SCALING LAWS FOR COLLAPSE IN ASYNCHRONOUS GRPO

Anonymous authors
Paper under double-blind review

ABSTRACT

Asynchronous reinforcement learning improves the throughput of large language model post-training by decoupling rollout generation from policy optimization, but introduces a mismatch between the behavior and learner policies. How the resulting policy staleness couples with the learning rate to govern training stability and collapse time remains poorly understood. We investigate this coupling in vanilla GRPO through controlled sweeps of the synchronization interval S and constant learning rate η on Llama-3.2-1B/3B, complemented by experiments on Qwen3-8B. We identify two empirical scaling laws: *(i) Stability-boundary scaling:* the largest stable learning rate scales approximately as S^{-1}, yielding a stability boundary characterized by an approximately constant product $S\eta$. *(ii) Collapse-time scaling:* among collapsing runs, estimated collapse times scale approximately as η^{-1}, corresponding to a model- and setup-dependent cumulative learning-rate budget that aligns across synchronization intervals in the Llama sweeps. We interpret these laws through a local analysis of the behavior-dependent GRPO surrogate and a complementary mean-field model. Under local regularity conditions, the analysis yields an $O(S\eta)$ upper bound on the staleness-induced update bias that resets at synchronization. The mean-field model shows how sufficiently strong positive feedback can sustain directional drift when update directions persist across synchronization cycles. When drift speed saturates under optimizer normalization, this mechanism predicts exit from a local surrogate-validity region after an approximately fixed cumulative learning rate. Together, these findings motivate a practical calibration rule: estimate the stability threshold and collapse budget from a coarse sweep, then jointly select S and η for the intended training horizon.

1 INTRODUCTION

Reinforcement learning (RL) is a central component of large language model (LLM) post-training (Ouyang et al., 2022; Lambert et al., 2025), with verifiable rewards enabling substantial advances in reasoning (Guo et al., 2025; Kimi Team, 2025). Long, variable-length rollouts create a systems bottleneck: synchronous pipelines must wait for the slowest response in each batch, leaving computational resources idle. Asynchronous systems alleviate this bottleneck by decoupling rollout generation from policy optimization, allowing generation and learner updates to proceed concurrently (Noukhovitch et al., 2025; Fu et al., 2025; Wu et al., 2025; Piché et al., 2026). The resulting throughput gain comes with *policy staleness*: the learner updates on trajectories generated by an earlier behavior policy. This off-policy mismatch creates a tension between system efficiency and optimization stability.

Existing approaches mitigate staleness by bounding policy lag (Fu et al., 2025; Khatri et al., 2026) or modifying policy updates to accommodate stale trajectories (Li et al., 2026; Zheng et al., 2026; Tian et al., 2026; Xu et al., 2026). Complementary analyses establish sufficient conditions for policy improvement when the behavior and learner policies remain close (Mroueh et al., 2026; Yao et al., 2026). However, lag measured in update steps does not determine policy mismatch on its own: the displacement accumulated during those steps also depends on the learning rate. A lag limit chosen in isolation therefore cannot characterize either the stability of a run or how long an unstable run may improve before collapsing. We ask: *How do staleness and learning rate jointly determine the stability boundary and the time to collapse beyond it?*

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.

We investigate this question in vanilla Group Relative Policy Optimization (GRPO) (Shao et al., 2024), without additional staleness corrections. Under periodic weight synchronization, we vary the synchronization interval S, measured in learner updates, and a constant learning rate η. Controlled sweeps on Llama-3.2-1B/3B reveal two empirical scaling laws: **(i) Stability-boundary scaling.** Within the evaluated training horizons, the stability boundary is approximately characterized by a constant product $S\eta$. The largest stable learning rate therefore scales approximately as S^{-1}: larger synchronization intervals require proportionally smaller learning rates. **(ii) Collapse-time scaling.** Among collapsing runs, the estimated onset time t_{obs}, measured in learner updates, scales approximately as η^{-1}. Equivalently, $t_{\text{obs}}\eta \approx c_2$, where the cumulative learning-rate budget c_2 is model- and setup-dependent but approximately invariant to S within each Llama sweep. Complementary Qwen3-8B experiments at $S = 16$ support this timing relationship with a different fitted budget. These laws distinguish the stability of a configuration from its failure time. Reducing the learning rate can move training below the stability boundary, but it can also merely delay collapse while leaving the configuration unstable. A longer period of improving reward is therefore not, by itself, evidence that instability has been resolved.

To interpret this separation, we analyze the GRPO update field while explicitly retaining its dependence on the behavior policy. Unlike delayed-gradient analyses for a fixed objective (Lian et al., 2015; Zhang et al., 2016; Mishchenko et al., 2022), this formulation accounts for changes in both the sampling distribution and the surrogate weighting. Under local regularity conditions, we derive an $O(S\eta)$ upper bound on the staleness-induced update bias, with a sawtooth envelope that resets at synchronization. A complementary mean-field model connects this local discrepancy to persistent drift: when the bias reinforces recent displacement and update directions persist across synchronization cycles, positive feedback can sustain coherent motion above a threshold controlled by learning rate and average lag. If directional coherence and the characteristic normalized step size saturate (Kingma & Ba, 2015; Balles & Hennig, 2018; Kunstner et al., 2023), the learner advances at an approximately constant speed per unit of cumulative learning rate. Modeling collapse as exit from a local region where the surrogate remains reliable then yields an approximately fixed budget, provided the effective exit geometry remains comparable across synchronization intervals.

These findings motivate joint calibration of staleness and learning rate for a fixed model and training setup. A coarse sweep can bracket the stability boundary and estimate the collapse budget, guiding the selection of (S, η) while distinguishing reduced instability from delayed failure. Selected configurations should then be validated over the intended training horizon.

Contributions.

1. **Two empirical scaling laws.** We characterize an approximately $S\eta$-controlled stability boundary and an inverse-learning-rate collapse clock through controlled GRPO sweeps (Section 4).
2. **Local dynamical analysis.** We derive a local bias bound and develop a conditional model connecting staleness feedback, persistent drift, and optimizer normalization to the two scaling relationships (Section 3).
3. **Joint calibration.** We formulate a rule for selecting synchronization intervals and learning rates using an empirical stability bracket and cumulative learning-rate budget (Section 3.5).

2 Preliminaries

2.1 Group Relative Policy Optimization

Let π_θ be an autoregressive policy with parameters $\theta \in \mathbb{R}^d$, initialized at θ_0. For a prompt $x \sim \mathcal{D}$, GRPO (Shao et al., 2024) samples G responses $y_{1:G} \sim \pi_\phi(\cdot \mid x)$ from a behavior policy π_ϕ, assigns binary verifiable rewards $r_i = r(x, y_i)$, and computes the group-normalized advantage

$$\hat{A}_i = \frac{r_i - \text{mean}(r_{1:G})}{\text{std}(r_{1:G}) + \epsilon_A}, \tag{1}$$

where $\epsilon_A > 0$ stabilizes the normalization. With token-level importance ratios $\rho_{i,l}(\theta, \phi) = \pi_\theta(y_{i,l} \mid x, y_{i,<l})/\pi_\phi(y_{i,l} \mid x, y_{i,<l})$, the standard GRPO surrogate for a group instance $z = (x, y_{1:G})$ is

$$\ell(z; \theta, \phi) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y_i|}\sum_{l=1}^{|y_i|}\Big[\min\big(\rho_{i,l}\hat{A}_i,\ \text{clip}(\rho_{i,l}, 1-\epsilon, 1+\epsilon)\hat{A}_i\big) - \beta\widehat{D}_{\text{KL}}^{i,l}(\theta)\Big], \tag{2}$$

where ϵ is the clipping parameter and β weights the sampled KL penalty relative to the fixed reference policy $\pi_{\text{ref}} = \pi_{\theta_0}$.

Writing p_ϕ for the distribution of group instances induced by $\mathcal{D}$ and π_ϕ, we define the *surrogate-gradient field* and its on-policy counterpart as

$$\mathcal{H}(\theta, \phi) \triangleq \mathbb{E}_{z\sim p_\phi}\big[\nabla_\theta \ell(z; \theta, \phi)\big], \qquad \mathcal{H}_{\text{on}}(\theta) \triangleq \mathcal{H}(\theta, \theta). \tag{3}$$

Differentiation holds the behavior policy and sampled advantages fixed; the diagonal field $\mathcal{H}_{\text{on}}$ need not be the gradient of a fixed scalar objective.

Let θ_t denote the parameters after t learner updates. A mini-batch $\mathcal{B}_t$ sampled from p_{ϕ_t} gives the gradient estimate and parameter update

$$g_t \triangleq \frac{1}{|\mathcal{B}_t|} \sum_{z\in\mathcal{B}_t} \nabla_\theta \ell(z; \theta_t, \phi_t), \qquad \theta_{t+1} = \theta_t + \eta_t u_t, \tag{4}$$

where u_t is the optimizer update in ascent notation (AdamW in our experiments). With $\mathcal{F}_t$ denoting the history before sampling $\mathcal{B}_t$, the sampling model gives $\mathbb{E}[g_t \mid \mathcal{F}_t] = \mathcal{H}(\theta_t, \phi_t)$. The idealized on-policy reference sets $\phi_t = \theta_t$, recovering $\mathcal{H}_{\text{on}}(\theta_t)$. We write $\Lambda_t \triangleq \sum_{s<t} \eta_s$ for the *cumulative learning rate*; our experiments use a constant rate η, so $\Lambda_t = t\eta$.

2.2 Staleness and Collapse

Stale rollouts. Asynchronous pipelines decouple rollout generation from policy optimization. We consider periodic weight synchronization every S learner updates, holding the behavior policy fixed between synchronizations. Abstracting away additional queueing delays, the behavior parameters at step t are

$$\phi_t \triangleq \theta_{t-k_t}, \qquad k_t \triangleq t - S\lfloor t/S \rfloor \in \{0, \dots, S-1\}. \tag{5}$$

Here S is the *synchronization interval*, while the *policy lag* k_t grows within each cycle, resets at synchronization, and has cycle average $\bar{k} = (S-1)/2$. The case $S = 1$ recovers the idealized on-policy reference. The resulting parameter discrepancy accumulates the intervening optimizer updates:

$$\theta_t - \phi_t = \sum_{j=t-k_t}^{t-1} \eta_j u_j. \tag{6}$$

Thus, lag counts updates, whereas the parameter discrepancy also depends on their learning rates and directions. The deviation $\mathcal{H}(\theta_t, \phi_t) - \mathcal{H}_{\text{on}}(\theta_t)$ is the *staleness bias* analyzed in Section 3.

Collapse. We define *training collapse* as an abrupt, severe drop in mean verifier reward on training rollouts, with validation accuracy providing corroborating evidence. Under a fixed prompt distribution and decoding configuration, each behavior-policy snapshot has a fixed expected reward, although empirical batch rewards fluctuate. We report a *policy-indexed collapse time* in learner-update units:

$$t_{\text{obs}} \triangleq S M_{\text{collapse}}, \tag{7}$$

where M_{collapse} indexes the first synchronized behavior-policy snapshot exhibiting this drop, with initialization indexed by zero. A run is *stable* if no collapse is observed within its training horizon.

3 Understanding the Scaling Laws

We interpret *stability-boundary scaling* and *collapse-time scaling* through the within-cycle policy discrepancy, the coherence of the resulting drift, and the distance traveled per unit of cumulative learning rate. The analysis separates a local property of the GRPO field from a conditional dynamical model: regularity controls the magnitude of stale feedback, while its sign, persistence, and normalized speed determine whether that feedback can produce a collapse clock. All norms below are per-coordinate RMS norms, $\|v\| = \|v\|_2/\sqrt{d}$, with inner product $\langle v, w\rangle_d = v^\top w/d$. We distinguish the update bound $U_{\max}$ from the characteristic normalized coordinate step $u_\star$; neither is assumed to equal one.

3.1 Staleness-Induced Bias

The behavior-dependent field in equation 3 gives

$$g_t = \mathcal{H}_{\text{on}}(\theta_t) + \delta_t + \xi_t, \qquad \delta_t = \mathcal{H}(\theta_t, \phi_t) - \mathcal{H}_{\text{on}}(\theta_t), \tag{8}$$

where $\xi_t = g_t - \mathcal{H}(\theta_t, \phi_t)$ satisfies $\mathbb{E}[\xi_t \mid \mathcal{F}_t] = 0$. In the periodic-synchronization model, $\delta_t = 0$ whenever $k_t = 0$.

Assumption 1 (Local regularity). *On a working neighborhood $\mathcal{U}$, the relevant gradients and updates satisfy $\|\nabla_\theta \ell(z; \theta, \phi)\| \le \Gamma$ and $\|u_t\| \le U_{\max}$. Moreover, $D_{\text{KL}}(p_\phi \| p_{\phi'}) \le C_\pi \|\phi - \phi'\|^2$ and $\|\mathbb{E}_{p_\phi}[\nabla_\theta \ell(z; \theta, \phi) - \nabla_\theta \ell(z; \theta, \phi')]\| \le L_\rho \|\phi - \phi'\|$.*

Proposition 1 (A resettable bias envelope). *Under Assumption 1, let $C_{\text{stale}} = L_\rho + \Gamma\sqrt{2C_\pi}$. While the relevant states remain in $\mathcal{U}$,*

$$\|\delta_t\| \le C_{\text{stale}} \|\theta_t - \phi_t\| \le C_{\text{stale}} U_{\max} \sum_{j=t-k_t}^{t-1} \eta_j. \tag{9}$$

Thus, at constant η, the bias has a sawtooth upper envelope *proportional to $k_t\eta$, with cycle average at most $C_{\text{stale}} U_{\max} \bar{k}\eta$.*

Proof sketch and scope. Adding and subtracting the on-policy integrand under p_{ϕ_t} splits δ_t into a surrogate-weighting shift and a sampling shift. Their norms are bounded by $L_\rho \|\theta_t - \phi_t\|$ and $\Gamma\sqrt{2C_\pi}\|\theta_t - \phi_t\|$, respectively, the latter by Pinsker's inequality. Summing the intervening updates gives equation 9 (Appendix A.2). With a locally Lipschitz behavior-policy Jacobian, $\delta_t = B_t(\theta_t - \phi_t) + O(\|\theta_t - \phi_t\|^2)$, where $B_t = -\partial_\phi \mathcal{H}(\theta_t, \theta_t)$. The bound controls magnitude, not sign: group normalization can reinforce displacement, while other GRPO terms can oppose it. Appendix A.3 isolates these contributions; positive net feedback remains a separate condition.

