# GCD: Correcting Hidden-State Bias in Off-Policy Agentic RL

**Changyuan Chen**[1,2*] **Jianyu Xiang**[1,*] **Jiasheng Luo**[1,*] **Ziye Wang**[1,*]
**Nallappan Gunasekaran**[1†]
[1]Eastern Michigan University [2]Tsinghua University
{cchen22, jxiang1, jluo6, zwang43}@emich.edu gunasmaths@gmail.com
https://github.com/EricChan2028/GCD-KV-Drift

## Abstract

Off-policy reinforcement learning (RL) for large language models (LLMs) bottlenecks at the rollout stage: every parameter update invalidates the key-value (KV) cache of in-flight agent trajectories, forcing either an expensive full recompute or staleness that destabilizes training. Recent methods such as M2PO tolerate this staleness by reweighting the policy gradient with importance ratios. However, these ratios assume that the reused forward pass computes the new policy's probability. Stale keys and values make it compute a hybrid of the old and new policies instead, leaving a *hidden-state bias* that skews the policy gradient. This bias, the log-probability gap between the hybrid and the new policy, evades every ratio-based diagnostic. In this work, we measure it on two backbones, Qwen3.5-9B and the pure Transformer Llama-3.1-8B. It grows with staleness and tracks a cheap estimate: how far the weights moved times how much a token attends to stale context. Building on this estimate, we introduce *Gradient-Coupled KV Drift* (GCD), which divides the bias out of the importance ratio and recomputes only the tokens where the estimate fails. The mechanism is architecture-independent, while our numbers are scoped to the single-node configurations we tested. On Llama-3.1-8B with MATH-500, GCD stays stable at a staleness of $512$ updates, at $1.27\times$ the throughput of full recompute but $5\%$ below naive reuse. On Qwen3.5-9B, we demonstrate that GCD cuts the ratio error of naive reuse by $46$–$56\%$ and survives staleness $512$ on MATH-500, where M2PO collapses. It also raises rollout throughput $1.38\times$ over full recompute on $\tau$-bench retail and lifts retail $\mathrm{pass}^4$ by $5.7$ points over M2PO at staleness $256$.

## 1 Introduction

Reinforcement learning (RL) post-training has become the standard route to elicit complex reasoning and tool use from large language models (LLMs), with Group Relative Policy Optimization (GRPO) [Shao et al., 2024] and its descendants powering recent advances such as DeepSeek-R1 [Guo et al., 2025]. The bottleneck in this regime is no longer the gradient step but the rollout. Each policy update invalidates the key-value (KV) cache of every in-flight agent trajectory, and on agentic benchmarks such as $\tau$-bench [Yao et al., 2024] those trajectories run for thousands of tokens. Production frameworks such as verl [Sheng et al., 2024] and OpenRLHF [Hu et al., 2025] therefore aggressively reuse stale KV cache across updates. A parallel literature on second-moment correction [Zheng et al., 2025], decoupled clipping [Yu et al., 2025], and variance-controlled asynchrony [Huang et al., 2026, Xu et al., 2026] attempts to keep the resulting off-policy gradient stable.

*Equal contribution (co-first authors). This work was completed by all authors at Eastern Michigan University; the rebuttal was completed by Changyuan Chen during an internship at Tsinghua University.

†Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This combination has become so ubiquitous that an assumption underlying all token-level importance correction has gone unexamined: *after an update, the log-probability computed for a token is the new policy's log-probability*. We show that this assumption fails whenever a KV cache is reused across an update. The forward pass consumes stale keys and values at every full-attention layer, but it uses the new query, output, and feed-forward projections. The resulting distribution is neither $\pi_{\theta_{\text{old}}}$ nor $\pi_{\theta_{\text{new}}}$ but a hybrid $\pi_{\text{hyb}}$ whose log-probability gap to $\pi_{\theta_{\text{new}}}$ is a hidden, uncorrected source of policy-gradient bias. We call this gap the *hidden-state bias*, and the importance ratio cannot detect it. Its denominator is a number stored at rollout time and its numerator comes from the hybrid pass. Clipping, second-moment bounds, and KL penalties therefore all act on a well-behaved ratio for the wrong target policy. Empirically on Qwen3.5-9B [Yang et al., 2025], the median gap reaches $0.034$ nats per token at a staleness of $256$ updates and $0.108$ at $512$, with a long tail on agent-reasoning spans. M2PO [Zheng et al., 2025] identifies staleness $256$ as its ceiling, and we attribute a substantial fraction of this ceiling to the uncorrected bias.

We propose *Gradient-Coupled KV Drift* (GCD), a hidden-state-aware importance correction. To first order, the bias is how far the weights moved times how much a token leans on stale context. Training already exposes both factors, so GCD can estimate the bias and divide it out of the ratio (Figure 1). Rather than pursue a worst-case bound, which breaks down once attention concentrates on a single position, we construct an *exception-safe* framework. A high-probability estimate covers typical attention. A tractable detector flags exactly the inputs on which that estimate is uninformative. Beyond identifying the hybrid distribution (Section 3), our contributions are:

- **A high-probability bias bound.** To our knowledge, Theorem 5 is the first off-policy bound that distinguishes positional KV drift from recurrent-state drift while remaining honest about its regime of validity.
- **Tail-event detector and selective recompute.** GCD trusts a correction only when the estimate and the detector agree (Theorem 6 and Corollary 7). Flagged tokens recompute only their offending layer-position pairs, adding $1.9\%$ of the FLOPs of full recompute on Qwen3.5-9B at staleness $512$.
- **Composition with M2PO.** On safe tokens the correction inflates the second moment by at most $10.5\%$, so the M2PO trust region applies with a small threshold adjustment (Theorem 8).
- **A ground-truth check.** For a token sampled from the hybrid, the gap is a KL divergence in expectation (Proposition 9), and full recompute confirms that the correction points the right way on average.
- **Capability claim, not system optimization.** GCD trains stably at staleness $512$ on MATH-500 [Hendrycks et al., 2021] and $\tau$-bench retail, beyond M2PO's collapse. On a pure Transformer stability survives, while the efficiency claim weakens (Section 6).

