1 - \mathbf{1}\{z \text{ is among the first } k_* \text{ sorted TR tokens}\}.$$

For M2PO, $\mathcal{U}$ is the microbatch, not the entire diagnostic pool; TR denotes the valid tokens in $\mathcal{T}$ for which $A_z d_z > 0$; all other tokens are retained. The recorded training sizes are 4 (instruction following) and 1 (math).

Advantage definitions. With prompt group $g(i)$, environment reward R_i, and $G_i = |g(i)|$,

$$A_{it}^{\mathrm{grp}} = \mathrm{sg}\left(R_i - \frac{1}{G_i}\sum_{j\in g(i)} R_j\right), \qquad \delta_{it}^{V} = r_{it}^{\mathrm{env}} + \gamma\widetilde{V}_{i,t+1} - V_{it}, \qquad A_{it}^{\mathrm{GAE}} = \mathrm{sg}\left[\sum_{u=t}^{T_i}(\gamma\lambda)^{u-t}\delta_{iu}^{V}\right].$$

The recorded Base critic training uses $\gamma = 1$, $\lambda = .95$ with Asym-Clip, No-IS, and TSM. $\widetilde{V}_{i,T_i+1} = 0$ at a true terminal boundary; at a truncation it denotes the declared bootstrap value. In particular,

$$A_{it}^{\mathrm{GAE}(\lambda=1)} = \mathrm{sg}\left[\sum_{u=t}^{T_i}\gamma^{u-t}r_{iu}^{\mathrm{env}} - V_{it} + \gamma^{T_i-t+1}\widetilde{V}_{i,T_i+1}\right],$$
$$A_{it}^{\mathrm{GAE}(\gamma=\lambda=1)} = \mathrm{sg}(R_i - V_{it}) \quad \text{(terminal-only reward, zero terminal bootstrap).}$$

The main matrices evaluated on critic-trained hosts use A^{grp}. The critic comparisons use stale A^{GAE} from the frozen training critic against a fresh A^{grp} reference.

C.2 Baseline changes and gate sensitivity

For a fixed response and reward, let $A_o = R - b_o$, $A_f = R - b_f$ and $\delta = b_f - b_o$. Then $A_o - A_f = \delta$, and $A_o A_f < 0$ exactly when $R \in (\min(b_o, b_f), \max(b_o, b_f))$. For $A_f \neq 0$, $|A_o - A_f|/|A_f| = |\delta|/|R - b_f|$, so a small old margin is sensitive to a baseline change. A sign flip changes a clipping gate only when the ratio also lies outside its active band; advantage changes can alter coefficient magnitude without changing the gate.

C.3 Exact-importance-sampling boundary

For a full sequence ratio $r(y \mid x) = \pi(y \mid x)/\mu(y \mid x)$, common support and a frozen action-independent shift $\delta(x)$,

$$\mathbb{E}_\mu[r(y \mid x)\delta(x)\nabla \log \pi(y \mid x)] = \delta(x)\mathbb{E}_\pi[\nabla \log \pi(y \mid x)] = 0.$$

This identity requires the usual integrability and differentiation-interchange conditions. Token ratios, length normalization, self-including group means, random group scales, finite clipping and gates need not inherit it. Detaching a baseline removes an autograd path but not its statistical dependence on the sampled actions. We therefore use the coefficient framework without attributing every measured direction change to baseline staleness.

D Directional Controls and Robustness

D.1 Low-drift comparisons

The slice $D_k \leq 0.02$ is a descriptive low-drift window. Measurements are matched by task, host rule, learning rate and checkpoint; matrix gradients and actor-ladder KL remain distinct recorded products at the matched state.

D.2 Coefficient totals and nested deletion

With the same old responses, detached advantages, score vectors and reduction, tightening Asym-Clip to Sym-Clip removes only $\mathcal{S} = \{z : A_z > 0,\ 1.2 < r_z < 1.28\}$, away from boundary ties:

$$b_{\mathrm{Asym}} - b_{\mathrm{Sym}} = \sum_{z\in\mathcal{S}} \nu_z A_z r_z q_z. \tag{10}$$

The matched low-drift Base-math cohort contains 47 states on seven hosts. The median deletion is 0.139% of tokens and 0.568% of Asym-Clip's absolute coefficient mass; the median OGA change is $+0.020$ and the maximum is $+0.117$. Most changes are modest, but the range shows that the location of a small deletion matters through the affected score vectors.