3.2 Stability-Boundary Scaling

Consider a coherent block $\mathcal{C}$ containing a fraction f of coordinates, with fixed signs $\chi_j \in \{\pm 1\}$ and common raw-noise scale $\sigma_{\mathcal{C}}$. Its RMS-unit direction is $e = \chi/\sqrt{f}$, with $\chi = 0$ outside $\mathcal{C}$. A normalized signed coordinate step $u_\star s$ then gives drift speed $\sqrt{f}\, u_\star s$ along e. If the stale response reinforces displacement with coefficient $b > 0$, its mean-lag signal-to-noise ratio is $b\bar{k}\eta u_\star s/\sigma_{\mathcal{C}}$.

Assumption 2 (Persistent mean-field feedback). *The drift direction persists across synchronization cycles, and its cycle-level coherence obeys the mean-lag closure*

$$s_{n+1} = \Phi(\gamma s_n + \rho_n), \qquad \gamma = \frac{b u_\star}{\sigma_{\mathcal{C}}} \bar{k}\eta, \qquad 0 \le s_n \le 1, \tag{10}$$

where $\rho_n \ge 0$ tends to zero. The response Φ is odd, increasing, L_Φ-Lipschitz, $\Phi'(0) > 0$, and $0 < \Phi(r) < 1$ for $r > 0$, with $\Phi(r) \to 1$.

The gain γ measures stale signal relative to raw-gradient noise. For example, replacing a coordinate's second moment by its stationary value gives $\Phi(r) = r/\sqrt{1 + r^2}$: its normalized mean saturates even as the signal grows. This motivates the bounded response, not an exact identity for AdamW. The closure also requires directional memory across resets; neither condition follows from the bias bound (Appendix A.6).

Proposition 2 (Mean-field stability boundary). *Define*

$$\gamma_c = \inf_{r>0} \frac{r}{\Phi(r)}, \qquad c_1 = \frac{\sigma_{\mathcal{C}} \gamma_c}{b u_\star}. \tag{11}$$

For $\gamma < \gamma_c$, the zero-signal map has no positive fixed point and $s_n \to 0$. For $\gamma > \gamma_c$, its largest fixed point $s_+(\gamma) > 0$ is nondecreasing and tends to one as $\gamma \to \infty$. If $s_{n_0} \ge s_+(\gamma)$, then $s_n \ge s_+(\gamma)$ thereafter and $s_n \to s_+(\gamma)$. The mean-field boundary is therefore

$$\bar{k}\eta = c_1, \qquad \eta_{\text{mf}}(S) = \frac{2c_1}{S - 1}, \quad S > 1. \tag{12}$$

Origin of the threshold. At a positive fixed point, $r = \gamma s$ gives $\gamma = r/\Phi(r)$, excluding such points below γ_c. Above γ_c, some $s \in (0, 1)$ satisfies $\Phi(\gamma s) > s$; since $\Phi(\gamma) < 1$, continuity gives a positive root. For comparable response parameters across S, the resulting boundary yields law (i), $\eta_{\mathrm{mf}}(S) \propto (S-1)^{-1} \approx S^{-1}$, conditional on entry into the persistent branch. Convergence, onset types, and finite-S corrections are treated in Appendix A.4.

3.3 Collapse-Time Scaling

Snapshot exit model. To connect drift to the reward-defined event, introduce a local surrogate-validity region $\mathcal{N}_R = \{\theta : \|\theta - \theta_0\| < R\}$, $R > 0$, and its first *synchronized-snapshot exit*,

$$\tau_S \triangleq \inf\{mS : \theta_{mS} \notin \mathcal{N}_R\}. \tag{13}$$

The modeling correspondence is $t_{\mathrm{obs}} = \tau_S$: the first failed behavior snapshot is the first snapshot outside $\mathcal{N}_R$. This is distinct from the empirical collapse criterion and avoids identifying delayed reward logs with learner-update times.

From coherence to parameter motion. Write $a_t = \mathbb{E}[u_t \mid \mathcal{F}_t]$ and $D_t = \sum_{j<t} \eta_j a_j$ for cumulative conditional *optimizer* drift. At constant η, the cycle-average conditional mean $\bar{a}_n = S^{-1} \sum_{k=0}^{S-1} a_{nS+k}$ connects the scalar coherence to parameter motion through

$$\bar{a}_n = \sqrt{f}\, u_\star s_n e + r_n, \qquad v = \sqrt{f}\, u_\star s_{\mathrm{sat}}. \tag{14}$$

Here r_n collects motion outside the coherent block, directional variation, and optimizer effects omitted by the closure. If s_n approaches s_{sat} and residual motion is small, summing cycles yields the cumulative-drift condition below; the initial transient and incomplete cycles enter a. Appendix A.5 quantifies this bridge from coherence to approximately constant *mean velocity*.

For deterministic nonnegative learning rates and a horizon $T \in S\mathbb{N}$, let $V_T = \sum_{j<T} \eta_j^2$ and $h_S = \max_{0 \le m < T/S} \sum_{j=mS}^{(m+1)S-1} \eta_j$. For $\delta \in (0, 1)$, a uniform martingale bound gives the noise radius $q_T(\delta) = 2U_{\max}\sqrt{2V_T \log(2/\delta)}$ (Appendix A.5).

Theorem 1 (Snapshot-clock bounds)**.** *Suppose $\|u_t\| \le U_{\max}$ before $T \wedge \tau_S$. For an RMS-unit direction e and constants $v > 0$, $a \ge 0$, and $0 \le \varepsilon < 1$, assume*

$$\|D_t - v\Lambda_t e\| \le a + \varepsilon v \Lambda_t, \qquad t \le T \wedge \tau_S. \tag{15}$$

Let $E_T = a + q_T(\delta) < R$ and assume $(1-\varepsilon)v\Lambda_T \ge R + E_T$. Then, with probability at least $1 - \delta$, $\tau_S \le T$ and

$$\frac{R - E_T}{(1+\varepsilon)v} \le \Lambda_{\tau_S} \le \frac{R + E_T}{(1-\varepsilon)v} + h_S. \tag{16}$$

Proof sketch. Decompose $\theta_t - \theta_0 = D_t + N_t$ into conditional drift and centered-update fluctuations. A uniform martingale bound controls $\|N_t\| \le q_T(\delta)$ for all $t \le T \wedge \tau_S$, giving

$$(1-\varepsilon)v\Lambda_t - E_T \le \|\theta_t - \theta_0\| \le (1+\varepsilon)v\Lambda_t + E_T. \tag{17}$$

The upper envelope prevents an earlier exit; the lower envelope forces exit by its first scheduled crossing of R, with at most h_S additional cumulative learning rate. Appendix A.5 proves both bounds on the same uniform event, including at the random exit time.

Saturation removes lag from the leading clock. At constant learning rate, $h_S = S\eta$, so

$$\eta t_{\mathrm{obs}} = c_2\left[1 + O\left(\varepsilon + \frac{a + q_T(\delta)}{R}\right)\right] + O(S\eta), \qquad c_2 = \frac{R}{\sqrt{f}\, u_\star s_{\mathrm{sat}}}. \tag{18}$$

When the effective exit distance and saturated speed are comparable across S, this yields law (ii), with explicit transient, noise, and synchronization-resolution errors. Before saturation, the leading budget instead scales as $R/(\sqrt{f}\, u_\star s_+(\gamma))$ and can depend on S. Thus a shared clock reflects a nearly constant distance-to-speed ratio, not the existence of a coherent branch alone. Exact snapshot quantization and its resolution error are given in Appendix A.5.

Why stable runs can outlive the clock. If cumulative mean drift instead remains bounded by $a_0 < R$, the same concentration argument prevents snapshot exit whenever $a_0 + q_T(\delta) < R$. At constant η and fixed confidence and margin, this permits an $\Omega(\eta^{-2})$ non-exit horizon rather than the coherent η^{-1} clock. Appendix A.5 gives the explicit bound and sufficient conditions controlling residual drift; decay of scalar coherence alone is not enough.

3.4 Interpreting the Gradient Cosine

Let $\mu_t = \mathcal{H}(\theta_t, \phi_t)$, $\sigma_{t,j}^2 = \mathbb{E}[\xi_{t,j}^2 \mid \mathcal{F}_t] > 0$, $\rho_{t,j} = \mu_{t,j}/\sigma_{t,j}$, and $w_{t,j} = \sigma_{t,j}^2 / \sum_i \sigma_{t,i}^2$. When adjacent means vary slowly and the relevant norms and inner products concentrate, Appendix A.7 gives

$$c_t \triangleq \cos(g_t, g_{t-1}) \approx \frac{\mathrm{SNR}_t}{1 + \mathrm{SNR}_t}, \qquad \mathrm{SNR}_t = \sum_j w_{t,j} \rho_{t,j}^2. \tag{19}$$

Unlike coordinate-normalized updates, this readout weights each coordinate by its share of raw-gradient noise.

A block can therefore sustain coherent normalized updates while contributing little to the raw cosine (Proposition 3, Appendix A.7). A subsequent rise additionally requires a stronger raw-gradient signal near failure. A thin high-SNR boundary layer can yield a short warning interval, conditional on the readout and trajectory assumptions; its bound is given in Appendix A.7. Thus a low raw-gradient cosine alone does not certify stability.

3.5 Practical Calibration

The two laws concern different aspects of a run: $\bar{k}\eta$ controls the feedback regime, while R/v controls the failure clock on a persistent branch. A coarse sweep can bracket the empirical stability boundary and estimate a cumulative-rate budget from several failed snapshots. Increasing S therefore calls for an approximately inverse change in η, followed by validation at the intended horizon. The budget is not a universal limit on stable training, nor does a single failed run determine it exactly. For varying learning rates, the clock interpretation applies only while the drift conditions persist; changing η can also change the feedback regime. For $S = 1$, synchronization-induced bias vanishes, but other sources of optimization failure remain possible.

4 Experiments

We examine the collapse laws through joint (S, η) sweeps on two Llama models, complemented by extended-horizon stability tests and a cross-family study of collapse timing.

4.1 Experimental Setup

Models and data. We train Llama-3.2-1B-Instruct and Llama-3.2-3B-Instruct (Meta AI, 2024) on the 7,500-problem training split of MATH (Hendrycks et al., 2021) and evaluate on its 5,000-problem test split. For Qwen3-8B (Yang et al., 2025), we train on DAPO-Math-17k (Yu et al., 2025) and report avg@1 on MATH-500 (Lightman et al., 2024) and Minerva Math (Lewkowycz et al., 2022), and avg@8 on AMC 2023/2024 (Art of Problem Solving, 2024).

Protocol. We use the asynchronous pipeline of Section 2.2, implemented in verl (Sheng et al., 2025) with vLLM (Kwon et al., 2023) rollout workers synchronized every S learner updates. All runs use vanilla GRPO without additional staleness correction (Appendix B). The main experiment crosses $S \in \{8, 16, 32\}$ with constant $\eta \in \{10^{-6}, 5 \times 10^{-7}, 2 \times 10^{-7}, 10^{-7}\}$ for both Llama models. We record the collapse index M_{collapse} and policy-indexed time $t_{\mathrm{obs}} = SM_{\mathrm{collapse}}$, and compute the gradient cosine c_t following Xu et al. (2026). Main-sweep runs stop after collapse is identified or after a collapse-free horizon exceeding the empirical reference clock c_2/η. The main sweeps use one seed per configuration; targeted 3B repeats at $S = 32$ are reported in Appendix C.

Table 1: Collapse indices M_{collapse} for the Llama sweeps and Qwen3-8B extension. "Stable" denotes no observed collapse within the training horizon; t_{obs} rows give the corresponding learner-update times for collapsing configurations.

Model	S	$\eta = 10^{-6}$	5×10^{-7}	2×10^{-7}	10^{-7}
Llama-3.2-1B / 3B	8	4	8	stable	stable
	16	2	4	10	stable
	32	1	2	5	10
	t_{obs}	32	64	160	320
Qwen3-8B	16	4	8	20	stable
	t_{obs}	64	128	320	—

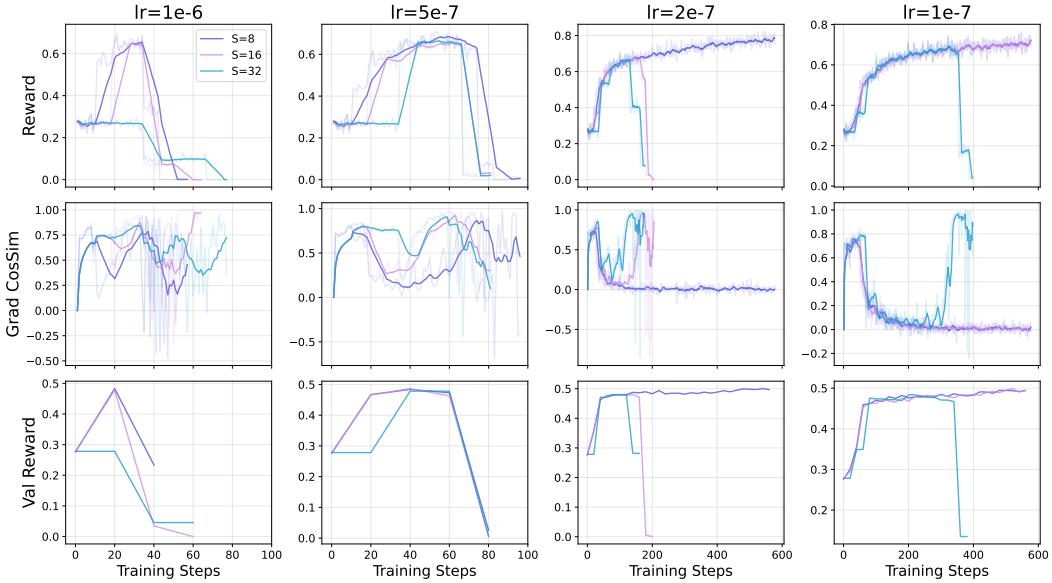


Figure 1: Llama-3.2-3B-Instruct (S, η) sweep. Columns vary η; curves vary S. Rows show training reward, gradient cosine c_t, and MATH validation reward. On the tested grid, runs with $S\eta \leq 1.6 \times 10^{-6}$ remain stable; others have reported collapse times $t_{\text{obs}} = 3.2 \times 10^{-5}/\eta$, independent of S (Table 1). At $\eta \geq 5 \times 10^{-7}$, c_t remains elevated through much of training; near the boundary, collapsing runs can initially resemble stable ones before c_t rises near collapse. The 1B sweep has identical collapse indices (Figure 6). Horizontal-axis ranges are column-specific.