## 2 Background and Related Work

**Off-policy importance correction.** PPO [Schulman et al., 2017], GRPO [Shao et al., 2024], and DAPO [Yu et al., 2025] are the algorithmic backbone of RL for LLMs. Each stale token carries a log-probability stored at rollout, and the importance ratio $w_t = \pi_{\theta_{\text{new}}}/\pi_{\theta_{\text{old}}}$ compares it with the one the current model computes. The most direct precursor is M2PO [Zheng et al., 2025], which clips the second moment $\mathbb{E}[w_t^2]$ rather than $w_t$, sustaining staleness up to $256$ on Qwen2.5-32B. SOUP [Yang et al., 2026] mixes on- and off-policy gradients per token; GAC [Xu et al., 2026] aligns gradient directions across asynchrony; Stable Asynchrony [Huang et al., 2026] provides variance-controlled estimators. Preference-optimization variants [Xiao et al., 2024, Khaki et al., 2024, Xie et al., 2024, Wallace et al., 2024] adapt the same framework to bandit-style training. All of these take $w_t$ as atomic. We instead identify a bias invisible at the level of $w_t$, and GCD composes with these methods.

**KV caches.** PagedAttention [Kwon et al., 2023], Mell [Liu et al., 2025], and FlashAttention [Dao et al., 2022, Shah et al., 2024] treat the cache as a faithful function of fixed parameters. Our setting is the opposite: the cache is a snapshot of old parameters whose reuse biases the gradient.

**Hybrid architectures.** Qwen3.5 [Yang et al., 2025] interleaves three Gated DeltaNet layers [Yang et al., 2024b] with one full-attention layer. Gated DeltaNet maintains a recurrent state rather than

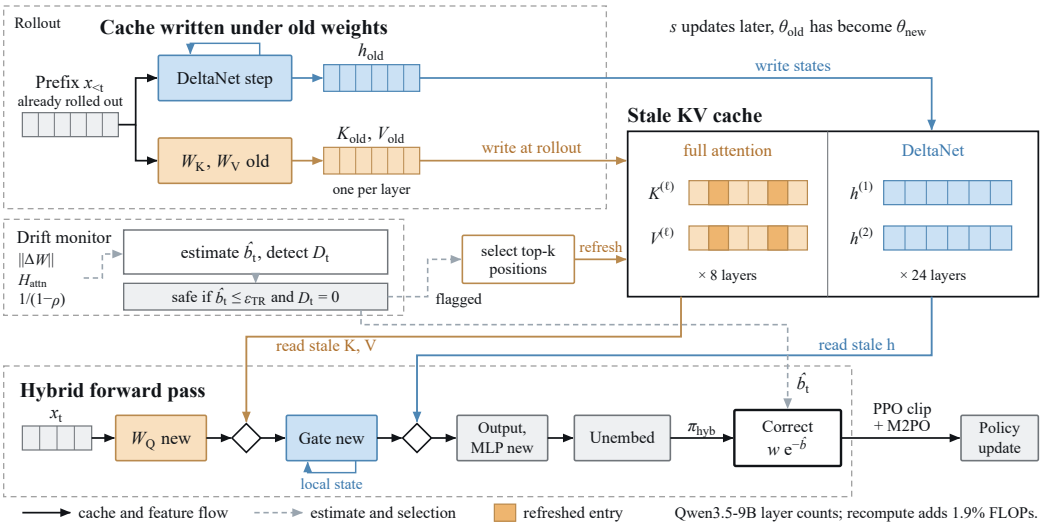



Figure 1: **GCD in one training step.** After $s$ updates, the forward pass reads the stale cache at two merge points, so it computes $\pi_{\text{hyb}}$ rather than $\pi_{\theta_{\text{new}}}$; the bottom row traces one token. The drift monitor estimates the gap from per-layer statistics, refreshes the top positions of flagged tokens, and corrects the importance ratio. Orange marks the full-attention path and blue the DeltaNet path.

Table 1: **Weight provenance of the operations that generate $y_t$ after an update.** Naive reuse mixes old and new weights inside one pass, so it computes neither $\log \pi_{\theta_{\text{old}}}$ nor $\log \pi_{\theta_{\text{new}}}$.

| Factor | Naive reuse | $\log \pi_{\theta_{\text{old}}}$ needs | $\log \pi_{\theta_{\text{new}}}$ needs |
|---|---|---|---|
| Query projection $W_Q$ | new | old | new |
| Keys and values $K_p, V_p$ at every cached $p < t$ | **old** | old | new, by a full re-run |
| DeltaNet recurrent state $h$ | **old** | old | new, by a full re-run |
| Output projection, MLP, gate, unembedding | new | old | new |

per-position KV, changing the bias from positional drift to per-layer state drift; our bound treats both. Llama-3.1-8B [Grattafiori et al., 2024], our second backbone, uses full attention only.

**Asynchronous systems and agentic RL.** On the systems side, asynchronous RLHF [Noukhovitch et al., 2024] and HybridFlow [Sheng et al., 2024] pioneered decoupled rollout-train queues, treating full recompute as a mandatory cost; GCD makes that cost a function of the trust region. For evaluation, $\tau$-bench [Yao et al., 2024] provides the canonical retail and airline benchmark with $\text{pass}^k$ reliability, the share of tasks an agent solves in all of $k$ attempts. Modecrua et al. [2026] show that long trajectories make rollout cost the dominant scaling axis. Section C reviews further work.

## 3 The Hidden Bias: A Hybrid Distribution

After an update we need $\log \pi_{\theta_{\text{new}}}(y_t \mid x_{<t})$ for a token that was already rolled out. *Full recompute* re-runs the prefix through the new weights, exact but at a cost of $\Theta(tdL)$. *Naive reuse* keeps the stale keys $K$, values $V$, and recurrent states $h$, applying the new query, output, and MLP. At $\Theta(dL)$, it is what production systems use.