Domain	Rule	n	Median cos	Mean cos	cos < 0	Mean Δ	$- \to +$
Instruction	No-IS	40	0.769	0.743	0/40	+1.108	34
Instruction	CISPO	40	0.790	0.770	0/40	+1.135	34
Instruction	Asym-Clip	40	-0.429	-0.365	34/40	+0.000	0
Instruction	TSM	40	0.491	0.462	0/40	+0.827	34
Instruction	M2PO	40	0.759	0.496	7/40	+0.861	27
Instruction	GSPO	40	0.466	0.463	0/40	+0.828	34
Instruction	TIC-GRPO	40	0.622	0.602	0/40	+0.967	34
Math (Base)	No-IS	96	0.590	0.565	1/96	+0.431	13
Math (Base)	CISPO	96	0.589	0.568	0/96	+0.435	13
Math (Base)	Asym-Clip	96	0.109	0.134	13/96	+0.000	0
Math (Base)	Sym-Clip	96	0.154	0.173	11/96	+0.040	2
Math (Base)	TSM	96	0.117	0.141	31/96	+0.007	10
Math (Base)	M2PO	96	0.508	0.439	0/96	+0.305	13
Math (Base)	GSPO	96	0.308	0.298	1/96	+0.165	12
Math (Base)	TIC-GRPO	96	0.556	0.545	0/96	+0.411	13

Table D.1: Common-state comparisons at sampled forward KL per token at most 0.02. Δ is relative to Asym-Clip; $- \to +$ counts a positive evaluated rule where Asym-Clip is negative. The slice contains 40 instruction-following and 96 mathematical states.

Math (Base)	States/hosts	Deleted tokens (%)	Deleted mass (%)	ΔOGA
All matched	47/7	0.1390 [0.1142, 0.1511] (0.0802, 0.1604)	0.5682 [0.4628, 0.6224] (0.3277, 0.6603)	+0.020098 [+0.007901, +0.044602] (-0.000652, +0.117212)
Step ≤ 256	39/7	0.1332 [0.1123, 0.1489] (0.0802, 0.1604)	0.5182 [0.4549, 0.6102] (0.3277, 0.6509)	+0.026530 [+0.009750, +0.049328] (-0.000369, +0.117212)

Table D.2: Asym-Clip $\to$ Sym-Clip on matched low-drift Base-math states. Cells give median, [IQR] and (minimum, maximum). Deleted mass is relative to Asym-Clip's absolute coefficient mass; $\Delta\mathrm{OGA} = \mathrm{OGA}_{\mathrm{Sym}} - \mathrm{OGA}_{\mathrm{Asym}}$.

Near-equal scalar totals provide a different, weaker diagnostic. Asym-Clip and No-IS differ in both ratio weighting and gates: their absolute coefficient mass can be close while their per-token coefficients and OGA differ. The table reports this comparison without treating equal totals as coefficient overlap.

Cohort	States/hosts	Flow (%)	Mass (%)	Median ΔOGA	IQR of Δ
Math (Base)	47/7	99.1679/100	98.6967/100	+0.434626	[+0.355835, +0.508217]
Math ≤ 256	39/7	99.2718/100	98.8597/100	+0.414346	[+0.335537, +0.497179]
Instruction	31/12	98.2706/100	98.6022/100	+1.262273	[+1.006933, +1.335580]

Table D.3: Scalar coefficient totals alongside alignment differences. Flow is the fraction of valid tokens with an open derivative gate; slash pairs are Asym-Clip/No-IS medians. Here $\Delta\mathrm{OGA} = \mathrm{OGA}_{\mathrm{NoIS}} - \mathrm{OGA}_{\mathrm{Asym}}$, and both weighting and gating change.

D.3 Cross-host drift coverage

Figure 3 contains 1,051 instruction-following and 1,232 mathematical rule–state cells. Pooled correlations mix hosts and all checkpoints are repeated measurements from one seed. Instruction-following Sym-Clip covers eight hosts, whereas the other instruction-following rules cover 14 or 16. Eight mathematical host states use 1,600 old responses rather than 12,480; excluding them leaves all eight pooled correlations negative, with maximum absolute change 0.092. Missing forward observations and invalid matrix gradients remain missing rather than being imputed.

D.4 Own-rule training trajectories

The separate actor-ladder diagnostic follows chronological checkpoints of the rule that trained each host. It contains 310 valid points on 32 trajectories; 27 within-trajectory Spearman correlations are negative. Positive exceptions are both instruction-following GSPO rates, lower-rate instruction-following Asym-Clip and Sym-Clip, and higher-rate mathematical TSM. These measurements use their own fresh-reference computations and are not merged with the cross-rule matrix diagonal.