4.2 A Stability Boundary in $S\eta$

Both Llama sizes exhibit identical stability patterns (Table 1), with joint dynamics in Figures 1 and 6 and per-configuration reward traces in Appendix E. Reducing η expands the stable region: all three synchronization intervals collapse at $\eta \geq 5 \times 10^{-7}$; only $S = 8$ is stable at 2×10^{-7}; and both $S = 8$ and 16 are stable at 10^{-7}, while $S = 32$ still collapses. In the 3B sweep, unstable runs can initially improve training and validation reward before losing these gains, whereas stable runs sustain them. Across the tested grid, stable configurations satisfy $S\eta \leq 1.6 \times 10^{-6}$ and collapsing configurations have $S\eta \geq 3.2 \times 10^{-6}$. Thus, doubling S from 8 to 16 halves the largest tested stable learning rate, supporting the approximately constant-$S\eta$ boundary predicted by Proposition 2.

4.3 A Cumulative-Learning-Rate Clock for Collapse

All nine collapsing configurations for each Llama size satisfy $t_{\text{obs}}\eta = c_2 = 3.2 \times 10^{-5}$. The reported collapse times are 32, 64, 160, and 320 updates for $\eta = 10^{-6}, 5 \times 10^{-7}, 2 \times 10^{-7}, 10^{-7}$, respectively.

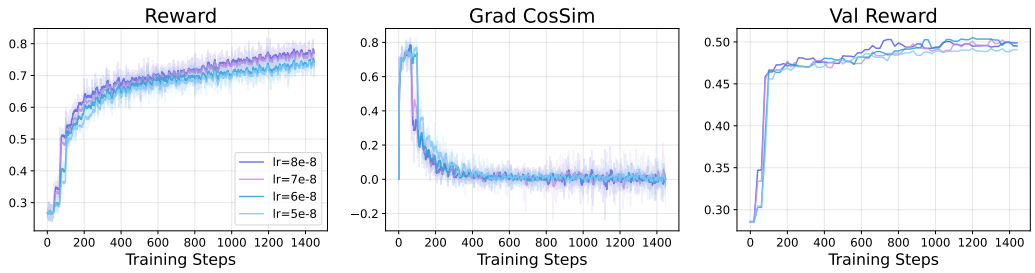


Figure 2: Extended-horizon Llama-3.2-3B-Instruct runs at $S = 32$ with $\eta \in \{8, 7, 6, 5\} \times 10^{-8}$. Left to right: training reward, gradient cosine c_t, and MATH validation reward. All four runs remain collapse-free for approximately $1{,}400$ learner updates, exceeding twice their reference times c_2/η, with c_t settling near zero.

Within each learning-rate column, these times are unchanged across the values of S that collapse. For example, at $\eta = 10^{-6}$, $S = 8, 16, 32$ yield collapse indices $4, 2, 1$, respectively, but all correspond to $t_{\text{obs}} = 32$. Larger staleness therefore reduces the number of synchronizations before collapse without shortening the recorded learner-update horizon. Even configurations at $S\eta = 3.2 \times 10^{-6}$ share this clock with settings with up to ten times larger $S\eta$. For the 3B model at $S = 32$, all three tested seeds reproduce the same four recorded collapse indices (Table 3). At the policy-snapshot resolution, these observations agree with Proposition 2 and Theorem 1 (equation 18): $S\eta$ separates the regimes, while the cumulative learning rate governs collapse timing within the unstable regime.

4.4 Gradient Alignment as a Late Warning

The cosine trajectories show distinct temporal patterns (Figure 1, middle row). Stable runs can exhibit an initial alignment transient during the reward increase, after which c_t approaches zero as reward remains high. At $\eta \geq 5 \times 10^{-7}$, collapsing runs instead show elevated, though fluctuating, cosine over much of their short trajectories. Near the boundary, at $S = 32$ and $\eta = 10^{-7}$, c_t first decays toward the stable-run level, then rises near the reward drop across all three tested seeds (Figure 4). Thus, early alignment need not signal failure, and a prolonged period of low cosine does not establish stability. The late rise is consistent with Proposition 3 and Corollary 2 (Appendix A.7), but offers limited warning near the boundary.

4.5 Extended-Horizon Stability

To distinguish stability from delayed collapse, we train the 3B Llama model at $S = 32$ with $\eta \in \{8, 7, 6, 5\} \times 10^{-8}$ for approximately $1{,}400$ learner updates. All four runs remain collapse-free (Figure 2): training reward continues to improve after its initial rise, validation reward settles near 0.5, and c_t approaches zero without a sustained late resurgence. Each run exceeds twice its reference time c_2/η, whereas $\eta = 10^{-7}$ at the same S collapses after 320 updates. This survival is not explained by extrapolating the unstable-regime clock to smaller learning rates. Separate three-seed checks remain collapse-free through 500 updates for all four rates (Figure 5), exceeding the reference clock at 8×10^{-8} and 7×10^{-8}. Stability at $\eta = 8 \times 10^{-8}$ and collapse at $\eta = 10^{-7}$ further bracket the transition between $S\eta = 2.56 \times 10^{-6}$ and 3.2×10^{-6}. Survival beyond the reference clock is consistent with the conditional non-collapse bound of Corollary 1.

4.6 A Fixed-Staleness Test on Qwen3-8B

We test Qwen3-8B at $S = 16$ using the main learning-rate grid. The three larger rates show rapid early reward gains followed by abrupt drops and declining validation accuracy (Figure 3). Their policy-indexed collapse times, $t_{\text{obs}} = 64, 128, 320$, satisfy $t_{\text{obs}}\eta = 6.4 \times 10^{-5}$, twice the Llama calibration (Table 1). By contrast, $\eta = 10^{-7}$ learns more gradually but remains collapse-free for approximately 800 updates, beyond its reference time of 640, with training reward and validation accuracy reaching about 0.65 and 0.67, respectively. At $\eta = 2 \times 10^{-7}$, the cosine stays near the

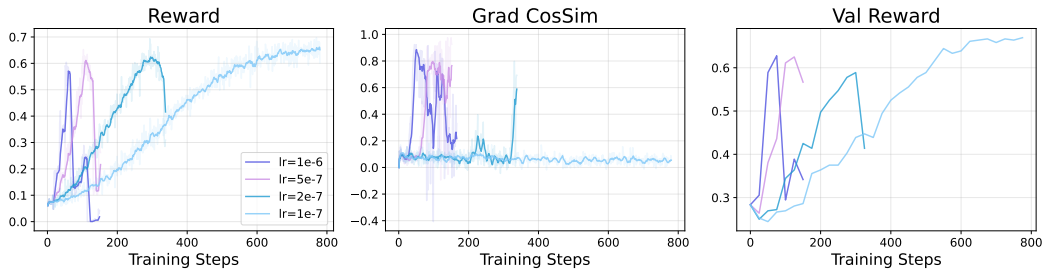


Figure 3: Qwen3-8B at $S = 16$, with one curve per learning rate. Left to right: training reward, raw-gradient cosine c_t, and mean validation accuracy. The $\eta = 10^{-7}$ run remains collapse-free over the displayed horizon, whereas the three larger learning rates have reported collapse times satisfying $t_{\text{obs}}\eta = 6.4 \times 10^{-5}$ (Table 1). Near the boundary ($\eta = 2 \times 10^{-7}$), c_t stays near the stable-run level before rising sharply around collapse.

stable-run level until rising near collapse, reproducing the late-warning pattern. This experiment supports inverse-learning-rate timing in a different model–data setting with a different clock constant; independence from S is tested by the Llama sweeps.

5 Related Work

Staleness in asynchronous RL for LLMs. Recent systems improve throughput by decoupling rollout generation from learning (Noukhovitch et al., 2025; Fu et al., 2025; Piché et al., 2026; Song et al., 2026). They manage policy lag through empirical bounds (Fu et al., 2025; Khatri et al., 2026) or objective-level corrections (Li et al., 2026; Zheng et al., 2026; Tian et al., 2026; Xu et al., 2026). We retain vanilla GRPO and characterize how admissible staleness depends on the learning rate, providing a baseline for evaluating such corrections.

Guarantees for delayed and off-policy updates. Distributed RL mitigates policy mismatch through importance weighting (Munos et al., 2016; Espeholt et al., 2018). Delayed-SGD and off-policy GRPO analyses relate convergence or improvement to delay, step size, and policy proximity (Lian et al., 2015; Mishchenko et al., 2022; Mroueh et al., 2026; Yao et al., 2026). We complement these guarantees with an approximately constant-$S\eta$ stability boundary and a $1/\eta$ collapse clock, insensitive to S across the tested unstable configurations.

Optimizer dynamics. Adam's coordinatewise normalization can reduce sensitivity to gradient magnitude (Balles & Hennig, 2018; Kunstner et al., 2023), while transformer gradient scales vary by parameter block (Zhang et al., 2024). Our feedback model combines these properties to explain the cumulative-learning-rate clock and late gradient-cosine rise near collapse (Appendix D).

6 Conclusion

We identify two empirical scaling laws that distinguish the stability of asynchronous GRPO from its collapse timing. In the Llama sweeps, the stability boundary is approximately characterized by constant $S\eta$, while collapsing runs follow an inverse-learning-rate clock with an approximately shared cumulative-rate budget across synchronization intervals. Qwen3-8B supports the timing relationship at fixed S with a different budget. Extended-horizon runs remain collapse-free beyond the fitted clock, while gradient cosine provides only a late warning near the boundary. Our local $O(S\eta)$ bias bound and conditional dynamical model connect these observations to staleness feedback, persistent drift, and optimizer normalization. The practical implication is to calibrate S and η jointly for each model and training setup, then validate the selected configuration over the intended horizon rather than mistaking delayed collapse for stability.

Reproducibility Statement

Section 2.2 defines the synchronization protocol, collapse criterion, and timing convention; Section 4.1 specifies the models, data, and experimental grid. Appendix B lists all training and evaluation settings, stopping rules, and the gradient-cosine measurement. Multi-seed checks and additional curves are in Appendices C and E, and all assumptions and proofs are in Appendix A. All models and datasets are publicly available.

Ethics Statement

This work studies the optimization stability of reinforcement learning for language models on public mathematical-reasoning benchmarks. It involves no human subjects, personal data, or newly collected data, and it introduces no new model capabilities. By helping practitioners avoid collapsing configurations, the findings may reduce wasted computation.

AI Use Statement

Large language models assisted with literature search, drafting and editing, checking derivations, and LaTeX formatting. The authors designed and ran all experiments, verified all AI-assisted content, and take full responsibility for the paper.

References

Art of Problem Solving. AMC problems and solutions. https://artofproblemsolving.com/wiki/index.php/AMC_Problems_and_Solutions, 2024.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients, 2018. URL https://openreview.net/forum?id=S1EwLkW0W.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In *ICML*, pp. 559–568, 2018. URL http://proceedings.mlr.press/v80/bernstein18a.html.

Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=jh-rTtvkGeM.

Sanghamitra Dutta, Gauri Joshi, Soumyadip Ghosh, Parijat Dube, and Priya Nagpurkar. Slow and stale gradients can win the race: Error-runtime trade-offs in distributed SGD. In *AISTATS*, pp. 803–812, 2018. URL http://proceedings.mlr.press/v84/dutta18a.html.

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 *ICML*, pp. 1406–1415, 2018. URL http://proceedings.mlr.press/v80/espeholt18a.html.

Lasse Espeholt, Raphaël Marinier, Piotr Stanczyk, Ke Wang, and Marcin Michalski. SEED RL: Scalable and efficient Deep-RL with accelerated central inference. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkgvXlrKwH.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. AReaL: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=X9diEuva9R.

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=7Bywt2mQsCe.

Jacob Hilton, Karl Cobbe, and John Schulman. Batch size-invariance for policy optimization. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=lXuZaxEaI7.

Jian Hu, Xibin Wu, Wei Shen, Jason Klein Liu, Weixun Wang, Songlin Jiang, Haoran Wang, Hao Chen, Bin Chen, Wenkai Fang, Xianyu, Yu Cao, Haotian Xu, and Yiming Liu. OpenRLHF: A Ray-based easy-to-use, scalable and high-performance RLHF framework. In *EMNLP (System Demonstrations)*, pp. 656–666, 2025. URL https://doi.org/10.18653/v1/2025.emnlp-demos.48.

Shengyi Huang, Jiayi Weng, Rujikorn Charakorn, Min Lin, Zhongwen Xu, and Santiago Ontanon. Cleanba: A reproducible and efficient distributed reinforcement learning platform. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Diq6urt3lS.

Devvrit Khatri, Lovish Madaan, Rishabh Tiwari, Rachit Bansal, Sai Surya Duvvuri, Manzil Zaheer, Inderjit S. Dhillon, David Brandfonbrener, and Rishabh Agarwal. The art of scaling reinforcement learning compute for LLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=FMjeC9Msws.

Kimi Team. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025. URL https://arxiv.org/abs/2501.12599.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Anastasia Koloskova, Sebastian U Stich, and Martin Jaggi. Sharper convergence guarantees for asynchronous SGD for distributed and federated learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4_oCZgBIVI.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between SGD and Adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, pp. 611–626, 2023. URL https://doi.org/10.1145/3600006.3613165.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James Validad Miranda, Alisa Liu, Nouha Dziri, Xinxi Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Christopher Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=i1uGbfHHpH.

Aitor Lewkowycz, Anders Johan Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Venkatesh 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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=IFXTZERXdM7.