Naive reuse does not compute $\log \pi_{\theta_{\text{new}}}$. It composes new projections with old memory, as Table 1 shows, and that composition defines the hybrid distribution $\pi_{\text{hyb}}$. We define the *hidden-state bias* of a token as the gap

$$b_t := \log \pi_{\text{hyb}}(y_t \mid x_{<t}) - \log \pi_{\theta_{\text{new}}}(y_t \mid x_{<t}). \tag{1}$$

Figure 2 shows the whole problem: the ratio assumes two clean passes, each with one set of weights, and the real pass is neither. Reuse is not equally harmful everywhere, though. Where attention spreads over many positions, drift arrives thinly and a cheap estimate can bound it; where attention peaks on

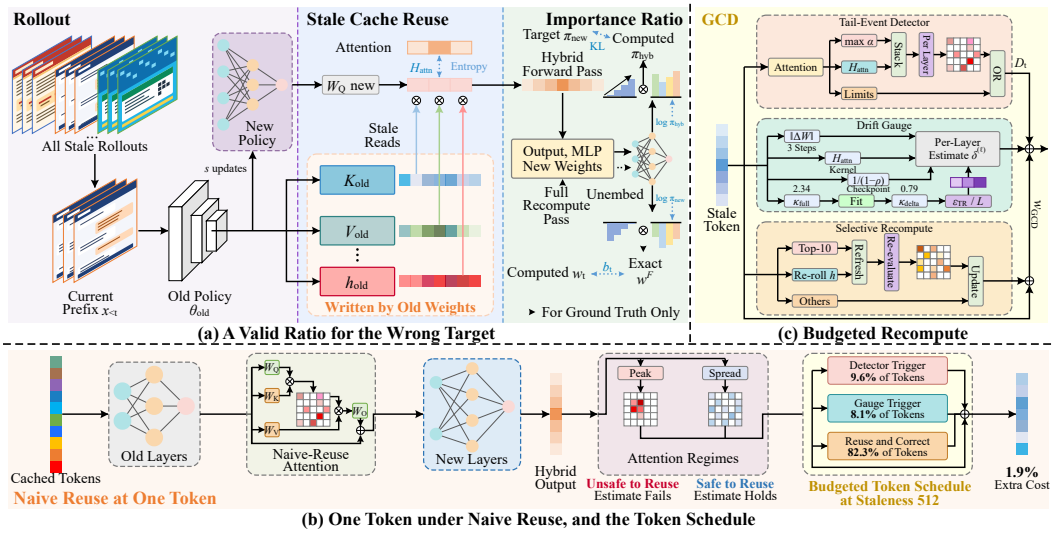



Figure 2: **Why reusing a stale cache fails, and how GCD budgets the fix.** (a) The old policy writes the cache and the new query reads it after $s$ updates, so the ratio is computed on $\pi_{\text{hyb}}$ while the target is $\pi_{\theta_{\text{new}}}$. (b) Inside one token, query and output are new while keys and values are old. Peaked attention makes stale entries unsafe to reuse; spread attention leaves them safe. (c) GCD detects tail tokens, gauges drift per layer, and recomputes only what it flags. Thumbnails, heat maps, and charts are illustrative; token shares and 1.9% extra cost are measured on Qwen3.5-9B at staleness 512.

a drifted entry, the estimate fails. GCD exploits this asymmetry, correcting where the estimate holds and recomputing where it does not.

**A worked example with two tokens.** Let token 1 be cached under the old weights, so that $K_{\text{old},1} = W_K^{\text{old}} x_{\text{old},1}$ and $V_{\text{old},1} = W_V^{\text{old}} x_{\text{old},1}$, and let an update arrive before token 2 is generated. Naive reuse forms the new query $q_2 = W_Q^{\text{new}} x_2$, attends over the stale entry and token 2's own fresh entry, and passes $\alpha_{2,1} V_{\text{old},1} + \alpha_{2,2} V_{\text{new},2}$ through the new output projection, MLP, gate, and unembedding. Reaching $\log \pi_{\theta_{\text{old}}}$ would need $W_Q^{\text{old}}$. Reaching $\log \pi_{\theta_{\text{new}}}$ would need $K_{\text{new},1}$ and $V_{\text{new},1}$, which means re-running token 1, and over a long prefix that is full recompute. To first order, $b_2 \approx \langle \nabla_{K_1} \log \pi_{\text{hyb}}, K_{\text{old},1} - K_{\text{new},1} \rangle + \langle \nabla_{V_1} \log \pi_{\text{hyb}}, V_{\text{old},1} - V_{\text{new},1} \rangle$. The gap therefore scales with the weight movement and with the attention $\alpha_{2,1}$ paid to the stale entry.

**Remark: why token-level correction is blind.** Since $\log \pi_{\theta_{\text{old}}}$ is a scalar stored at rollout and the numerator is the hybrid pass, the computed ratio is exactly $\pi_{\text{hyb}}/\pi_{\theta_{\text{old}}}$, a well-behaved importance ratio for the wrong target policy. The gap does not inflate $\mathbb{E}[w_t^2]$; it displaces the *estimand*. Every diagnostic built on $w_t$, whether clip fraction, second moment, or KL, is blind by construction, which is why this failure mode can survive in production.