E Robustness to Critic-Based Stale Advantages

E.1 Construction

We trained Qwen3-1.7B-Base mathematical runs with a frozen value model and GAE ($\gamma = 1$, $\lambda = .95$) under Asym-Clip, No-IS and TSM at both learning rates, seed 42. Direction comparisons use the three lower-rate runs (12 checkpoints each). In each host–checkpoint block, No-IS, Asym-Clip, Sym-Clip and TSM use the same old batch under two stale-side advantage constructions: token GAE from the frozen training critic, and group-relative advantages. The fresh reference is group-relative in both cases, so the comparison measures agreement with that declared target rather than ranking the advantage constructions intrinsically. The evaluations share checkpoints and stored fresh responses; fresh-reference norms agree within a median of 0.19% (maximum 1.0%).

At a true terminal, the GAE bootstrap is zero; at a truncation it uses the declared bootstrap value. Even at $\lambda = 1$, token GAE includes $-V_t$ and need not collapse to a response-constant return, so it can change the signed branches of (8).

E.2 Paired fixed-state comparison

Table E.1 holds the checkpoint fixed while changing the stale advantage or correction rule. Relative to the group-relative fresh reference, GAE lowers OGA for all four rules in every block (mean changes -0.39 to -0.46). Under GAE, Asym-Clip and Sym-Clip differ from No-IS by -0.005 on average and TSM by -0.018, compared with -0.040, -0.042 and $+0.010$ under group-relative stale advantages. These are repeated measurements from one seed.

Outcome comparisons between critic-trained and group-relative runs follow different optimization paths and are therefore not used to attribute a training effect to the advantage construction.

Eval rule	N	Diag.	Host–LR	ρ_{pool}	Med. ρ_{host}	U
Instruction following						
No-IS	156	17	16	-0.826	-0.624	2
CISPO	136	20	14	-0.923	-0.885	2
Asym-Clip	156	19	16	+0.504	+0.352	2
Sym-Clip	76	20	8	+0.455	+0.345	2
TSM	156	20	16	-0.796	-0.709	2
M2PO	136	20	14	-0.428	-0.782	2
GSPO	136	20	14	-0.501	-0.048	2
TIC-GRPO	99	13	14	-0.770	-0.719	39
Math (Base)						
No-IS	154	20	15	-0.858	-0.685	0
CISPO	154	19	15	-0.912	-0.867	0
Asym-Clip	154	12	15	-0.632	-0.385	0
Sym-Clip	154	20	15	-0.703	-0.420	0
TSM	154	20	15	-0.478	-0.853	0
M2PO	154	23	15	-0.856	-0.857	0
GSPO	154	20	15	-0.819	-0.797	0
TIC-GRPO	154	20	15	-0.935	-0.888	0
Math (Base): 12,480 old responses only						
No-IS	146	12	14	-0.917	-0.745	0
CISPO	146	19	14	-0.936	-0.880	0
Asym-Clip	146	12	14	-0.609	-0.420	0
Sym-Clip	146	20	14	-0.686	-0.465	0
TSM	146	20	14	-0.570	-0.863	0
M2PO	146	23	14	-0.853	-0.881	0
GSPO	146	20	14	-0.802	-0.780	0
TIC-GRPO	146	20	14	-0.932	-0.881	0

N: valid plotted cells; Diag.: actor rule = evaluation rule; Host–LR: distinct source/trajectory/host/LR groups with a plotted cell. U: undefined matrix OGA cells with eligible forward KL, shown as bottom rugs. Other missing/provenance exclusions are in the CSV audit.

Spearman coefficients are descriptive: pooled over plotted cells, or the median across within-host coefficients. Each coefficient requires at least three points and nonconstant KL and cosine. Checkpoints share trajectories; all records use seed 42. No independent-experiment interpretation or significance test.

Main math has 8 states/64 cells with 1,600 old responses and 146 states/1,168 cells with 12,480; the final block retains only the latter. Responses are sampling metadata, not independent repeats. Instruction Sym-Clip is evaluated on its own two hosts and on the six Asym-Clip, No-IS and TSM hosts; missing cross-host cells are not fabricated.

Cosines and fresh references come from the cross-rule matrix. Forward KL is the primary old-side actor-record field, joined by exact state and critic/seed/line/reward guards. Invalid own-rule OGA and reverse/JS/fresh-side flags do not veto a valid cross-rule cell.

Table D.4: Numerical companion to the cross-host drift view, including the 12,480-response mathematical sensitivity cohort. Correlations are descriptive and computed over correlated checkpoints.