Xiaocan Li, Zheng Shen, and Shiliang Wu. A-3PO: Accelerating asynchronous LLM training with staleness-aware proximal policy approximation. In *The 1st Workshop on Scaling Post-training for LLMs*, 2026. URL https://openreview.net/forum?id=CgtMJGtYnr.

Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *ICLR*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Jian Lu and Yi Luo. Periodic asynchrony: An on-policy approach for accelerating LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2511.18871.

Sadhika Malladi, Kaifeng Lyu, Abhishek Panigrahi, and Sanjeev Arora. On the SDEs and scaling rules for adaptive gradient algorithms. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=F2mhzjHkQP.

Meta AI. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024. Accessed: 2026-09-26.

Konstantin Mishchenko, Francis Bach, Mathieu Even, and Blake Woodworth. Asynchronous SGD beats minibatch SGD under arbitrary delays. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=4XP0ZuQKXmV.

Ioannis Mitliagkas, Ce Zhang, Stefan Hadjis, and Christopher Ré. Asynchrony begets momentum, with an application to deep learning. In *Allerton*, pp. 997–1004, 2016. URL https://doi.org/10.1109/ALLERTON.2016.7852343.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *ICML*, pp. 1928–1937, 2016. URL http://proceedings.mlr.press/v48/mniha16.html.

Youssef Mroueh, Nicolas Dupuis, Brian Belgodere, Apoorva Nitsure, Mattia Rigotti, Kristjan Greenewald, Jiri Navratil, Jarret Ross, and Jesus Rios. Revisiting group relative policy optimization: Insights into on-policy and off-policy training. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XXfOf22o3K.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G. Bellemare. Safe and efficient off-policy reinforcement learning. In *NIPS*, pp. 1046–1054, 2016. URL http://papers.nips.cc/paper/6538-safe-and-efficient-off-policy-reinforcement-learning.

Feng Niu, Benjamin Recht, Christopher Re, and Stephen J. Wright. HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent, 2011. URL https://arxiv.org/abs/1106.5730.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, Arian Hosseini, Rishabh Agarwal, and Aaron Courville. Faster, more efficient RLHF through off-policy asynchronous learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FhTAG591Ve.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Alexandre Piché, Ehsan Kamalloo, Rafael Pardinas, Xiaoyin Chen, and Dzmitry Bahdanau. PipelineRL: Faster on-policy reinforcement learning for long sequence generation. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=A35ak14Cyp.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *Ann. Probab.*, 22(4):1679–1706, 1994. URL http://dml.mathdoc.fr/item/1176988477.

James Queeney, Ioannis Paschalidis, and Christos Cassandras. Generalized proximal policy optimization with sample reuse. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=in_RVSTqYxK.

Nicolas Le Roux, Marc G Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alexandre Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Tóth, and Sam Work. Tapered off-policy REINFORCE - stable and efficient reinforcement learning for large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gFFgCWiXWI.

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, 2024. URL 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*, EuroSys '25, pp. 1279–1297. ACM, March 2025. doi: 10.1145/3689031.3696075. URL http://dx.doi.org/10.1145/3689031.3696075.

Jingwei Song, Meng Chen, Jie Xiao, Qingnan Ren, Jiaqi Huang, Yangshen Deng, Chris Tong, Wanyi Chen, Suli Wang, Zhisheng Chen, Ziqian Bi, Shuo Lu, Yiqun Duan, Xu Wang, Rymon Yu, Lynn Ai, Eric Yang, and Tianyu Shi. ECHO-2: A large-scale distributed rollout framework for cost-efficient reinforcement learning, 2026. URL https://arxiv.org/abs/2602.02192.

Minghao Tian, Yunfei Xie, and Chen Wei. How off-policy can GRPO be? Mu-GRPO for efficient LLM reinforcement learning, 2026. URL https://arxiv.org/abs/2605.17570.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning near-perfect PointGoal navigators from 2.5 billion frames. In *ICLR*, 2020. URL https://openreview.net/forum?id=H1gX8C4YPr.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie E Everett, Alexander A Alemi, Ben Adlam, John D Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=d8w0pmvXbZ.

Bo Wu, Sid Wang, Yunhao Tang, Jia Ding, Eryk Helenowski, Liang Tan, Tengyu Xu, Tushar Gowda, Zhengxing Chen, Chen Zhu, Xiaocheng Tang, Yundi Qian, Beibei Zhu, and Rui Hou. LlamaRL: A distributed asynchronous reinforcement learning framework for efficient large-scale LLM training, 2025. URL https://arxiv.org/abs/2505.24034.

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. URL https://arxiv.org/abs/2603.01501.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Chaorui Yao, Yanxi Chen, Yuchang Sun, Yushuo Chen, Wenhao Zhang, Xuchen Pan, Yaliang Li, and Bolin Ding. Group-relative REINFORCE is secretly an off-policy algorithm: Demystifying some myths about GRPO and its friends. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7CFlXvCoN6.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=2a36EMSSTp.

Wei Zhang, Suyog Gupta, Xiangru Lian, and Ji Liu. Staleness-aware Async-SGD for distributed deep learning. In *IJCAI*, pp. 2350–2356, 2016. URL http://www.ijcai.org/Abstract/16/335.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need Adam: A Hessian perspective. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=X6rqEpbnj3.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=IIgl5MWelz.

Yinmin Zhong, Zili Zhang, Xiaoniu Song, Hanpeng Hu, Chao Jin, Bingyang Wu, Nuo Chen, Yukun Chen, Yu Zhou, Changyi Wan, Hongyu Zhou, Yimin Jiang, Yibo Zhu, and Daxin Jiang. StreamRL: Scalable, heterogeneous, and elastic RL for LLMs with disaggregated stream generation, 2025. URL https://arxiv.org/abs/2504.15930.

A Theory Details and Proofs

This appendix separates three logical steps. The local bias bound is a consequence of regularity of the GRPO field. The feedback threshold is a property of an explicitly specified mean-field model. The collapse clock follows from concentration around cumulative *optimizer* drift, together with a reward-to-snapshot-exit correspondence. None of the latter assumptions is inferred solely from agreement between the tabulated collapse times and a fitted clock. Throughout, $\|v\| = \|v\|_2/\sqrt{d}$, $\langle v, w\rangle_d = v^\top w/d$, and matrix operator norms are induced by this norm. Ordinary dot products and $\|\cdot\|_2$ are used when stated explicitly.

A.1 Geometry, Conditional Means, and the Observation Model

The quantities being separated. The raw-gradient mean is $\mu_t = \mathcal{H}(\theta_t, \phi_t)$, while the conditional mean of the applied optimizer update is $a_t = \mathbb{E}[u_t \mid \mathcal{F}_t]$. These are different for a nonlinear, stateful optimizer. The first determines the bias δ_t and raw-gradient readout; the second determines parameter displacement. The pre-sampling filtration $\mathcal{F}_t$ includes the learner, behavior policy, and optimizer state. The sampling model in the preliminaries assumes a fresh mini-batch conditional on this state; using cached or adaptively selected data would require a corresponding change to the conditional field.

RMS geometry. For $\mathcal{C} \subseteq \{1, \dots, d\}$ with $|\mathcal{C}| = fd$, let $\chi_j \in \{\pm 1\}$ on $\mathcal{C}$ and $\chi_j = 0$ elsewhere. Then $e = \chi/\sqrt{f}$ satisfies $\|e\| = 1$ and $|\langle u, e\rangle_d| \le \|u\|$. If the signed conditional coordinate means are $u_\star s$ on $\mathcal{C}$ and zero elsewhere, then

$$a = u_\star s \chi = \sqrt{f}\, u_\star s\, e, \qquad \|a\| = \langle a, e\rangle_d = \sqrt{f}\, u_\star s. \tag{20}$$

The constants $U_{\max}$ and $u_\star$ serve different purposes: the former bounds stochastic steps, while the latter is a typical coordinate scale in the drift model. The absence of an explicit factor d in the resulting inequalities does not make their constants independent of model dimension.

Working domain versus failure region. The regularity neighborhood $\mathcal{U}$ need not equal $\mathcal{N}_R$. The former is a domain on which the field can be controlled; the latter is a model of the effective reward-failure boundary. Snapshot exit may occur after a learner has crossed that boundary between synchronizations. At constant η, the ball of radius $R + S\eta U_{\max}$ suffices to cover an intervening cycle starting inside $\mathcal{N}_R$, provided the update bound holds there. All local statements require the relevant learner and behavior parameters to belong to $\mathcal{U}$. Smoothness of a clipped, finite-sample surrogate is not automatic: Assumption 1 is imposed on the averaged field in the region used by the analysis.

Policy-indexed observations. The empirical event is the abrupt training-reward degradation defined in Section 2.2. The exit model identifies its generating snapshot with the first snapshot outside $\mathcal{N}_R$:

$$t_{\text{obs}} = \tau_S, \qquad M_{\text{collapse}} = \tau_S/S. \tag{21}$$

This is a modeling correspondence, not a reward-labeling rule or a claim that failure is irreversible. It refers to the synchronization step of the generating policy, not to when its reward is delivered to the logger.

For comparison, define the learner's first exit $\tau = \inf\{t : \theta_t \notin \mathcal{N}_R\}$. In general, only $\tau \le \tau_S$ follows: a learner may exit and re-enter before a synchronization. If the first exit remains visible at the next synchronization, then

$$\tau_S = S\lceil \tau/S \rceil, \qquad \tau \in (\tau_S - S, \tau_S]. \tag{22}$$

The main snapshot-clock theorem does not require this additional no-re-entry condition. In particular, an empirical timestamp must not be rounded according to a fitted theoretical clock to enforce equation 22.

A.2 Proof of the Local Bias Bound

Fix $\theta, \phi, \phi' \in \mathcal{U}$ and set $h(z) = \nabla_\theta \ell(z; \theta, \phi')$. Adding and subtracting $\mathbb{E}_{p_\phi} h$ gives

$$\begin{aligned}\mathcal{H}(\theta, \phi) - \mathcal{H}(\theta, \phi') &= \mathbb{E}_{p_\phi}[\nabla_\theta \ell(z; \theta, \phi) - \nabla_\theta \ell(z; \theta, \phi')] \\ &\quad + (\mathbb{E}_{p_\phi} - \mathbb{E}_{p_{\phi'}})h(z).\end{aligned} \tag{23}$$

The first term is bounded by $L_\rho\|\phi - \phi'\|$. For the second, boundedness of h and Pinsker's inequality yield

$$\|(\mathbb{E}_{p_\phi} - \mathbb{E}_{p_{\phi'}})h\| \le 2\Gamma\ \mathrm{TV}(p_\phi, p_{\phi'}) \le \Gamma\sqrt{2D_{\mathrm{KL}}(p_\phi\|p_{\phi'})} \le \Gamma\sqrt{2C_\pi}\|\phi - \phi'\|.$$

Hence $\phi \mapsto \mathcal{H}(\theta, \phi)$ is locally Lipschitz with constant C_{stale}. Taking $\phi = \phi_t$ and $\phi' = \theta = \theta_t$, and using $\theta_t - \phi_t = \sum_{j=t-k_t}^{t-1} \eta_j u_j$, proves Proposition 1. For constant η, averaging the bound over a complete cycle uses $S^{-1}\sum_{k=0}^{S-1} k = (S-1)/2$. Only the bound is sawtooth-shaped; the realized bias need not grow monotonically within a cycle. □

Directional expansion. Suppose additionally that the behavior-policy Jacobian of $\mathcal{H}$ exists and is locally L_B-Lipschitz. With $\Delta = \theta - \phi$, integration along the segment from θ to ϕ gives

$$\mathcal{H}(\theta, \phi) - \mathcal{H}(\theta, \theta) = B(\theta)\Delta + r(\theta, \phi), \quad B(\theta) = -\partial_\phi \mathcal{H}(\theta, \theta), \quad \|r(\theta, \phi)\| \le \tfrac{1}{2}L_B\|\Delta\|^2. \tag{24}$$

The Lipschitz field bound gives $\|B(\theta)\|_{\mathrm{op}} \le C_{\mathrm{stale}}$. Differentiability alone would give an $o(\|\Delta\|)$ remainder, not the displayed quadratic bound. The sign of $\langle B(\theta)\Delta, \Delta\rangle_d$ remains unrestricted.

A.3 GRPO Channels: Exact Cancellation, Clipping, and Finite Groups

These examples isolate individual channels rather than represent the full token-level, clipped, KL-regularized objective. They show why reinforcement of drift is possible, and why its net sign must remain a separate premise.

Exact importance weighting. For a behavior-independent advantage $A(x, y)$, a sequence-level ratio $\rho(y) = \pi_\theta(y \mid x)/\pi_\phi(y \mid x)$, and no clipping,

$$\mathbb{E}_{y\sim\pi_\phi}[\rho(y)A(x, y)\nabla_\theta \log \pi_\theta(y \mid x)] = \mathbb{E}_{y\sim\pi_\theta}[A(x, y)\nabla_\theta \log \pi_\theta(y \mid x)]. \tag{25}$$

Thus this isolated policy-gradient term has no stale-policy bias. This conclusion does not automatically extend to a sampled KL term, behavior-dependent group statistics, or token-level ratios that leave the prefix distribution uncorrected.

The clipping contribution. At a fixed stale sampling distribution, compare the clipped and unclipped policy-gradient terms. Away from clipping boundaries, their difference is

$$\delta^{\mathrm{clip}} = -\mathbb{E}_{p_\phi} \sum_{\mathrm{clipped}} w_{i,l}\hat{A}_i\rho_{i,l}\nabla_\theta \log \pi_\theta(y_{i,l} \mid x, y_{i,<l}), \tag{26}$$

where $w_{i,l} > 0$ is the response/token normalization. Clipping occurs when $\hat{A}_i > 0$, $\rho_{i,l} > 1 + \epsilon$, or $\hat{A}_i < 0$, $\rho_{i,l} < 1 - \epsilon$. In either case $\hat{A}_i \log \rho_{i,l} > 0$. If the score changes smoothly along $\Delta = \theta - \phi$, then

$$\Delta^\top \nabla_\theta \log \pi_\theta = \log \rho_{i,l} + O(\|\Delta\|^2),$$

with the norm convention absorbed into the local remainder constant. Assuming the weighted absolute coefficients have bounded expectation,

$$\Delta^\top \delta^{\mathrm{clip}} \le O(\|\Delta\|^2). \tag{27}$$

Clipping therefore removes updates that, to this local accuracy, push further along the already favored displacement. This is not a global negative-semidefiniteness claim. For a fixed nonzero clip width and uniformly bounded scores, clipping is inactive in a sufficiently small neighborhood of $\theta = \phi$; its derivative there is zero, rather than necessarily a nonzero component of $B(\theta)$.