## 4 Method: Gradient-Coupled KV Drift

**How large is the gap?** Start from one cached key. After an update, the fresh key differs from the stale one by about $\Delta W_K \, x_p$, so the drift in a key is proportional to how far the weights moved. A token feels this drift only through the attention it pays to stale entries. The wider that attention spreads over cached positions, the more stale context flows in, and attention entropy measures the spread. A DeltaNet layer has no per-position cache, but its state carries drift forward for about $1/(1-\rho)$ steps, where $\rho$ is the spectral radius of its linearized recurrence. Summing over layers gives

$$\widehat{b}_t \;=\; \kappa_{\text{full}} \sum_{\ell \in \mathcal{L}_{\text{full}}} \left\| \Delta W^{(\ell)} \right\|_{\text{op}} \mathcal{H}_{\text{attn}}^{(\ell)}(x_{<t}) \;+\; \kappa_{\text{delta}} \sum_{\ell \in \mathcal{L}_{\text{delta}}} \frac{\left\| \Delta W^{(\ell)} \right\|_{\text{op}}}{1 - \rho^{(\ell)}}. \tag{2}$$

**Algorithm 1** One GCD training step on a batch of stale rollouts.

**Require:** Stale rollouts with stored $\log \pi_{\theta_{\text{old}}}$ and cached $K_{\text{old}}, V_{\text{old}}, h_{\text{old}}$; last update $\Delta W^{(\ell)}$; advantages $\hat{A}_t$
**Require:** Tolerance $\epsilon_{\text{TR}}$, threshold $\widetilde{\epsilon}_{M2}$, clip $\epsilon$, weight $\beta$, regularity $(H_0, \alpha_{\max})$

For each layer, $\sigma^{(\ell)} \leftarrow$ three steps of power iteration on $\Delta W^{(\ell)}$, and read $\rho^{(\ell)}$ if $\ell \in \mathcal{L}_{\text{delta}}$
**for** each token $t$ **do** ▷ hybrid pass
  Compute $\log \pi_{\text{hyb}}(y_t \mid x_{<t})$ from stale $K, V, h$, and keep $\mathcal{H}^{(\ell)}_{\text{attn}}$ and the largest attention weight
  $\widehat{b}_t \leftarrow$ Equation (2) with $\sigma^{(\ell)}$ for $\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$, and $D_t \leftarrow$ Equation (6)
**for** each token $t$ with $D_t = 1$ or $\exists \ell : \delta^{(\ell)}_t > \epsilon_{\text{TR}}/L$ **do** ▷ selective recompute
  In each flagged full-attention layer, refresh $K^{(\ell)}_p, V^{(\ell)}_p$ at the top-$k$ attention positions
  In each flagged DeltaNet layer, re-roll $h^{(\ell)}$ from the most recent clean checkpoint
  Re-evaluate $\log \pi_{\text{hyb}}(y_t \mid x_{<t})$, $\widehat{b}_t$, and $D_t$
$w^{\text{GCD}}_t \leftarrow \exp\big(\log \pi_{\text{hyb}}(y_t \mid x_{<t}) - \log \pi_{\theta_{\text{old}}}(y_t \mid x_{<t}) - \widehat{b}_t\big)$ for every token
Update $\theta_{\text{new}}$ by one ascent step on $\mathcal{J}_{\text{GCD}}$ of Equation (4) with threshold $\widetilde{\epsilon}_{M2}$

Every term is weight movement times sensitivity to stale context, and on a Transformer-only model the second sum is empty. The inputs are nearly free: three steps of power iteration per layer give $\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$, the attention kernel gives $\mathcal{H}^{(\ell)}_{\text{attn}}$, and one power iteration per checkpoint gives $\rho^{(\ell)}$.

**Why does subtracting it work?** By definition $\log \pi_{\theta_{\text{new}}} = \log \pi_{\text{hyb}} - b_t$, so dividing the computed ratio by $\exp(b_t)$ recovers the true ratio. We do not know $b_t$ exactly, but we know its sign on average, because its mean is a KL divergence (Proposition 9). A non-negative estimate is therefore the right correction on average, which gives the GCD weight

$$w^{\text{GCD}}_t := \frac{\pi_{\text{hyb}}(y_t \mid x_{<t})}{\pi_{\theta_{\text{old}}}(y_t \mid x_{<t})} \cdot \exp(-\widehat{b}_t). \tag{3}$$

This is a statement about the mean. On tokens with $b_t < 0$ the factor moves the wrong way, and we measure the cost of that in Section 6.

**Can the estimate be trusted?** The estimate scales with attention entropy, but a token that attends to one position has almost no entropy and still reads that position's drift in full. The first-order estimate therefore fails exactly when attention peaks. Peaks, however, are easy to see: the detector $D_t$ of Equation (6) reads the largest attention weight and the entropy that the pass has already produced. GCD calls a token safe only when both $\widehat{b}_t \le \epsilon_{\text{TR}}$ and $D_t = 0$.

### 4.1 What to do with flagged tokens

Write $\widehat{b}_t = \sum_\ell \delta^{(\ell)}_t$, with one term per layer. GCD recomputes when $\delta^{(\ell)}_t > \epsilon_{\text{TR}}/L$ or when $D_t = 1$, working only on the offending layers, because that is where the drift enters. A few positions hold most of the attention, so a full-attention layer refreshes $K$ and $V$ at the top-$k$ attention positions with $k = \lceil 1/(1-\alpha_{\max}) \rceil$. A DeltaNet layer instead re-rolls $h^{(\ell)}$ from the most recent clean checkpoint.

We report cost with four quantities. $T_{\text{det}} = \Pr[D_t{=}1]$ is the detector's trigger rate and $T_{\text{gauge}} = \Pr[\exists \ell : \delta^{(\ell)}_t > \epsilon_{\text{TR}}/L,\ D_t{=}0]$ the gauge's. Their sum $T_{\text{tok}} = T_{\text{det}} + T_{\text{gauge}}$ is the share of tokens entering selective recompute, whereas $C_F$ is the extra FLOPs over naive reuse as a share of full recompute. These differ widely: on Qwen3.5-9B at staleness 512, $T_{\text{tok}} = 17.7\%$ yet $C_F = (5.1 - 4.3)/42.7 = 1.9\%$. The KV-refresh part of $C_F$ scales with $|\mathcal{L}_{\text{full}}|/L$, DeltaNet re-rolls add to it, and at an extreme trigger rate recompute can cost more than a full re-run (Section B.7).