Rule	Domain	n	ρ_s	Median cos	$n_{\cos<0}$	Max fwd
No-IS	Instruction following	7/9	-0.500/-0.950	+0.251/+0.174	1/1	10.52/7.857
No-IS	Math (Base)	8/11	-0.095/-0.327	+0.395/+0.350	0/0	77.4/30.91
CISPO	Instruction following	9/10	-0.867/-0.879	+0.305/+0.419	0/1	5.897/3.758
CISPO	Math (Base)	8/12	-0.881/-0.867	+0.227/+0.167	0/3	13.18/10.86
Asym-Clip	Instruction following	9/10	-0.017/+0.285	-0.509/-0.312	8/7	0.2799/0.04112
Asym-Clip	Math (Base)	8/12	-0.690/-0.888	+0.031/+0.136	2/0	0.01506/0.002667
Sym-Clip	Instruction following	10/10	-0.576/+0.624	-0.524/-0.432	9/8	0.2378/0.01446
Sym-Clip	Math (Base)	8/12	-0.810/-0.580	+0.088/+0.256	0/0	0.007192/0.003665
TSM	Instruction following	10/10	-0.685/-0.758	+0.119/+0.189	0/0	4.431/1.254
TSM	Math (Base)	8/12	+0.167/-0.329	+0.076/+0.071	2/4	2.176/3.194
M2PO	Instruction following	10/10	-0.200/-0.733	-0.003/+0.059	5/5	0.3616/0.131
M2PO	Math (Base)	12/12	-0.203/-0.937	+0.057/+0.315	0/0	0.174/0.01347
GSPO	Instruction following	10/10	+0.236/+0.224	+0.452/+0.598	0/0	0.04802/0.01307
GSPO	Math (Base)	8/12	-0.929/-0.643	+0.306/+0.330	0/0	0.01632/0.008256
TIC-GRPO	Instruction following	3/10	-1.000/-0.455	+0.325/+0.259	0/0	0.1457/0.2306
TIC-GRPO	Math (Base)	8/12	-0.976/-0.839	+0.329/+0.489	0/0	0.08349/0.03151

Table D.5: Own-rule direction–drift trajectories. Slash pairs distinguish the two learning rates. Correlations exclude zero or invalid gradients and are descriptive; higher-rate instruction-following TIC-GRPO has three valid points.

	Same rule, two advantages			Rule minus No-IS, same advantage	
Rule	Group-relative	GAE	GAE lower	Group-relative	GAE
No-IS	$+0.494$	$+0.066$	36/36	baseline	baseline
Asym-Clip	$+0.454$	$+0.061$	36/36	-0.040 (2/36)	-0.005 (10/36)
Sym-Clip	$+0.452$	$+0.061$	36/36	-0.042 (1/36)	-0.005 (8/36)
TSM	$+0.503$	$+0.048$	36/36	$+0.010$ (19/36)	-0.018 (1/36)

Table E.1: Critic-based versus group-relative stale advantages on the same 36 lower-rate mathematical host–checkpoint blocks. Both are evaluated against the group-relative fresh reference. Left: mean OGA and paired change; right: mean difference to No-IS under the same stale advantage. Checkpoints are repeated measurements from one seed.

F Experimental Details and Stopping Protocol

F.1 Models, data and evaluation sets

Mathematics. Qwen3-1.7B-Base answers with raw prompt completions and is scored by binary math-verifier correctness. The initial model answered 39,386 DeepScaleR problems (Luo et al., 2025) eight times each; 55.1% of the groups were all wrong and 1.6% all correct, and the remaining 17,085 problems (136,680 responses) form the training pool. Evaluation uses the held-out problems in Table F.1, eight responses each. The assembled set has 1,574 problems, of which 1,560 are scored in every reported evaluation. No evaluation prompt occurs in the 17,085-problem training pool; the 200 DeepScaleR items come from its separate evaluation split.

Source	Problems	Notes
DeepScaleR evaluation split (Luo et al., 2025)	200	held out
MATH-500 (Hendrycks et al., 2021; Lightman et al., 2023)	500	levels 1–5
OlympiadBench (He et al., 2024)	674	competition, English
Minerva Math (Lewkowycz et al., 2022)	100	seed-42 sample
AMC 2023	40	
AIME 2024	30	
AIME 2025	30	
Assembled / scored	1,574 / 1,560	

Table F.1: Mathematical evaluation set.