An exact finite-group normalization calculation. Fix a prompt, let rewards be in $\{0, 1\}$, and write $q_\theta = \Pr_{\pi_\theta}(r = 1) \in (0, 1)$. Consider $G \ge 2$ independent responses, sequence-level importance ratios, no clipping, no length weighting, and no KL term. Use the within-group population standard deviation and stabilizer $\epsilon_A \ge 0$; all-identical groups have zero advantage. For a selected response let

K be the number of successful peers, so $K \sim \mathrm{Bin}(G-1, q_\phi)$. Its normalized advantage conditional on success or failure is

$$A_+(k) = \frac{G-1-k}{\sqrt{(k+1)(G-1-k)+G\epsilon_A}}, \qquad A_-(k) = \frac{-k}{\sqrt{k(G-k)+G\epsilon_A}}, \tag{28}$$

with zero-over-zero endpoints interpreted as zero. Define the polynomial

$$C_{G,\epsilon_A}(p) = \sum_{k=0}^{G-1} \binom{G-1}{k} p^k (1-p)^{G-1-k} [A_+(k) - A_-(k)]. \tag{29}$$

Lemma 1 (Finite-group stale normalization)**.** *For the isolated estimator above,*

$$\mathcal{H}_G(\theta, \phi) = C_{G,\epsilon_A}(q_\phi) \nabla_\theta q_\theta. \tag{30}$$

Consequently, if q is twice continuously differentiable,

$$\delta_G = -C'_{G,\epsilon_A}(q_\theta)(\nabla_\theta q_\theta)(\nabla_\theta q_\theta)^\top \Delta + O(\|\Delta\|^2). \tag{31}$$

Proof. Exchangeability reduces the group mean to a selected response. Importance weighting changes that response's law to π_θ while leaving its peers under π_ϕ. The score identities

$$\mathbb{E}_{\pi_\theta}[\mathbf{1}\{r=1\} \nabla \log \pi_\theta] = \nabla q_\theta, \qquad \mathbb{E}_{\pi_\theta}[\mathbf{1}\{r=0\} \nabla \log \pi_\theta] = -\nabla q_\theta$$

then give $[A_+(K) - A_-(K)] \nabla q_\theta$ conditional on the peers. Averaging over K proves equation 30. Subtract the on-policy expression and expand $q_\phi = q_\theta - \nabla q_\theta^\top \Delta + O(\|\Delta\|^2)$ to obtain equation 31. □

The channel is reinforcing where $C'_{G,\epsilon_A}(q) < 0$ and opposing where that derivative is positive. Finite groups cannot be replaced by their informative-group probability alone: for example, $C_{2,\epsilon_A}(p) = 1/(1+2\epsilon_A)$ is constant, so this channel vanishes identically at $G = 2$ in the stated estimator. For $G = 3$, it is also constant, $C_{3,\epsilon_A}(p) = 2/(\sqrt{2} + 3\epsilon_A)$. These examples preclude a universal finite-group reinforcement claim.

Population normalization. In the population-statistics variant, with $\varsigma(q) = \sqrt{q(1-q)}$, the coefficient is $C_\infty(q) = 1/(\varsigma(q) + \epsilon_A)$. The first-order response becomes

$$B^{\mathrm{pop}} = \frac{1-2q}{2\varsigma(q)(\varsigma(q)+\epsilon_A)^2} (\nabla q)(\nabla q)^\top. \tag{32}$$

It is positive semidefinite for $q < 1/2$ and negative semidefinite for $q > 1/2$. This provides an explicit reinforcing channel on difficult prompts, not a sign theorem for the full GRPO field. Prefix-distribution shifts, clipping, finite-group effects, and the sampled KL penalty can change the net response.

A.4 Proof and Scope of the Feedback Threshold

Let $F_\gamma(s) = \Phi(\gamma s)$ on $[0,1]$. Continuity, monotonicity, $F_\gamma(0) = 0$, and $F_\gamma(1) < 1$ ensure that a largest fixed point $s_+(\gamma)$ exists.

Variational threshold. At any positive fixed point, $r = \gamma s$ satisfies $\gamma = r/\Phi(r)$. Therefore no such point exists when $\gamma < \gamma_c = \inf_{r>0} r/\Phi(r)$. Conversely, for $\gamma > \gamma_c$, choose $r > 0$ with $r/\Phi(r) < \gamma$. Then $s = r/\gamma < \Phi(r) < 1$ and $F_\gamma(s) > s$. Because $F_\gamma(1) < 1$, continuity gives a positive fixed point between s and one. The Lipschitz bound and differentiability at zero give

$$\frac{1}{L_\Phi} \le \gamma_c \le \gamma_{\mathrm{lin}} \triangleq \frac{1}{\Phi'(0)}. \tag{33}$$

No statement at equality $\gamma = \gamma_c$ is needed.

Monotonicity and saturation. If $\gamma' > \gamma$, then $F_{\gamma'}(s_+(\gamma)) \ge s_+(\gamma)$. A root of $F_{\gamma'}(s) - s$ therefore lies at or above $s_+(\gamma)$, proving monotonicity of the largest root. For any fixed $s_0 \in (0,1)$, $F_\gamma(s_0) \to 1 > s_0$ as $\gamma \to \infty$. It follows that $s_+(\gamma) \ge s_0$ for sufficiently large γ, and hence $s_+(\gamma) \to 1$.

Dynamics with a vanishing task signal. If $s_n \geq s_+$, then $s_{n+1} = \Phi(\gamma s_n + \rho_n) \geq F_\gamma(s_+) = s_+$. This proves forward invariance above the largest root. Let $L = \limsup_n s_n$ and choose a subsequence along which $s_{n_k+1} \to L$ and $s_{n_k} \to L' \leq L$. Since $\rho_n \to 0$, continuity gives $L = F_\gamma(L') \leq F_\gamma(L)$. Above the largest root, $F_\gamma(s) < s$; therefore $L \leq s_+$. Together with invariance, this yields $s_n \to s_+$. Below threshold, $F_\gamma(s) < s$ for every $s > 0$, so the same limsup argument gives $s_n \to 0$ for every initialization. This proves Proposition 2. □

Branch entry is a separate condition. The result does not assert that every high-lag run reaches the upper branch. A persistent task signal or optimizer memory can seed a common direction, but their presence alone does not prove that $s_{n_0} \geq s_+(\gamma)$. Without cross-cycle directional persistence, coherent individual cycles can cancel in their cumulative displacement. The closure in equation 10 is therefore a mean-field closure with an explicit basin condition, not an exact reduction of the AdamW recursion.

Continuous versus discontinuous onset. If Φ is strictly concave on $[0, \infty)$, then $F_\gamma(s) - s$ is strictly concave and $\gamma_c = \gamma_{\text{lin}}$. For $\gamma > \gamma_c$ it has one positive fixed point; that point approaches zero as $\gamma \downarrow \gamma_c$. To see continuity, any positive subsequential limit would be a positive fixed point at γ_c, where strict concavity gives $F_{\gamma_c}(s) < s$ for $s > 0$.

More generally, if $\gamma_c < \gamma_{\text{lin}}$, then the onset is discontinuous. Indeed, differentiability at zero gives an $s_0 > 0$ and a neighborhood of γ_c in which $F_\gamma(s) < s$ for $0 < s \leq s_0$. Every positive fixed point just above threshold must exceed s_0. This proves the possibility of a jump, not its occurrence in the LLM experiments. The observed budget constrains the combined ratio R/v, and cannot identify coherence alone without comparable exit geometry, transients, and measurement precision.

An explicit nonlinear example. For $r \geq 0$, consider the effective response

$$\Psi_\kappa(r) = \frac{r + \kappa r^3}{\sqrt{1 + (r + \kappa r^3)^2}}, \qquad \kappa > 0, \tag{34}$$

extended oddly. It is increasing, bounded, globally Lipschitz, and $\Psi'_\kappa(0) = 1$. Choose $\gamma_0 < 1$ and $s_0 \in (0, 1)$. If

$$\kappa > \frac{s_0/\sqrt{1 - s_0^2} - \gamma_0 s_0}{(\gamma_0 s_0)^3},$$

then $\Psi_\kappa(\gamma_0 s_0) > s_0$, so $\gamma_c < \gamma_0 < \gamma_{\text{lin}} = 1$ and a discontinuous branch exists. This is an illustrative composite of superlinear feedback and a saturating optimizer response; it is not derived from the measured GRPO response or used to fit the reported collapse labels.

Retaining every lag. Keeping the full lag distribution in the homogeneous closure gives, for $S > 1$,

$$F_{S,\gamma}(s) = \frac{1}{S} \sum_{k=0}^{S-1} \Phi\left(\frac{2\gamma k}{S-1} s\right). \tag{35}$$

This is a lag-resolved closure, not an exact stochastic optimizer recursion. Its derivative at zero is still $\gamma\Phi'(0)$. Its largest-root threshold $\gamma_c(S)$ lies in $[1/L_\Phi, 1/\Phi'(0)]$ by the same contraction and intermediate-value arguments. For a strictly concave response, $\gamma_c(S) = \gamma_{\text{lin}}$; for a nonconcave response the nonlinear threshold can depend on S. The saturated root tends to $(S-1)/S$, because the $k = 0$ contribution is zero. Persistent optimizer memory can supply a nonzero reset-step contribution, but that is an additional model, not an automatic property of equation 35.

At fixed γ, let $F_{\infty,\gamma}(s) = \int_0^1 \Phi(2\gamma u s)\, du$. A left-Riemann-sum bound, using the $2L_\Phi\gamma$ Lipschitz constant in u, gives

$$\sup_{s \in [0,1]} |F_{S,\gamma}(s) - F_{\infty,\gamma}(s)| \leq \frac{L_\Phi \gamma + 1}{S}. \tag{36}$$

Thus a common constant-product boundary is a useful leading-order reduction, not an exact finite-S identity for every response. Replacing $\bar{k}$ by $S/2$ adds the further relative correction $1/S$.

A.5 Uniform Drift Bounds and the Snapshot Clock

We first establish one uniform event and then evaluate it at the random exit time. This avoids substituting a stopping time into a bound proved only at a fixed time. Learning rates are deterministic and nonnegative, and $\delta \in (0,1)$ is a fixed failure probability. The constant-rate experiments are a special case.

Lemma 2 (A uniform stochastic tube). *Let ζ be a stopping time and suppose $\|u_t\| \le U_{\max}$ for $t < T \wedge \zeta$. With $a_t = \mathbb{E}[u_t \mid \mathcal{F}_t]$ and $D_t = \sum_{j<t} \eta_j a_j$, with probability at least $1-\delta$,*

$$\sup_{0 \le t \le T \wedge \zeta} \|\theta_t - \theta_0 - D_t\| \le q_T(\delta) = 2U_{\max}\sqrt{2V_T \log(2/\delta)}. \tag{37}$$

Proof. Use the stopped martingale $N_t = \sum_{j<t} \eta_j \mathbf{1}\{j < \zeta\}(u_j - a_j)$. The indicator is $\mathcal{F}_j$-measurable, and each increment has RMS norm at most $2U_{\max}\eta_j$. The Hilbert-space maximal inequality of Pinelis (1994, Theorem 3.5), applied to the scaled inner product $\langle \cdot, \cdot \rangle_d$, gives

$$\Pr\left(\max_{t \le T} \|N_t\| \ge r\right) \le 2\exp\left(-\frac{r^2}{8U_{\max}^2 V_T}\right).$$

For $t \le \zeta$, this martingale equals $\theta_t - \theta_0 - D_t$. Choosing $r = q_T(\delta)$ proves the claim. When $V_T = 0$, all displacements vanish and the result is immediate. □

A sharper fluctuation radius. If the centered updates satisfy $\|u_t - a_t\| \le \sigma_u$ before stopping, the same argument replaces q_T by $\sigma_u\sqrt{2V_T \log(2/\delta)}$ throughout the clock bounds. The default choice $\sigma_u = 2U_{\max}$ needs no extra information, but can be conservative because $U_{\max}$ also bounds the mean motion. A smaller noise bound must be established separately, not selected to match the observed collapse times.

Proof of Theorem 1. Apply Lemma 2 with $\zeta = \tau_S$. On its event, for every $t \le T \wedge \tau_S$,

$$(1-\varepsilon)v\Lambda_t - E_T \le \|\theta_t - \theta_0\| \le (1+\varepsilon)v\Lambda_t + E_T, \qquad E_T = a + q_T(\delta). \tag{38}$$

Let $\bar{t}$ be the first synchronization step with $(1-\varepsilon)v\Lambda_{\bar{t}} \ge R + E_T$. The horizon condition ensures $\bar{t} \le T$. If $\tau_S > \bar{t}$, the lower bound at $\bar{t}$ gives $\|\theta_{\bar{t}} - \theta_0\| \ge R$, contradicting the definition of τ_S. Thus $\tau_S \le \bar{t} \le T$. At exit, the upper radial bound gives $R \le (1+\varepsilon)v\Lambda_{\tau_S} + E_T$, proving the lower clock bound. By minimality of $\bar{t}$, its preceding synchronization has cumulative rate strictly below $(R + E_T)/((1-\varepsilon)v)$. The intervening cycle contributes at most h_S, so

$$\Lambda_{\tau_S} \le \Lambda_{\bar{t}} \le \frac{R + E_T}{(1-\varepsilon)v} + h_S.$$

Both bounds hold on the same event. □

Normalized error and finite resolution. Writing $c_2 = R/v$, the theorem is equivalently

$$\frac{1 - E_T/R}{1+\varepsilon} \le \frac{\Lambda_{\tau_S}}{c_2} \le \frac{1 + E_T/R}{1-\varepsilon} + \frac{h_S}{c_2}. \tag{39}$$

For ε bounded away from one and small E_T/R, this gives equation 18 under the observation correspondence. At constant η and a comparison horizon $T = O(c_2/\eta)$,

$$\frac{q_T(\delta)}{R} = O\left(\frac{U_{\max}}{v}\sqrt{\frac{\eta \log(2/\delta)}{c_2}}\right).$$

The error is controlled by normalized noise, not by parameter count alone. The bound can be loose, especially when collapse occurs after only one or a few synchronization cycles. In particular, $S\eta/c_2$ need not be small in such cells.