### 4.2 The objective and calibration

With $\hat{A}_t$ the GRPO group-relative advantage, GCD plugs into the usual objective over stale rollouts:

$$\mathcal{J}_{\text{GCD}}(\theta) = \mathbb{E}_\tau \left[ \tfrac{1}{T} \sum_t \min\Big( w^{\text{GCD}}_t \hat{A}_t,\ \text{clip}(w^{\text{GCD}}_t, 1{-}\epsilon, 1{+}\epsilon) \hat{A}_t \Big) \right] - \beta\, \mathbb{E}[(w^{\text{GCD}}_t)^2] \cdot \mathbf{1}[\mathbb{E}[w^2_t] > \widetilde{\epsilon}_{M2}], \tag{4}$$

where $\widetilde{\epsilon}_{M2} := \epsilon_{M2} \exp(2B_b)$. For $B_b \le 0.05$, which holds on every safe token, this raises the M2PO threshold by at most $10.5\%$, from $1.5$ to $1.66$. A factor bounded by $\exp(B_b)$ can only move the second moment that far (Theorem 8). Setting $\widehat{b}_t = 0$ recovers M2PO, and setting $\beta = 0$ gives GCD alone. Algorithm 1 lists one full step.

The constants are fitted once per backbone by least squares on $512$ probe prompts, comparing full recompute with naive reuse on tokens with $D_t = 0$. This takes $0.93$ H100-hours on Qwen3.5-9B. Calibration buys efficiency, not validity: Theorem 5 holds with the analytical constants, and fitting only tightens them to a usable scale. Over-estimating $\kappa$ costs compute; under-estimating it is what $D_t$ and $\epsilon_{\mathrm{TR}}$ must contain. Re-fits deviate by at most $9\%$, most at early RL (Table 6), so we recommend calibrating at the earliest checkpoint.

## 5 Theory: Exception-Safe, Not Worst-Case

We prove a bound for typical attention, a detector for the rest, and composition with M2PO, leaving the proofs to Section A. Three assumptions are standard, apart from the upstream condition in Assumption 1. The fourth is not, so we measure its tail $\hat{q}$ empirically.

**Assumption 1** (Smoothness)**.** Each layer is $C^2$ in its weights on a convex neighborhood $\mathcal{N}_\theta \ni \{\theta_{\mathrm{old}}, \theta_{\mathrm{new}}\}$. The Hessian of $\log \pi_\theta$ has spectral norm at most $G$ on $\mathcal{N}_\theta$, and residual-stream norms are at most $B$. Upstream drift is dominated: at every layer, the drift in $K$, $V$, or $h$ caused by the change in the input stream is at most $\lambda$ times the drift caused by $\Delta W^{(\ell)}$.

**Assumption 2** (DeltaNet contractivity)**.** For $\ell \in \mathcal{L}_{\mathrm{delta}}$, the linearized state transition has spectral radius $\rho_\theta^{(\ell)} \le \rho_{\max} < 1$ on $\mathcal{N}_\theta$, and $\theta \mapsto \rho_\theta^{(\ell)}$ is $L_\rho$-Lipschitz.

**Assumption 3** (RoPE sensitivity)**.** Full-attention layers use RoPE [Su et al., 2024] with base $\omega$, and the attention Jacobian with respect to keys at position $p$ has operator norm at most $C_{\mathrm{RoPE}}(\omega)\, \alpha_p$.

**Assumption 4** (Regularity)**.** Per layer, $\boldsymbol{\alpha}_t^{(\ell)} \in \mathcal{R}(H_0, \alpha_{\max}) := \{\boldsymbol{\alpha} : \mathcal{H}_{\mathrm{attn}} \ge H_0,\ \max_p \alpha_p \le \alpha_{\max} < 1\}$ with probability at least $1 - q$.

**Theorem 5** (High-probability concentration)**.** *Under Assumptions 1 to 4, on the regular event* $\mathcal{E} := \bigcap_\ell \{\boldsymbol{\alpha}_t^{(\ell)} \in \mathcal{R}(H_0, \alpha_{\max})\}$,

$$|b_t|\, \mathbf{1}[\mathcal{E}] \;\le\; \kappa_{\mathrm{full}} \sum_{\ell \in \mathcal{L}_{\mathrm{full}}} \left\| \Delta W^{(\ell)} \right\|_{\mathrm{op}} \mathcal{H}_{\mathrm{attn}}^{(\ell)}(x_{<t}) \;+\; \kappa_{\mathrm{delta}} \sum_{\ell \in \mathcal{L}_{\mathrm{delta}}} \frac{\left\| \Delta W^{(\ell)} \right\|_{\mathrm{op}}}{1 - \rho^{(\ell)}} \;+\; \tfrac{1}{2} G \, \|\Delta W\|_{\mathrm{op}}^2 , \tag{5}$$

*with* $\kappa_{\mathrm{full}} = (1+\lambda)\, C_{\mathrm{RoPE}}(\omega)\, B\, [(1-\alpha_{\max}) \log(1/\alpha_{\max})]^{-1} \sqrt{2/g}$ *for GQA head ratio* $g$*, and* $\kappa_{\mathrm{delta}} = (1+\lambda)\, G_h L_\rho$ *with* $G_h = O(\sqrt{r})$ *for kernel rank* $r$*. Moreover* $\Pr[\mathcal{E}] \ge 1 - Lq$.

The bound makes the two forms of drift explicit. Its DeltaNet term diverges as a layer's spectral radius approaches the contractivity boundary, so we verify $\rho_{\max} = 0.83$ on Qwen3.5-9B. The bound holds on $\mathcal{E}$ only, and the detector below handles the rest. With these analytical constants it is a valid inequality; with fitted constants, Equation (2) is an estimate, not a certified bound.