Instruction following. Qwen3-1.7B (non-Base) answers writing prompts that we generate with rule templates following IFEval (Zhou et al., 2023): each prompt specifies a word-count range and 4, 6, 8 or 10 hard constraints (about a quarter each) drawn from ten IFEval instruction types—keyword inclusion, forbidden words, a title, an end phrase, no commas, lowercase text, bullet lists, a postscript, quotation marks and placeholders. The pool has 9,000 prompts with 16 responses each; training consumes the first 8,192 (131,072 responses). Evaluation uses 800 prompts from the remaining 808 (pool indices 8,192–8,991), which no run trains on, with 16 responses each. "Bandit" in source files refers to this task. The initial evaluation scores are 0.2311 for mathematics and 0.472992 for instruction following.

For instruction following, let h_i of $n_i > 0$ hard constraints pass, let w_i be the post-thinking answer's word count, and let $[L_i, U_i]$ be the prompt-specific length range. The recorded reward has the multiplicative form

$$R_i = \min\{1, (h_i/n_i)\ell_i\}, \qquad \ell_i = \begin{cases} w_i/L_i, & w_i < L_i, \\ 1, & L_i \le w_i \le U_i, \\ \max\{0, 1 - (w_i - U_i)/b_i\}, & w_i > U_i. \end{cases}$$

Here $b_i \ge U_i/2$ is the prompt's overflow buffer. The delivered score is the response mean of R_i. It therefore differs from both the mean hard-constraint pass fraction and the all-constraints-pass rate of standard IFEval; length enters as a multiplicative factor. Reported scores are the recorded evaluator's outputs.

The fixed behavior pool is consumed in order, with no fresh-data replacement. Training minibatches, old diagnostic samples, fresh-reference responses and score evaluations are separate objects. Diagnostics and scores use the evaluation prompts above. Fresh gradients generated independently on the same prompts can still differ. Group-relative advantages subtract a prompt-group reward mean without standard-deviation normalization in the reported diagnostic contract.

The common stopping horizon is 256 updates. Instruction-following batch size and horizon correspond to 131,072 responses. This horizon is a reporting choice; mathematical observations beyond it appear only in the supplementary full-range record.

F.2 Measurement scope and eligibility

The archived measurements retain source, host rule, learning rate, checkpoint and seed, which define all joins. Common-state comparisons require valid cells for every included rule; a direction

Setting	Instruction following	Mathematics
Model	Qwen3-1.7B	Qwen3-1.7B-Base
Reward	Continuous	Binary correctness
Training group size	16	8
Global batch (responses)	512	256
Microbatch (responses)	4	1
Maximum training length	4096	8192
Learning rates	5×10^{-6} and 2.5×10^{-6}	
Schedule / seed	Constant, zero warmup / 42	
Optimizer	AdamW, $\beta = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay .01	
Gradient clipping	Global norm 1	

Table F.2: Recorded settings of the main training runs. Directional diagnostics use the declared common reduction in Appendix C.1 and measure raw-gradient geometry rather than the executed AdamW update.

is eligible only when its cosine is finite and in $[-1, 1]$, both norms are positive and no invalid flag applies. A stored zero produced by a zero gradient is undefined, and OGA is reported without split-half ceiling rescaling. Instruction-following Sym-Clip extensions reuse the same checkpoints and stored fresh responses; mathematics uses the Base-only matrix. Actor-ladder and cross-rule matrix directions remain distinct measurement products. Appendix E describes the separate critic-GAE robustness check. Row-level provenance and input hashes accompany the released artifacts.

F.3 Drift estimates

The drift estimate D_k (Section 3.3) is the mean old-sample negative log ratio after response-length normalization. Reverse-side or nonlinear drift warnings leave this old-side quantity well defined. We preserve those warnings in the exports and mark affected direction–drift points. Finite-sample forward estimates can be negative; the direction analysis keeps them. No eligible critic-free plotted point has a negative forward value. The stored pseudo-JS applies a nonlinear transformation after length normalization, differs from standard sequence JS, and is not used for stopping.

F.4 Leave-family-out stopping replay

The replay includes the eight critic-free rules at both learning rates, separately in each domain. There are seven outer families: the two clipping variants together and six other rules. All members and rates of a target family are excluded from calibration, which uses only critic-free trajectories of the same task.

For each outer fold, the KL candidates are the distinct, eligible calibration values at checkpoints up to 256, plus zero. Fixed-step candidates are the calibration trajectories' recorded update counts up to 256, plus zero. Each candidate is evaluated by mean gain across calibration trajectories with equal trajectory weights. Ties within 10^{-10} scaled gain points choose the smaller threshold or earlier fixed step. Candidate sets use no target scores, KL values or peaks; the calibration is a direct one-dimensional outcome search.