For an exactly straight deterministic trajectory $\theta_t - \theta_0 = vt\eta e$, the snapshot model reduces to

$$M_{\text{collapse}} = \left\lceil \frac{c_2}{S\eta} \right\rceil, \qquad c_2 \le \eta t_{\text{obs}} < c_2 + S\eta. \tag{40}$$

This exact ceiling formula is a special case, not an identity obtained by discarding the error terms in equation 18.

Connecting cycle coherence to cumulative drift. At constant η, let $\bar{a}_n = S^{-1}\sum_{k=0}^{S-1} a_{nS+k}$. The bridge from the mean-field model to actual parameter motion is

$$\bar{a}_n = \sqrt{f}\, u_\star s_n e + r_n, \tag{41}$$

where r_n includes other coordinates, directional variation, weight decay, and deviations of AdamW from the scalar closure. This is an explicit approximation condition on conditional *optimizer* means, not on raw gradients. Suppose that from a cycle n_0 onward $|s_n - s_{\mathrm{sat}}| \le \varepsilon_s s_{\mathrm{sat}}$ and $\|r_n\| \le \varepsilon_r v$, where $v = \sqrt{f}\, u_\star s_{\mathrm{sat}}$. Then $\|\bar{a}_n - ve\| \le (\varepsilon_s + \varepsilon_r)v$. Writing $t_0 = n_0 S$, summing complete cycles and bounding a partial cycle yields equation 15 with

$$\varepsilon = \varepsilon_s + \varepsilon_r, \qquad a = (U_{\max} + v)(\Lambda_{t_0} + S\eta). \tag{42}$$

An incomplete cycle at a time before τ_S uses only pre-exit updates; τ_S itself is a cycle boundary. The initial segment is absorbed into a without assuming it already follows the persistent direction. If the mean-field branch converges but does not saturate, the appropriate velocity is instead $v(S,\eta) = \sqrt{f}\, u_\star s_+(\gamma)$, which generally produces an S-dependent budget. A shared clock therefore needs comparable velocity and exit distance, not just the existence of a nonzero branch.

Corollary 1 (A sufficient non-collapse horizon). *If* $\sup_{t \le T\wedge\tau_S} \|D_t\| \le a_0 < R$ *and* $a_0 + q_T(\delta) < R$, *then* $\Pr(\tau_S > T) \ge 1 - \delta$. *In particular, at constant* η *this holds whenever*

$$T < \frac{(R - a_0)^2}{8U_{\max}^2 \eta^2 \log(2/\delta)}. \tag{43}$$

Proof. On the uniform-tube event, $\|\theta_t - \theta_0\| \le a_0 + q_T(\delta) < R$ for every $t \le T \wedge \tau_S$. If $\tau_S \le T$, this contradicts the defining exit inequality. Substitute $V_T = T\eta^2$ for the constant-rate bound. □

For fixed confidence and a margin $R - a_0$ bounded away from zero, this is a high-probability $\Omega(\eta^{-2})$ lower bound on the snapshot exit time. The zero-conditional-mean case has $a_0 = 0$. A condition $\|a_t\| \le \varepsilon U_{\max}$ instead gives $\|D_t\| \le \varepsilon U_{\max}\Lambda_t$; at $T = O(\eta^{-2})$ this term has order ε/η. Thus merely assuming that ε eventually tends to zero is not enough.

When the scalar subcritical model controls cumulative drift. A sufficient, stronger condition is $q = \gamma L_\Phi < 1$ and $\sum_n \rho_n < \infty$. Then $s_{n+1} \le q s_n + L_\Phi \rho_n$, and summing gives

$$\sum_{n\ge 0} s_n \le \frac{s_0 + L_\Phi \sum_{n\ge 0}\rho_n}{1 - q}. \tag{44}$$

Together with equation 41 and bounded cumulative residuals, this bounds the coherent contribution to D_t by $S\eta\sqrt{f}\, u_\star \sum_n s_n$, up to a partial cycle. It closes a sufficient route from a subcritical feedback model to Corollary 1; the entire subcritical region is not automatically a zero-drift regime.

Displacement predictions. The coherent tube predicts an approximately linear displacement in Λ_t when a, ε, and the martingale term are small. With negligible cumulative mean drift, the bound is instead of order $\sqrt{V_t} = \sqrt{\eta\Lambda_t}$ at constant rate. This is an upper scale, not a universal equality for every stable run. For an unstopped zero-mean process, orthogonality of martingale differences gives $\mathbb{E}\|\theta_t - \theta_0\|^2 = \sum_{j<t} \eta_j^2 \mathbb{E}\|u_j\|^2$; a matching mean-square diffusion scale additionally requires nondegenerate update variance. Reward curves alone do not measure either displacement regime.

A.6 AdamW: Bounds, Normalization, and Momentum

AdamW uses the moment recursions of Adam with decoupled weight decay (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). To keep time indexing explicit, write its ascent update after n gradient observations as

$$\nu_n = \frac{\widehat{m}_n}{\sqrt{\widehat{v}_n} + \epsilon_{\mathrm{Adam}}} - \lambda\theta_{n-1}, \qquad \widehat{m}_n = \sum_{r=1}^{n} \alpha_{n,r} h_r, \qquad \widehat{v}_n = \sum_{r=1}^{n} \omega_{n,r} h_r^2, \tag{45}$$

where $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$, $\epsilon_{\text{Adam}} > 0$, $\lambda \ge 0$, and h_r is the gradient after any gradient-norm clipping, all operations are coordinatewise, and

$$\alpha_{n,r} = \frac{(1-\beta_1)\beta_1^{n-r}}{1-\beta_1^n}, \qquad \omega_{n,r} = \frac{(1-\beta_2)\beta_2^{n-r}}{1-\beta_2^n}.$$

This paragraph uses ν_n for the applied update to distinguish one-based optimizer indexing from u_t in the main text.

A corrected worst-case bound. Cauchy–Schwarz gives, for every coordinate and every gradient history,

$$\frac{|\widehat{m}_{n,j}|}{\sqrt{\widehat{v}_{n,j}} + \epsilon_{\text{Adam}}} \le C_n, \qquad C_n^2 = \frac{(1-\beta_1)^2}{1-\beta_2}\frac{1-\beta_2^n}{(1-\beta_1^n)^2}\sum_{k=0}^{n-1}\left(\frac{\beta_1^2}{\beta_2}\right)^k. \tag{46}$$

When $\widehat{v}_{n,j} = 0$, the numerator is also zero and the inequality holds by continuity. If $0 < \beta_2 < 1$ and $\beta_1^2 < \beta_2$, then $C_\star = \sup_{n\ge 1} C_n < \infty$. On a parameter region $\|\theta\| \le B_\theta$, it follows that

$$\|\nu_n\| \le C_\star + \lambda B_\theta. \tag{47}$$

Thus a local $U_{\max}$ can include both bias correction and weight decay. It depends on the parameter region as well as optimizer hyperparameters, and is not the typical saturated velocity. To prove equation 46, apply $(\sum_r \alpha_r h_r)^2 \le (\sum_r \alpha_r^2/\omega_r)(\sum_r \omega_r h_r^2)$ and substitute the two weights. Finiteness follows from the geometric sum and convergence of the bias-correction factors.

What saturation means. For a constant nonzero coordinate gradient h, bias correction gives $\widehat{m} = h$, $\widehat{v} = h^2$; its normalized component is $h/(|h| + \epsilon_{\text{Adam}})$, approaching a sign step only when $|h| \gg \epsilon_{\text{Adam}}$. For Gaussian $g = \mu + \sigma Z$, a sign update has the exact mean $\mathbb{E}[\text{sign}(g)] = \text{erf}(\mu/(\sqrt{2}\sigma))$. Replacing Adam's random second moment by a fixed stationary value $\mu^2 + \sigma^2$ gives the illustrative response

$$\frac{\mathbb{E}[g]}{\sqrt{\mathbb{E}[g^2]}} = \frac{\rho}{\sqrt{1+\rho^2}}, \qquad \rho = \mu/\sigma. \tag{48}$$

This is exact for that deterministic-denominator approximation, not for AdamW's conditional mean with a random denominator and stored momentum. The distinction between sign, magnitude, and variance adaptation motivates this modeling choice (Balles & Hennig, 2018).

Clipping and scale invariance. The history-wise bound holds whether or not h_r is clipped. With $\epsilon_{\text{Adam}} = 0$, multiplying *every* gradient in a coordinate's history by the same positive constant leaves its normalized Adam component unchanged. Gradient-norm clipping generally applies a different factor at each step, so it can change the trajectory and its mean velocity. Weight decay can likewise alter both direction and speed. Neither operation justifies replacing $U_{\max}$ by one or treating v as invariant without qualification.

What a momentum bound actually establishes. For a fixed normalization, a constant learning rate, and $z_t = \beta_1 z_{t-1} + (1-\beta_1)h_t$ with $z_{-1} = 0$,

$$\eta\sum_{t=0}^{T-1} z_t = \eta\sum_{t=0}^{T-1} h_t - \eta\sum_{t=0}^{T-1}\beta_1^{T-t}h_t. \tag{49}$$

If $\|h_t\| \le U_h$, the final term has norm at most $\eta U_h \beta_1/(1-\beta_1)$. Bias-correcting this first-moment average adds at most another term of that size: since $\|z_t\| \le (1-\beta_1^{t+1})U_h$, $\|z_t/(1-\beta_1^{t+1}) - z_t\| \le \beta_1^{t+1}U_h$. These are cumulative-displacement identities along the same input sequence, not a proof that two independently evolved training trajectories differ by a fixed number of steps. A time-varying Adam preconditioner or learning rate needs additional control and is not covered by this identity.

The main stochastic tube uses $a_t = \mathbb{E}[u_t \mid \mathcal{F}_t]$ for the actual optimizer, so it remains valid with momentum under its stated hypotheses. However, zero-mean raw gradients do not imply $a_t = 0$: the optimizer state contains past noise. This is why the incoherent result explicitly requires control of cumulative optimizer drift.

A.7 Raw-Gradient Cosine: An Error-Controlled Readout

Let $\Sigma_t = \mathbb{E}[\xi_t \xi_t^\top \mid \mathcal{F}_t]$ and $Q_t = \|\mu_t\|_2^2 + \operatorname{tr}\Sigma_t > 0$. The following deterministic perturbation lemma states exactly what is required to replace the observed cosine by a signal fraction.

Lemma 3 (Cosine approximation). *For $j = t-1, t$, suppose $|\|g_j\|_2^2 - Q_j| \le \epsilon_N Q_j$, with $0 \le \epsilon_N < 1$, and $|g_t^\top g_{t-1} - \mu_t^\top \mu_{t-1}| \le \epsilon_I \sqrt{Q_t Q_{t-1}}$. Then*

$$\left| c_t - \frac{\mu_t^\top \mu_{t-1}}{\sqrt{Q_t Q_{t-1}}} \right| \le \frac{\epsilon_I + \epsilon_N}{1 - \epsilon_N}. \tag{50}$$

If the displayed signal-over-energy ratio differs from $\mathrm{SNR}_t/(1+\mathrm{SNR}_t)$ by at most ϵ_D, add ϵ_D to the right-hand side.

Proof. Write $r = \mu_t^\top \mu_{t-1}/\sqrt{Q_t Q_{t-1}}$, so $|r| \le 1$. The normalized numerator is $r + e_I$, $|e_I| \le \epsilon_I$; the normalized denominator is $d_N = \|g_t\|_2 \|g_{t-1}\|_2 / \sqrt{Q_t Q_{t-1}}$ lies in $[1-\epsilon_N, 1+\epsilon_N]$. Hence $|(r+e_I)/d_N - r| \le (\epsilon_I + |r|\epsilon_N)/(1-\epsilon_N)$. The final claim is the triangle inequality. □

Why the hypotheses matter. The exact numerator expansion is

$$g_t^\top g_{t-1} = \mu_t^\top \mu_{t-1} + \mu_t^\top \xi_{t-1} + \xi_t^\top g_{t-1}.$$

The last term has zero conditional mean given $\mathcal{F}_t$. The middle term need not: μ_t depends on the preceding update and can also change at a synchronization. Thus fresh sampling alone does not establish the inner-product hypothesis or slow variation across a reset. Likewise, a ratio of expected energies is not generally the expectation of a random cosine. Lemma 3 isolates these errors rather than silently discarding them.

When adjacent means and energies are close, ϵ_D is small, and the signal ratio becomes

$$\frac{\|\mu_t\|_2^2}{\|\mu_t\|_2^2 + \operatorname{tr}\Sigma_t} = \frac{\mathrm{SNR}_t}{1 + \mathrm{SNR}_t}.$$

For positive coordinate noise variances, $\mathrm{SNR}_t = \sum_j w_{t,j} \rho_{t,j}^2$ is an exact identity; zero-variance coordinates can instead be handled directly through $\|\mu_t\|_2^2 / \operatorname{tr}\Sigma_t$.

Proposition 3 (Noise-weighted masking). *If $|\rho_{t,j}| \le \rho_{\mathcal{C}}$ on $\mathcal{C}$ and $|\rho_{t,j}| \le \rho_{\mathcal{H}}$ elsewhere, then*

$$\mathrm{SNR}_t \le w_{\mathcal{C}} \rho_{\mathcal{C}}^2 + (1 - w_{\mathcal{C}}) \rho_{\mathcal{H}}^2, \tag{51}$$

where $w_{\mathcal{C}} = \sum_{j \in \mathcal{C}} w_{t,j}$. Thus a small weighted sum can mask coherent normalized updates on $\mathcal{C}$.