**Theorem 6** (Tail-event detector)**.** *Define*

$$D_t \;:=\; \mathbf{1}\left[ \exists\, \ell \in \mathcal{L}_{\mathrm{full}} :\; \max_p \alpha_{t,p}^{(\ell)} > \alpha_{\max} \;\vee\; \mathcal{H}_{\mathrm{attn}}^{(\ell)} < H_0 \right]. \tag{6}$$

*Then* $D_t = 1$ *on* $\mathcal{E}^c$*, and* $D_t = 0$ *implies that Theorem 5 applies. On* $\{D_t = 1\}$*, refreshing the top-*$k$ *positions of a layer that violates regularity leaves attention mass at most* $\bar{m} := \max\{1 - \alpha_{\max},\ H_0 / \log(k+1)\}$ *on its stale entries, so the first-order bound of that layer shrinks by at least a factor of* $(1+\lambda)/(\lambda + \bar{m})$*, which is* $1/\bar{m}$ *at* $\lambda = 0$*. The remainder* $\frac{1}{2} G \, \|\Delta W\|_{\mathrm{op}}^2$ *and the terms of layers that are not refreshed are unchanged.*

The detector reuses the attention statistics of the forward pass and does not depend on $\kappa$.

**Corollary 7** (Composite GCD safety)**.** *Let* $\widehat{b}_t$ *use constants that upper-bound the right side of Equation (5). Then for any* $\epsilon_{\mathrm{TR}} > \frac{1}{2} G \, \|\Delta W\|_{\mathrm{op}}^2$,

$$\Pr\big[\, |b_t| > \epsilon_{\mathrm{TR}} \,\big|\, \widehat{b}_t \le \epsilon_{\mathrm{TR}} \wedge D_t = 0 \,\big] \;\le\; Lq. \tag{7}$$

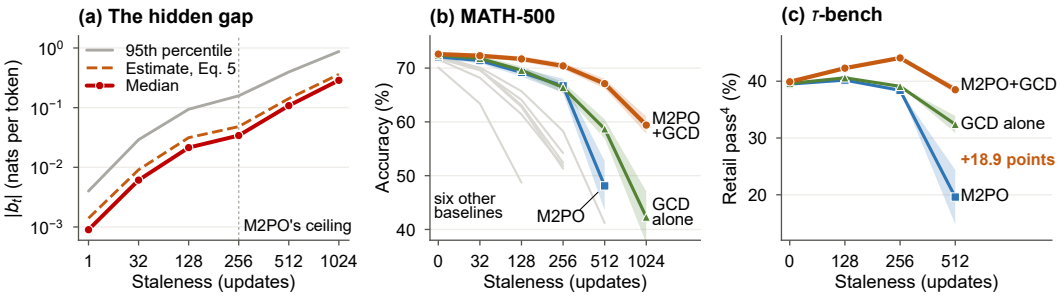



Figure 3: **The hidden gap grows with staleness, and correcting it keeps training stable.** (a) Bias on Qwen3.5-9B with the estimate of Equation (5), remainder included. (b) MATH-500 accuracy. (c) $\tau$-bench retail $\mathrm{pass}^4$. Bands show one standard deviation where reported.

**Theorem 8** (GCD preserves the M2PO trust region)**.** *For $|\widehat{b}_t| \le B_b$ and sub-Gaussian $w_t$ with parameter $\sigma^2$, the ratio $w_t^{\mathrm{GCD}} = w_t \exp(-\widehat{b}_t)$ is sub-Gaussian with parameter $\sigma^2 \exp(2B_b)$, and $\mathbb{E}[(w_t^{\mathrm{GCD}})^2] \le \exp(2B_b)\mathbb{E}[w_t^2]$. Applying M2PO with threshold $\epsilon_{M2} \exp(2B_b)$ to $w_t^{\mathrm{GCD}}$ yields the same contraction as M2PO with threshold $\epsilon_{M2}$ on $w_t$.*

None of these results uses the hybrid structure. Without DeltaNet layers the second sum of Equation (5) is empty, and $\kappa_{\mathrm{full}}$ is defined for standard softmax attention with RoPE. The 3:1 ratio buys efficiency, not validity (Remark A.2).

**Proposition 9** (Sign of the bias in expectation)**.** *For a fixed prefix $x_{<t}$ and $y_t \sim \pi_{\mathrm{hyb}}(\cdot \mid x_{<t})$, we have $\mathbb{E}[b_t \mid x_{<t}] = \mathrm{KL}\big(\pi_{\mathrm{hyb}}(\cdot \mid x_{<t}) \,\|\, \pi_{\theta_{\mathrm{new}}}(\cdot \mid x_{<t})\big) \ge 0$, with equality iff they coincide.*

Tokens generated after an update are sampled from $\pi_{\mathrm{hyb}}$ itself, so a non-negative correction matches the sign of the mean. A token generated before the update is sampled from $\pi_{\theta_{\mathrm{old}}}$ instead, and its mean gap is $\mathrm{KL}(\pi_{\theta_{\mathrm{old}}} \,\|\, \pi_{\theta_{\mathrm{new}}}) - \mathrm{KL}(\pi_{\theta_{\mathrm{old}}} \,\|\, \pi_{\mathrm{hyb}})$, which has no fixed sign. The proposition also names a better regression target: $\widehat{b}_t$ should estimate $\mathbb{E}[b_t \mid \text{features}]$, whereas fitting $|b_t|$ over-estimates it by $\mathbb{E}|b_t| - \mathbb{E}[b_t] \ge 0$. We call the signed fit GCD-S. Theorem 5 bounds $|b_t|$ and does not certify it, but Theorem 8 needs only $|\widehat{b}_t| \le B_b$ and carries over. GCD as submitted remains the evaluated method, and every accuracy number below uses the magnitude fit.