On the held-out trajectory, a strict $D_k > \tau$ triggers an alarm. The delivered model is the previous recorded saveable checkpoint, including one whose monitor is missing; first-observation alarms deliver the initial model. The first monitored step is 2 for instruction following and 8 for mathematics. No alarm delivers step 256. A missing, negative, or provenance-ineligible forward monitor skips the alarm test at that checkpoint without discarding its score. Two instruction-following step-256 monitors are missing; the fallback then continues training. Fixed-step stopping delivers the last saved checkpoint at or before its calibrated count. Executed updates for KL are counted through the alarm.

Oracle gain is the maximum over the initial model and the trajectory's actual checkpoints up to 256; it is a hindsight reference. The replay uses only the declared outcome-calibrated scalar rule. Counts exclude calibration runs, fresh-reference computation, log-probability monitoring cost and wall-clock overhead.

KL + cap (drift-aware budgeted stopping). The combined rule first selects a fixed-step budget K_{-h} on the donor trajectories using the procedure above, then freezes it. A residual KL threshold τ^{res}_{-h} is calibrated on the same donors under an inclusive-rollback delivery rule: if $D_k > \tau^{\text{res}}_{-h}$ at any checkpoint $k \le K_{-h}$, the rule delivers the preceding saved checkpoint; otherwise it delivers the cap checkpoint K_{-h}. The candidate set is zero, the donor KL values and $+\infty$ (no override); ties within 10^{-10} gain points select the *largest* (least aggressive) threshold. This sequential calibration measures the marginal value of KL over the fixed schedule without a two-dimensional search.

Held-out family	D/T	K	KL thr.	Residual thr.	Train F	Train KL	Train C	Test fire
CISPO	14/2	8	0.048024	0.084607	+2.53	+2.89	+2.87	1
Asym-/Sym-Clip	12/4	8	0.066635	0.084607	+1.94	+2.90	+2.88	0
GSPO	14/2	8	0.046645	0.084607	+2.02	+2.95	+2.82	0
M2PO	14/2	8	0.048024	0.041118	+1.75	+2.71	+2.62	1
No-IS	14/2	8	0.048024	0.084607	+2.20	+3.04	+3.02	1
TIC-GRPO	14/2	8	0.048024	0.084607	+2.41	+2.92	+2.92	1
TSM	14/2	8	0.048024	0.116948	+2.20	+2.94	+2.96	0

D/T gives donor/test trajectories. Fixed first selects K; KL-only selects τ_{KL}. With K frozen, KL+cap selects τ_{res} from zero, eligible donor KL values and $+\infty$; ties choose the largest threshold. Test fire counts held-out inclusive rollbacks at or before the cap. No target score or KL enters calibration.

Held-out family	D/T	K	KL thr.	Residual thr.	Train F	Train KL	Train C	Test fire
CISPO	14/2	128	1.03381	7.63921	+2.88	+2.74	+3.43	0
Asym-/Sym-Clip	12/4	96	0.045063	13.1794	+4.51	+3.00	+4.92	0
GSPO	14/2	96	0.045063	13.1794	+3.82	+2.91	+4.16	0
M2PO	14/2	96	0.045063	13.1794	+3.57	+2.60	+3.92	0
No-IS	14/2	96	0.032127	$+\infty$	+2.74	+2.78	+2.74	0
TIC-GRPO	14/2	96	0.071556	13.1794	+3.42	+2.87	+3.77	0
TSM	14/2	64	0.171131	$+\infty$	+3.76	+4.13	+3.76	0

D/T gives donor/test trajectories. Fixed first selects K; KL-only selects τ_{KL}. With K frozen, KL+cap selects τ_{res} from zero, eligible donor KL values and $+\infty$; ties choose the largest threshold. Test fire counts held-out inclusive rollbacks at or before the cap. No target score or KL enters calibration.

Table F.3: Fixed, KL-only and KL+cap calibration by held-out family. Both learning rates and both clipping variants are excluded together where applicable. Parameters use only donor trajectories from the same task; Test fire counts inclusive KL+cap rollbacks on the held-out trajectories.

G Supplementary Training and Delivery Results

G.1 Gain trajectories behind the peak and cutoff summary

Table 1 reports each trajectory's observed peak and step-256 gain. Figure G.1 restores checkpoint order for three illustrative rules. It is descriptive rather than a method ranking; the complete checkpoint data for all rules accompany the plotting sources.