Proof. Splitting $\mathrm{SNR}_t = \sum_j w_{t,j} \rho_{t,j}^2$ over $\mathcal{C}$ and its complement gives

$$\sum_{j \in \mathcal{C}} w_{t,j} \rho_{t,j}^2 + \sum_{j \notin \mathcal{C}} w_{t,j} \rho_{t,j}^2 \le w_{\mathcal{C}} \rho_{\mathcal{C}}^2 + (1 - w_{\mathcal{C}}) \rho_{\mathcal{H}}^2.$$

Under the cosine-approximation conditions, a small right-hand side therefore gives a small raw-gradient cosine without excluding large coordinate-normalized signals in $\mathcal{C}$. □

Masking is possible, not automatic. For example, the algebraic values $w_{\mathcal{C}} = 10^{-6}$, $\rho_{\mathcal{C}} = 10$, and $\rho_{\mathcal{H}} = 0$ give $\mathrm{SNR} = 10^{-4}$, even though $\rho_{\mathcal{C}}/\sqrt{1+\rho_{\mathcal{C}}^2}$ is close to one. This is an illustration, not a measurement from the experiments. Low raw variance alone does not guarantee high normalized stale signal: the response matrix B_t can scale with the same coordinates. Localization of coherent drift in a low-noise block requires evidence or an explicit model assumption.

A conditional late-warning model. To model the observed rise, suppose the cosine approximation has error at most ϵ_c and there is a thin layer $R - \Delta_R \le \|\theta_t - \theta_0\| < R$ in which $\mathrm{SNR}_t \ge \kappa^2$. This is an additional signal assumption; surrogate failure does not by itself imply a large raw-gradient signal or an aligned adjacent mean.

Corollary 2 (Boundary-layer warning)**.** *At times in this layer where the readout conditions hold, $c_t \ge \kappa^2/(1+\kappa^2) - \epsilon_c$. Suppose, at constant η, that $\|\theta_t - \theta_0 - vt\eta e\| \le E < R - \Delta_R$ through τ_S, with a horizon long enough for snapshot exit. Let t_{edge} be the first layer-entry-or-crossing time. Then*

$$0 \le \tau_S - t_{\mathrm{edge}} \le \frac{\Delta_R + 2E}{v\eta} + S. \tag{52}$$

Proof. Monotonicity of $z/(1+z)$ gives the cosine bound. At first entry or crossing, $vt_{\mathrm{edge}}\eta \ge R - \Delta_R - E$. The first synchronization with $vt\eta \ge R + E$ is necessarily outside the region, giving $v\tau_S\eta \le R + E + vS\eta$. Subtracting proves equation 52. Since reaching radius R requires first reaching radius $R - \Delta_R$, the time difference is nonnegative. □

In the small-error limit, the leading warning duration is $(\Delta_R/R)c_2/\eta + S$. A sustained warning additionally requires the trajectory not to leave the high-SNR layer again before failure. The result neither guarantees a visible precursor at every step nor applies the readout approximation across arbitrary synchronization jumps. The stopping protocol does not provide evidence about recovery or post-collapse gradient dynamics, which are outside this analysis.

A.8 Weaker Moment Assumptions

If only $\mathbb{E}[\|u_t\|^2 \mid \mathcal{F}_t] \le U_2^2$ is available, Jensen's inequality still gives

$$\mathbb{E}\|\theta_t - \phi_t\| \le U_2 \sum_{j=t-k_t}^{t-1} \eta_j, \qquad \mathbb{E}\|\theta_t - \theta_0\| \le U_2 \Lambda_t.$$

These are expectation bounds, not deterministic speed limits. For the stopped martingale in Lemma 2, Hilbert-space orthogonality gives $\mathbb{E}\|N_T\|^2 \le U_2^2 V_T$. Doob's maximal inequality then yields

$$\Pr\left(\max_{t \le T} \|N_t\| \ge r\right) \le \frac{U_2^2 V_T}{r^2}. \tag{53}$$

Accordingly, the snapshot-clock proof remains valid with $q_T = U_2\sqrt{V_T/\delta}$, subject to its other assumptions. Finite second moments alone do not provide the logarithmic-confidence radius obtained from bounded increments.

A.9 Calibration, Identifiability, and Experimental Checks

What a shared clock identifies. In the deterministic snapshot model, a failed cell identifies the interval

$$\eta(t_{\mathrm{obs}} - S) < c_2 \le \eta t_{\mathrm{obs}}. \tag{54}$$

Several cells constrain c_2 through their interval intersection; with stochastic and model residuals, those intervals must be widened using equation 16. For the reported Llama counts, the deterministic intervals intersect in $(2.88, 3.20] \times 10^{-5}$; for the three Qwen3 cells, in $(6.08, 6.40] \times 10^{-5}$. These are model-based resolution intervals, not confidence intervals or independently estimated parameter ranges. Equality of the reported products is therefore compatible with a common budget but does not determine a unique latent exit time. It cannot, on its own, establish saturation or a discontinuous feedback transition.

Scope of the empirical comparison. The Llama sweeps vary both S and η and therefore examine both scalings. Qwen3-8B varies η at $S = 16$, so it checks timing at that synchronization interval, not invariance across S. The datasets and evaluation setups also differ; different budgets cannot be attributed solely to model family. A finite-horizon stable label is not a proof of infinite-horizon stability or of the martingale idealization. The main-grid products bracket an empirical boundary; the extended 3B comparison at $S = 32$ uses products 2.56×10^{-6} and 3.20×10^{-6}, or mean-lag products 1.24×10^{-6} and 1.55×10^{-6}. A threshold inferred from these points is specific to the measured setup and must account for the unequal observation horizons.

Falsifiable links in the mechanism. The required checks are distinct. A stale-versus-fresh field comparison at matched learner states tests the sign of $\langle \delta_t, \theta_t - \phi_t \rangle_d$. Applied-update projections and cycle-to-cycle directions test persistence and saturated mean speed; raw-gradient cosine is not a substitute for either. RMS displacement against Λ_t tests the straight-drift tube and whether stable runs require a different drift model. Reward annotations and snapshot-version records test the observation correspondence independently of the fitted clock. These checks are predictions and validation requirements of the mechanism, not experiments asserted to have been completed.

B Implementation Details

System. All runs use the asynchronous recipe of verl (Sheng et al., 2025) with disaggregated rollout and training: generation runs on vLLM (Kwon et al., 2023) workers and optimization on FSDP2 workers, on a single node of eight A100 GPUs per run, four for each. The rollout workers generate from a weight snapshot, the learner takes one optimizer step per incoming batch, and the weights are synchronized every S learner steps, which realizes the lag sequence in equation 5. No stabilization mechanism is enabled: there is no off-policy correction beyond the GRPO clip, no gradient projection, and no staleness-aware learning-rate control.

Configuration. Table 2 lists the settings. Within a model family, all (S, η) configurations share the data order, initialization, and remaining hyperparameters, so that S and η are the only varying factors. Each configuration is run with seed 42; Appendix C repeats the $S = 32$ runs of the 3B policy with seeds 100 and 2026.

Table 2: Training and evaluation configuration, shared by all runs unless stated otherwise. Values marked † are framework defaults that the launch scripts do not override.

Setting	Value
Training data	MATH train, 7.5k (Llama); DAPO-Math-17k (Qwen3)
Validation data	MATH test, 5k (Llama); MATH-500, AMC 23/24, Minerva (Qwen3)
Reward	binary exact match, $\{0, 1\}$
Algorithm	vanilla GRPO
Group size G / prompts per step	8 / 256 (one update per rollout batch)
Clip range ϵ / loss aggregation	$0.2^\dagger$ / token-mean†
KL term	in the loss, $\beta = 10^{-3}$, k3 estimator
Entropy coefficient	0
Staleness S (sync interval)	$\{8, 16, 32\}$
Learning rate η	$\{10^{-6}, 5 \times 10^{-7}, 2 \times 10^{-7}, 10^{-7}\}$, constant, no warmup
Optimizer	AdamW, $(\beta_1, \beta_2) = (0.9, 0.999)^\dagger$, weight decay $0.01^\dagger$
Gradient-norm clipping	$1.0^\dagger$
Micro-batch per device	4
Max prompt / response length	512 / 4096 tokens
Rollout temperature / top-p	$1.0^\dagger$ / $1.0^\dagger$
Evaluation	every 25 steps, temperature 1.0, top-p 1.0
Seeds	42; also 100, 2026 for the runs of Appendix C

Stopping criterion. Runs have no fixed step budget. A run is stopped once collapse is identified (Section 2.2), or once it has clearly outlived the horizon c_2/η at which collapsing runs with the same learning rate fail; the curves therefore end at different steps.

Evaluation protocol. The Llama policies are evaluated on the 5k MATH test split, which is disjoint from the 7.5k training split and never used for rollouts, with the same rule-based checker as in training. The Qwen3 policies are evaluated on MATH-500 (Lightman et al., 2024), AMC 2023 and AMC 2024 (Art of Problem Solving, 2024), and Minerva Math (Lewkowycz et al., 2022), reporting avg@1 on MATH-500 and Minerva Math and avg@8 on the two AMC sets. Decoding settings and the number of samples per problem are fixed across configurations and evaluation steps.

Gradient cosine. The cosine c_t is taken between the raw gradients of consecutive learner steps, before the optimizer state is applied, following the protocol of Xu et al. (2026): the required inner products and norms are accumulated across data-parallel ranks instead of materializing the full gradient vectors.

C Robustness to the Random Seed

The sweeps of Section 4 use one seed per configuration. To test whether the regularities depend on the seed, we repeat the $S = 32$ runs of `Llama-3.2-3B-Instruct` with seeds 42, 100 and 2026. Nothing else changes: data, pipeline, and hyperparameters are those of Section 4 and Appendix B. We use $S = 32$ because it is the staleness at which the sweep samples both sides of the stability boundary: the four learning rates of the main grid collapse (Figure 4), and the four extended-horizon learning rates are stable (Figure 5).

Collapsing runs. Table 3 reports M_{collapse} for each run, using the abrupt reward drop defined in Section 2.2. At every learning rate, the three seeds collapse in the same rollout batch. The per-batch mean training rewards before the collapse agree to within 0.02, and even the two-stage drop at $\eta = 2 \times 10^{-7}$, a partial drop one batch before the collapse, is reproduced by all seeds. The seed therefore moves t_{obs} by less than the observation granularity S, so, for the 3B policy at $S = 32$, the clock of Section 4.3 is a property of (S, η) and not of a particular sample path. The gradient cosine is equally reproducible: at $\eta = 10^{-7}$, in every seed c_t decays to about 0.05 after the initial reward climb and rises again shortly before the collapse, as described in Section 4.4.

Table 3: First collapsed rollout batch M_{collapse} of `Llama-3.2-3B-Instruct` at $S = 32$ under three seeds, with $t_{\text{obs}} = M_{\text{collapse}} S$. All seeds agree in every column.

Seed	$\eta = 10^{-6}$	5×10^{-7}	2×10^{-7}	10^{-7}
42	1	2	5	10
100	1	2	5	10
2026	1	2	5	10
t_{obs}	32	64	160	320

Stable runs. At $\eta \in \{8, 7, 6, 5\} \times 10^{-8}$, all three seeds train stably for 500 steps, and their reward, gradient-cosine, and validation curves are nearly indistinguishable. At 8×10^{-8} and 7×10^{-8} this horizon exceeds the clock c_2/η of 400 and 457 steps, so every seed survives past the point at which a run above the boundary would have collapsed; the stable side of the bracket at $S\eta = 2.56 \times 10^{-6}$ is therefore reproduced across seeds. At 6×10^{-8} and 5×10^{-8} the 500-step horizon is shorter than the clock, and these runs show seed agreement only; their long-horizon behavior is covered by the single-seed runs of Figure 2.

Why the seeds agree. The seed changes the data order and the rollout samples, which enter the update through the zero-mean noise ξ_t. Above the boundary, the collapse time is set by the coherent part of the drift, and the noise perturbs it only at relative order $O(\tau^{-1/2})$ (Theorem 1); below the boundary, the noise-driven displacement is predicted to stay far from the edge of $\mathcal{N}_R$ over the horizons we run (Corollary 1). In both regimes the seed should therefore leave the observable outcome unchanged, which is what the runs show.

D Extended Related Work

This appendix expands Section 5. Our work sits between three lines of research that rarely meet: systems that make LLM post-training asynchronous and must live with stale rollouts, algorithms and analyses that make policy optimization tolerate off-policy data, and the optimization literature on delayed updates and adaptive optimizers. We review each in turn and, at the end of every paragraph, state how our analysis differs. In short, prior work either chooses the lag empirically, corrects it inside the objective, or guarantees stability under sufficient conditions; we keep the objective fixed and characterize which (S, η) configurations fail, and when.

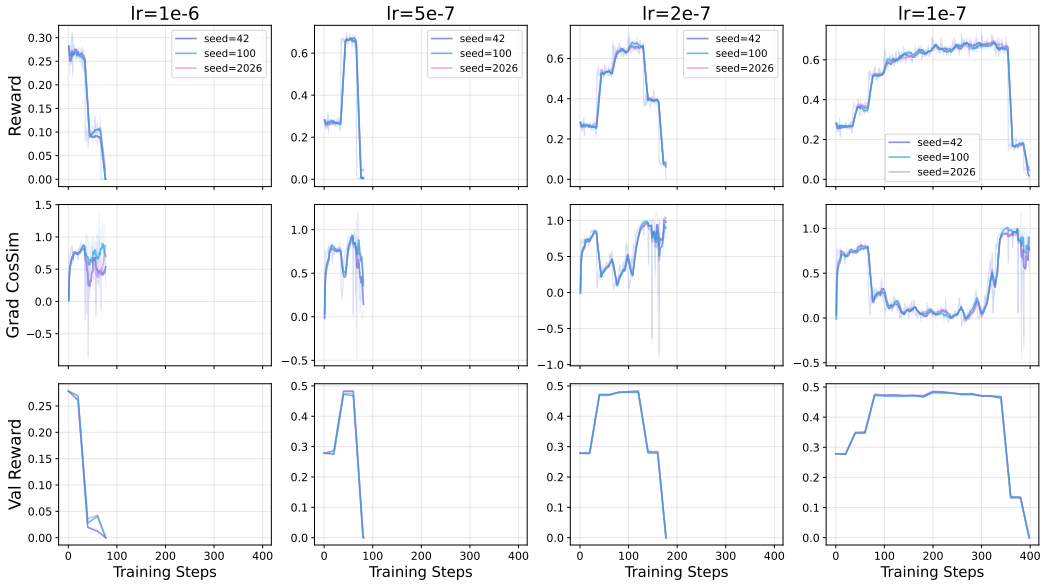


Figure 4: Seed robustness above the boundary: `Llama-3.2-3B-Instruct` at $S = 32$ with the four learning rates of the main grid, three seeds each. *Top:* training reward; *middle:* gradient cosine c_t; *bottom:* validation reward. All seeds collapse in the same rollout batch (Table 3), with nearly identical curves before the collapse.