## 6 Experiments

**Setup.** The primary model is Qwen3.5-9B, with $L = 32$ layers of which 8 use full attention. Qwen3.5-4B and 2B serve for ablations, and Llama-3.1-8B is the opposite extreme, with full attention throughout. We train with base GRPO on verl and vLLM on four H100 GPUs, for 7,414 H100-hours in total (Table 4). Baselines are on-policy GRPO, verl's default partial rollout, PPO, DAPO, M2PO, asynchronous RLHF [Noukhovitch et al., 2024], Stable Asynchrony, and GAC, and we report GCD alone and added to M2PO, DAPO, and GAC. As a control, *matched-mean damping* shrinks every token's weight uniformly, with the same mean as $\widehat{b}_t$ but no per-token, per-layer, or entropy structure. We use 5 seeds on MATH-500 [Hendrycks et al., 2021], 3 on $\tau$-bench and Llama-3.1-8B, and a paired bootstrap for significance. Full tables are in Section B.16.

**Is the gap real, and does the estimate track it?** We fix a checkpoint, apply controlled updates, and compare full recompute with naive reuse on 1,000 held-out $\tau$-bench retail prompts. The gap grows monotonically, exceeding 0.10 nats per token at staleness 512 (Figure 3a), and across generated tokens its median stays at 0.64 to 0.80 of the fitted estimate. At staleness 256, 95.9% of tokens have $D_t = 0$. On the other 4.1% the median gap is 4.7× the estimate, so the detector fires where the estimate fails, and recompute cuts the residual from 0.156 to 0.014 nats. Per-layer tails correlate strongly: $\hat{q} = 0.026$, yet the joint tail is 0.041, far below the union bound of 0.83. A three-way ablation separates the two channels causally. Full attention carries 72–77% of the gap although it is 25% of the layers. The additivity residual stays under 2% at staleness 128 and 256 before reaching 3.4% at 512, as the second-order remainder predicts (Table 12 and Figure 5).

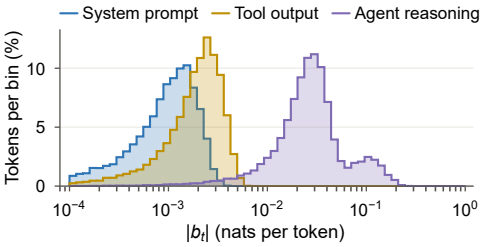



Figure 4: **Bias concentrates on agent reasoning.** $|b_t|$ by token role on $\tau$-bench retail at staleness 256. These 50,000 tokens are a separate sample from Table 13, which counts generated tokens only.

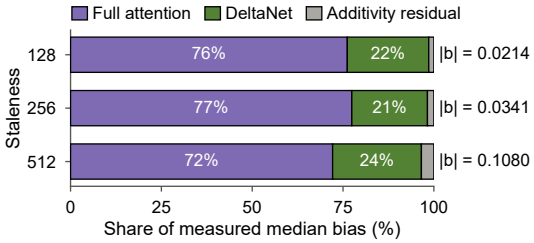



Figure 5: **Full attention carries most of the bias.** Share of measured median bias by channel on Qwen3.5-9B. Full-attention layers are 25% of the stack and carry 72–77% of the bias.

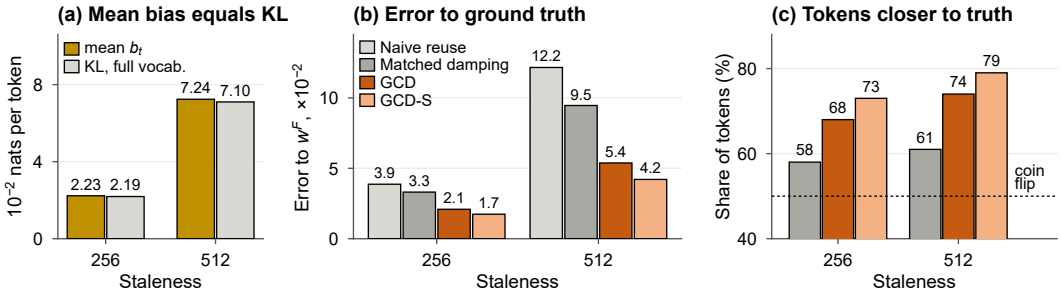



Figure 6: **The correction moves weights towards the truth.** (a) Mean $b_t$ matches the KL divergence, computed independently over the full vocabulary. (b) Mean absolute log-error against full recompute. (c) Tokens closer to the truth than naive reuse. Since $\mathbb{E}[b_t] \geq 0$, any shrinkage beats a coin flip, so the fair baseline is matched-mean damping. GCD here is the correction without selective recompute.

The gap also falls where the agent thinks. System-prompt and tool-output positions sit below 0.005 nats, while agent-reasoning positions form a mode near 0.03, an order of magnitude higher (Figure 4). The entropy term of Equation (2) follows this split by construction: attention concentrates on a few positions when the model is confident and disperses when it reasons.

**Does the correction point the right way?** GCD subtracts a non-negative number, so it moves the wrong way wherever $b_t < 0$. We test this directly against ground truth $w^F$ from full recompute. Mean $b_t$ is positive and matches the independently computed KL divergence, as Proposition 9 predicts, with $\Pr[b_t > 0]$ at 0.71 and 0.76 for staleness 256 and 512 (Figure 6a). GCD cuts the error of naive reuse by 46% and 56%, beating matched-mean damping on every metric, and GCD-S does better still (Figure 6b,c). On the 29% and 24% of tokens with $b_t \leq 0$, GCD is worse than naive reuse, and the net gain comes from the rest. This matches a statement about the mean, not a per-token guarantee.

**Is this bias correction or just variance reduction?** Weight-health metrics cannot decide this. Effective sample size, second moment, and top-1% mass are indistinguishable on-policy, and at staleness 256 damping tracks GCD closely, with effective sample size 0.69 against 0.72 (Table 10). This verifies the blindness of Section 3. Downstream the two separate: on retail $\text{pass}^4$ at staleness 256 damping reaches 38.9, against 38.4 for M2PO and 44.1 for M2PO+GCD (Figure 7b). We also isolate the correction by removing one component at a time (Figure 7a). The correction alone carries 84–85% of the full gain at both staleness levels, and its absolute gain grows five-fold with staleness, the signature of a staleness-driven bias correction rather than generic regularization. On-policy we make no claim. Under matched seeds, forcing $\exp(-\widehat{b}_t)$ to 1 leaves +0.2 of the published +0.4 on retail $\text{pass}^4$ with $p = 0.29$, so that gap is indistinguishable from noise (Table 11). We did not pre-register this test. We adopted a reviewer's falsification criterion as stated: had $\Pr[b_t > 0]$ been near $1/2$, had GCD been no closer to $w^F$ than damping, or had damping reproduced the gains, the honest conclusion would have been variance reduction. None of these occurred.