G.2 Stopping aggregates and per-trajectory alternatives

The tables report Fixed step, KL only and KL+cap on the same held-out trajectories. Gains are task-specific; a negative delivery is below the corresponding initial model even when it improves on the no-stop endpoint.

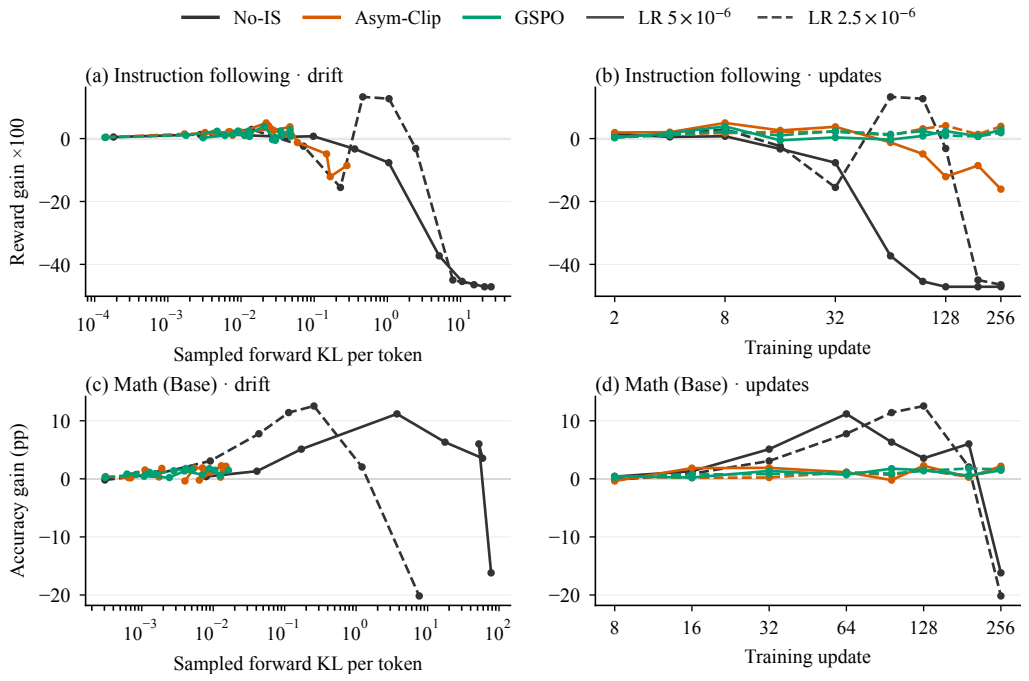

Figure G.1: **Obtaining gains and retaining them along the trajectory.** Gain against sampled forward KL per token (left) and update index (right) in the common step-256 window. Solid and dashed lines denote the two learning rates.

Domain	Rule	n	Gain	Δ vs N	W/T/L	Negative	Cost
Instruction	No stop	16	-17.813	+0.000	0/16/0	10	256.0
Instruction	Fixed	16	+2.152	+19.964	13/0/3	2	8.0
Instruction	KL	16	+2.307	+20.120	11/3/2	2	86.5
Instruction	KL+cap	16	+2.658	+20.471	13/0/3	0	8.0
Instruction	Oracle	16	+4.426	+22.239	15/1/0	0	–
Base math	No stop	16	-1.629	+0.000	0/16/0	4	256.0
Base math	Fixed	16	+2.248	+3.877	8/0/8	4	96.0
Base math	KL	16	+0.603	+2.233	4/10/2	4	211.0
Base math	KL+cap	16	+2.248	+3.877	8/0/8	4	96.0
Base math	Oracle	16	+4.963	+6.592	13/3/0	0	–

Retrospective leave-family-out replay. Gains use scaled continuous-reward points for instruction following and accuracy pp for Base math. W/T/L compares with no stopping; negative counts delivery below θ_0. KL+cap versus Fixed is 2/12/2 with mean $+0.506$ in instruction following and 0/16/0 with mean 0 in Base math. Cost counts executed updates and excludes calibration and monitoring overhead; Oracle is hindsight only.

Table G.1: Stopping-policy aggregates on the held-out trajectories. Cost counts executed updates and excludes calibration and monitoring overhead; W/T/L compares with no stopping.