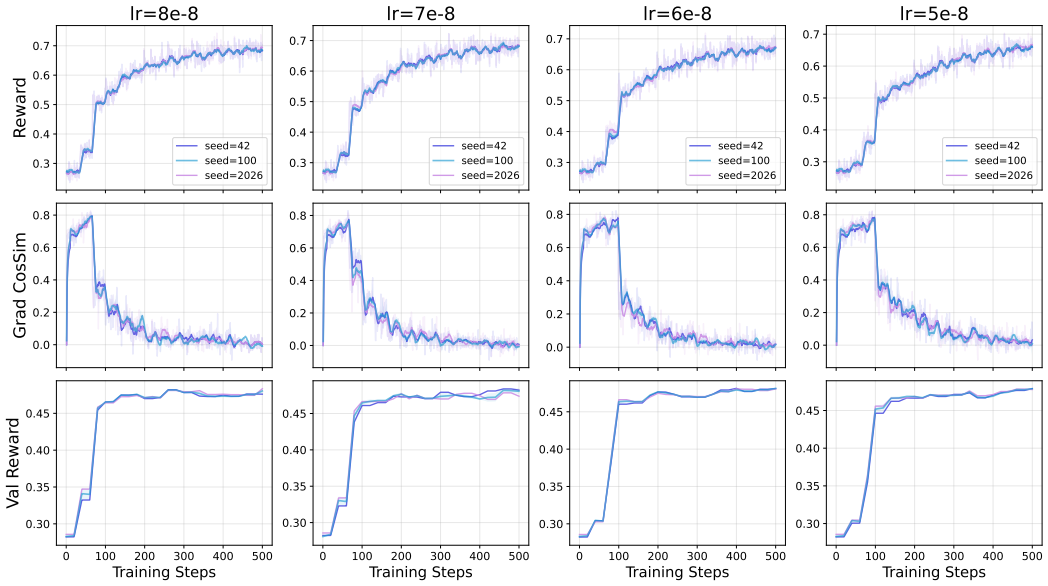


Figure 5: Seed robustness below the boundary: `Llama-3.2-3B-Instruct` at $S = 32$ with the four extended-horizon learning rates, three seeds each, over 500 steps. All runs are stable, and c_t decays to zero after the initial reward climb in every seed.

Asynchronous RL for LLMs: how staleness is handled. Because generation dominates the cost of RL for LLMs, recent systems decouple it from optimization (Sheng et al., 2025; Hu et al., 2025), through asynchronous actor–learner pipelines (Noukhovitch et al., 2025; Fu et al., 2025; Wu et al., 2025), disaggregated and streaming generation (Zhong et al., 2025), in-flight weight updates (Piché et al., 2026), or partial rollouts (Kimi Team, 2025). ECHO-2 (Song et al., 2026) distributes rollouts over wide-area inference workers and makes bounded staleness an explicit, user-controlled parameter, trading it against dissemination latency and cost. Such systems handle the resulting staleness in one of two ways. They either cap the lag at a hand-tuned value (Fu et al., 2025; Khatri et al., 2026; Song et al., 2026) or restore on-policy updates at every iteration (Lu & Luo, 2026), or they modify the objective so that stale data can be used, through decoupled proximal policies (Hilton et al., 2022; Fu et al., 2025; Li et al., 2026), tapered importance weights (Roux et al., 2025), second-moment trust regions (Zheng et al., 2026), relaxed clipping (Tian et al., 2026), or projection of stale-aligned gradients (Xu et al., 2026). *Difference:* we keep vanilla GRPO without additional staleness correction and characterize an approximately constant-$S\eta$ stability boundary in the Llama sweeps, together with a collapse clock in $\sum_t \eta_t$ under the stated drift conditions. These findings provide a baseline against which such corrections can be measured, while our conditional readout model explains how the consecutive-gradient cosine used by Xu et al. (2026) can provide only a late warning near the boundary.

Off-policy corrections in distributed RL. Decoupling acting from learning is the central design pattern of distributed deep RL, from A3C (Mnih et al., 2016) to IMPALA (Espeholt et al., 2018) and SEED RL (Espeholt et al., 2020), and the resulting policy lag is controlled with truncated importance weights such as Retrace and V-trace (Munos et al., 2016; Espeholt et al., 2018) or with sample-reuse variants of PPO (Queeney et al., 2021). Other systems avoid the lag altogether by remaining synchronous (Wijmans et al., 2020), and reproducibility studies show how stale rollouts perturb actor–learner PPO dynamics (Huang et al., 2024). *Difference:* our vanilla GRPO setup uses the clipped ratio and a KL penalty without additional staleness correction; we characterize its stability as a function of (S, η) with other settings fixed.

Analyses of delayed and off-policy updates: what is guaranteed. Asynchronous SGD is analyzed as gradient descent on a fixed objective with gradients evaluated at delayed iterates, and convergence holds when the step size shrinks with the maximum delay (Niu et al., 2011; Lian et al., 2015; Zhang et al., 2016; Dutta et al., 2018; Koloskova et al., 2022; Mishchenko et al., 2022); asynchrony has also been shown to act as an implicit momentum (Mitliagkas et al., 2016). For GRPO, recent analyses show that an update improves the objective when the behavior policy stays close to the learner (Mroueh et al., 2026), and that group-relative REINFORCE admits an off-policy interpretation in which regularizing the update is what tolerates the mismatch (Yao et al., 2026). *Difference:* these are sufficient conditions, for convergence or for a single improvement step. We complement these guarantees by examining collapse in vanilla GRPO, where delay also shifts the sampling distribution and importance ratios. The Llama sweeps exhibit an approximately constant-$S\eta$ stability boundary and inverse-learning-rate collapse timing that is insensitive to S across the tested unstable configurations.

Optimizer dynamics: interpreting the collapse clock. Adam can exhibit sign-like behavior through coordinatewise normalization (Bernstein et al., 2018; Balles & Hennig, 2018; Kunstner et al., 2023), and raw gradient scales in transformers differ widely across parameter blocks (Zhang et al., 2024). Learning-rate thresholds such as the edge of stability (Cohen et al., 2021), scaling rules for adaptive methods (Malladi et al., 2022), and learning-rate sensitivity studies of large transformers (Wortsman et al., 2024) tie stability to optimizer settings. *Difference:* our conditional model uses these properties to explain two staleness-specific effects: when persistent drift has comparable saturated speed and exit geometry across S, the leading collapse budget in $\sum_t \eta_t$ is approximately shared; heterogeneous gradient scales can mask coherent normalized updates in the raw-gradient cosine under the stated readout conditions.

E Additional Training Curves

Figure 6 shows the 1B sweep summarized in Table 1; the remaining figures split the training reward of both Llama sweeps by staleness.

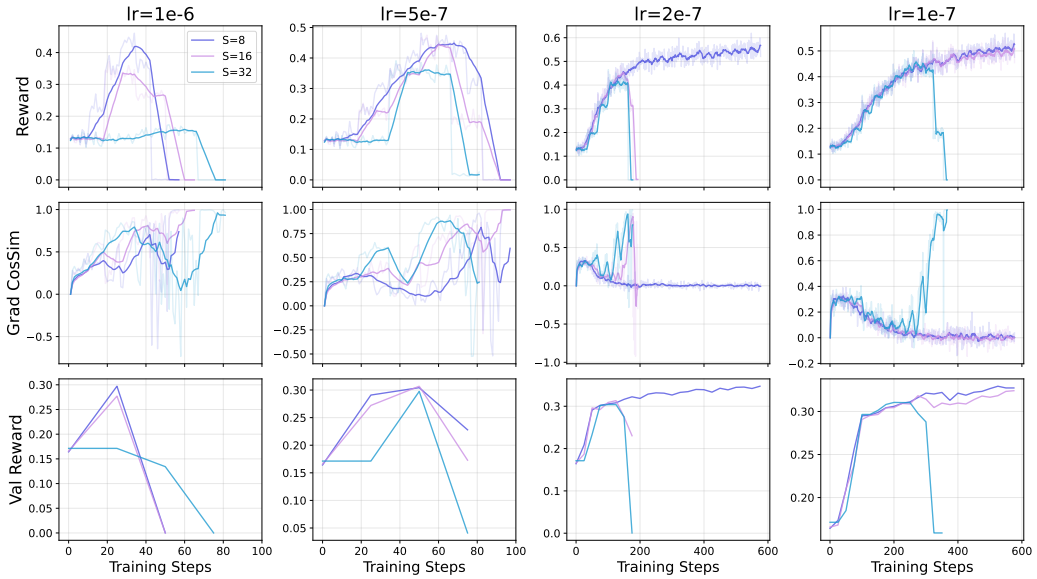


Figure 6: Staleness–learning-rate sweep on `Llama-3.2-1B-Instruct`, with the layout of Figure 1. The collapse counts coincide with those of the 3B policy (Table 1), so both sizes share the boundary in $S\eta$ and the clock constant $c_2 = 3.2 \times 10^{-5}$.

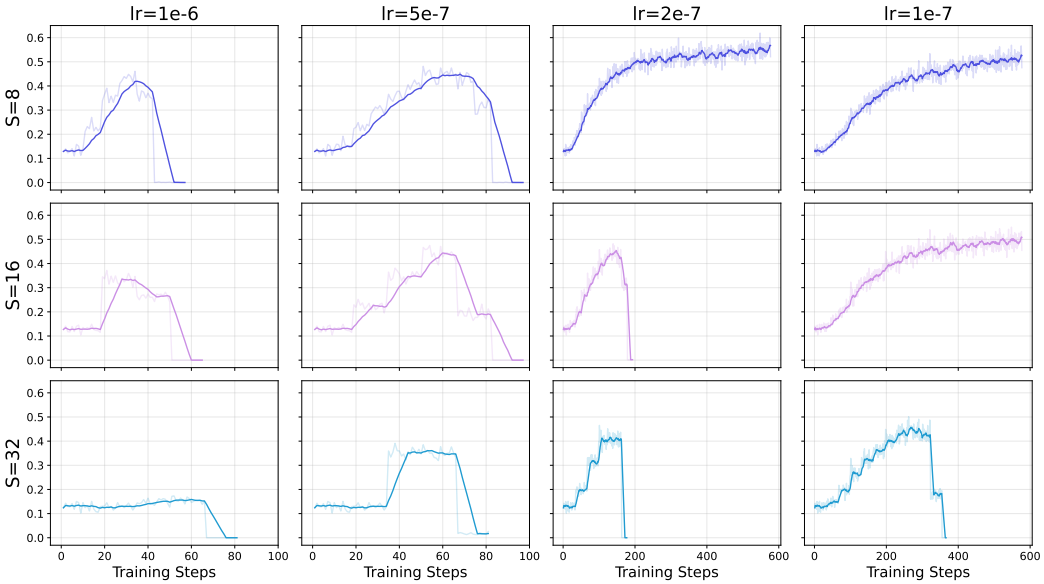


Figure 7: Per-staleness training-reward curves on `Llama-3.2-1B-Instruct`, separated into individual panels for legibility (companion to Figure 6). Each panel fixes one (S, η) pairing; collapse corresponds to an abrupt, severe drop in training reward. The onset of collapse is consistent with the collapse-time scaling $t_{\text{obs}}\eta \approx c_2$ reported in Table 1.

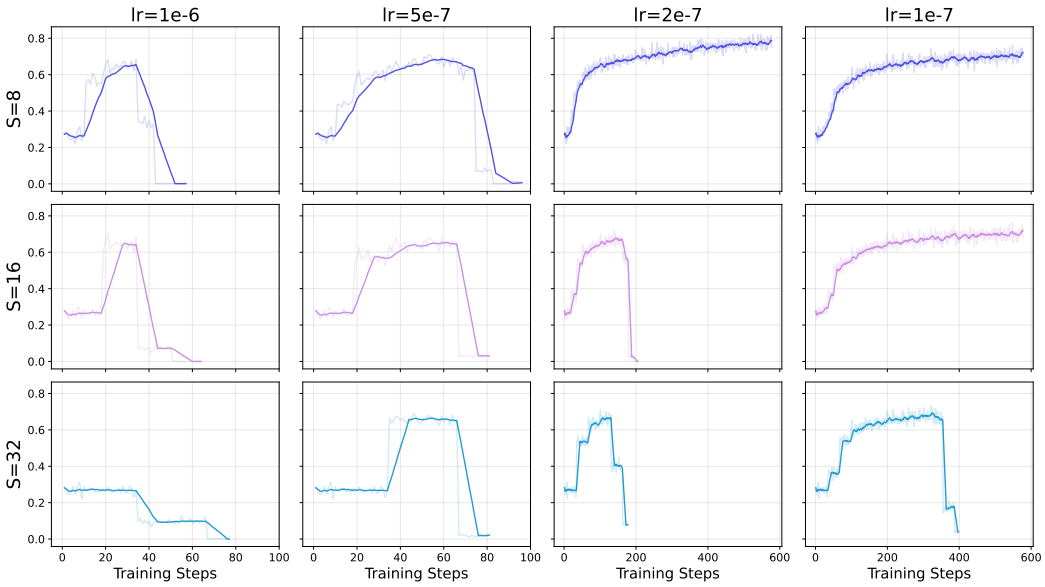


Figure 8: Per-staleness training-reward curves on `Llama-3.2-3B-Instruct`, separated into individual panels for legibility (companion to Figure 1). As with the 1B policy, the collapse onset across (S, η) is consistent with the approximately S-independent collapse-time scaling $t_{\text{obs}}\eta \approx c_2$ (Section 4.3).