Method	LR	N	F	KL	C	O	Δ(C–F)	C/D
No-IS	5	-47.12	+0.74	+0.54	+0.54	+1.49	-0.20	8/4
No-IS	2.5	-46.48	+2.91	+2.91	+2.91	+13.26	+0.00	8/8
CISPO	5	-46.50	-3.68	+2.82	+2.82	+2.82	+6.49	8/4
CISPO	2.5	-34.41	+2.72	+2.72	+2.72	+2.72	+0.00	8/8
Asym-Clip	5	-16.08	+4.98	-1.23	+4.98	+4.98	+0.00	8/8
Asym-Clip	2.5	+3.89	+1.73	+3.89	+1.73	+4.11	+0.00	8/8
Sym-Clip	5	-8.20	+3.56	-0.22	+3.56	+3.56	+0.00	8/8
Sym-Clip	2.5	+3.26	+0.94	+3.26	+0.94	+3.69	+0.00	8/8
TSM	5	-39.12	+1.07	+1.92	+1.07	+1.92	+0.00	8/8
TSM	2.5	+2.31	+2.59	+2.90	+2.59	+7.36	+0.00	8/8
M2PO	5	-3.18	+7.02	+4.67	+4.67	+7.02	-2.34	8/4
M2PO	2.5	+6.50	+2.97	+3.40	+2.97	+6.50	+0.00	8/8
GSPO	5	+2.79	+3.87	+2.37	+3.87	+3.87	+0.00	8/8
GSPO	2.5	+1.86	+2.33	+1.86	+2.33	+2.39	+0.00	8/8
TIC-GRPO	5	-45.83	-1.15	+2.99	+2.99	+2.99	+4.15	8/4
TIC-GRPO	2.5	-18.68	+1.82	+2.12	+1.82	+2.12	+0.00	8/8
Mean	–	-17.81	+2.15	+2.31	+2.66	+4.43	+0.51	8.0/7.0

Gains are scaled continuous-reward points. N: no stopping at step 256; F: Fixed; KL: KL only; C: KL+cap; O: hindsight Oracle. Δ(C–F) is KL's paired increment over the frozen cap. C/D gives KL+cap executed/delivered updates. Both learning rates (units 10^{-6}) and both clipping variants are held out together where applicable.

Table G.2: Instruction-following per-trajectory delivery for no stopping, Fixed, KL only, KL+cap and the hindsight oracle. Units are scaled continuous-reward points; C/D gives KL+cap executed/delivered updates.

Method	LR	N	F	KL	C	O	Δ(C–F)	C/D
No-IS	5	-16.19	+6.32	+0.38	+6.32	+11.18	+0.00	96/96
No-IS	2.5	-20.16	+11.41	+3.08	+11.41	+12.54	+0.00	96/96
CISPO	5	-15.17	-1.20	-1.20	-1.20	+6.65	+0.00	128/128
CISPO	2.5	+5.61	+10.83	+5.61	+10.83	+11.59	+0.00	128/128
Asym-Clip	5	+2.17	-0.22	+2.17	-0.22	+2.25	+0.00	96/96
Asym-Clip	2.5	+1.80	+0.66	+1.80	+0.66	+1.80	+0.00	96/96
Sym-Clip	5	+2.11	+1.67	+2.11	+1.67	+2.61	+0.00	96/96
Sym-Clip	2.5	+1.74	-0.15	+1.74	-0.15	+1.74	+0.00	96/96
TSM	5	+2.04	-15.48	-15.48	-15.48	+2.84	+0.00	64/64
TSM	2.5	-5.26	+5.21	-4.27	+5.21	+5.21	+0.00	64/64
M2PO	5	+2.60	+3.51	+2.60	+3.51	+3.96	+0.00	96/96
M2PO	2.5	+4.85	+2.55	+4.85	+2.55	+4.85	+0.00	96/96
GSPO	5	+1.47	+1.75	+1.47	+1.75	+1.75	+0.00	96/96
GSPO	2.5	+1.63	+0.90	+1.63	+0.90	+1.79	+0.00	96/96
TIC-GRPO	5	+1.40	+5.10	-0.11	+5.10	+5.22	+0.00	96/96
TIC-GRPO	2.5	+3.28	+3.12	+3.28	+3.12	+3.41	+0.00	96/96
Mean	–	-1.63	+2.25	+0.60	+2.25	+4.96	+0.00	96.0/96.0

Gains are accuracy pp. N: no stopping at step 256; F: Fixed; KL: KL only; C: KL+cap; O: hindsight Oracle. Δ(C–F) is KL's paired increment over the frozen cap. C/D gives KL+cap executed/delivered updates. Both learning rates (units 10^{-6}) and both clipping variants are held out together where applicable.

Table G.3: Mathematical per-trajectory delivery for the same policies. Units are accuracy pp; C/D gives KL+cap executed/delivered updates.