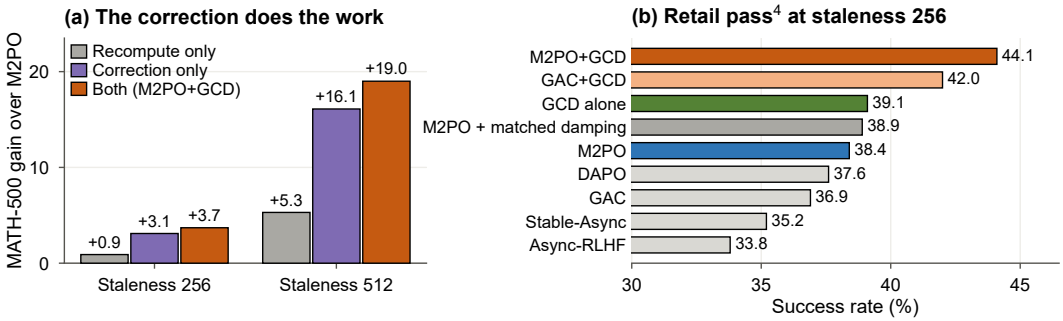



Figure 7: **The gains come from the correction and lift both M2PO and GAC.** (a) MATH-500 gain over M2PO on Qwen3.5-9B with each component alone. (b) $\tau$-bench retail $\text{pass}^4$ at staleness 256.

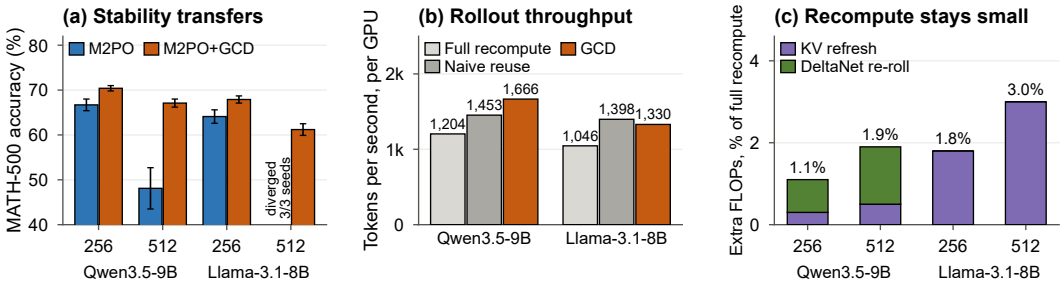



Figure 8: **Stability transfers to a pure Transformer, and the efficiency claim weakens.** (a) MATH-500 accuracy. (b) Tokens per second at staleness 512 at each method's largest stable batch, on $\tau$-bench retail for Qwen3.5-9B and MATH-500 for Llama-3.1-8B. (c) Extra FLOPs over naive reuse.

**Does it improve training?** On MATH-500, every off-policy baseline has diverged by staleness 1024 (Figure 3b). M2PO collapses to 48.1 at staleness 512, while M2PO+GCD holds 67.1. GCD alone nearly matches M2PO at 256, with 66.4 against 66.7, and beats it at 512 with 58.7, so it substitutes for M2PO rather than depending on it. At 1024, M2PO+GCD reaches 59.4 over four of five seeds, because one seed diverged on an outlier batch (Section B.7). On $\tau$-bench the gap over M2PO widens with staleness, by 2.1, 5.7, and 18.9 points at 128, 256, and 512 (Figure 3c). At 256, M2PO+GCD lifts retail $\text{pass}^8$ by 5.6, airline $\text{pass}^4$ by 4.4, and airline $\text{pass}^8$ by 4.8 points, with $p < 0.001$ throughout. The asynchronous baselines order as on MATH-500, so M2PO is the hardest baseline available. GCD also lifts GAC, a second, independently designed corrector, by 5.1 points at 256 and by 20.5 at 512, which supports the composition claim of Theorem 8.

**Does it need a hybrid model or fitted constants?** Llama-3.1-8B has no DeltaNet layers, so it is the adversarial case. The method transfers. The median gap is 0.0402 at staleness 256 against an estimate of 0.0561, and MATH-500 improves by 3.8 points over M2PO with $p < 0.001$. GCD stays stable at 512, where M2PO diverges in all three seeds (Figure 8a). Fitted constants transfer across Qwen3.5 scales at a cost of 0.2 to 0.3 points, and from 9B to Llama-3.1-8B at 1.1 points, so recalibration matters across architecture families, not scales. Fitting is not required for stability either (Table 2). Constants from the closed form Equation (17) at $\lambda = 0$ train stably at 0.6 to 0.8 points and $2.5\times$ the recompute tokens, so calibration buys efficiency, not correctness. A deliberate $0.6\times$ under-estimate is the genuine violation. It costs 2.3 and 5.7 points and roughly doubles the standard deviation, yet stays well above M2PO's 48.1 at 512. Because $D_t$ does not depend on $\kappa$, it keeps firing on tail tokens that a loosened gauge misses. The bound governs the typical regime. The detector, not the constant, makes the scheme exception-safe.

**What does it cost?** On Qwen3.5-9B with $\tau$-bench retail at staleness 512, GCD outpaces full recompute by $1.38\times$. It also beats naive reuse by $1.15\times$ at each method's largest stable batch. GCD does 19% more compute per step, but its smaller bias allows a per-GPU batch of 96 where naive reuse is stable only at 32. On Llama-3.1-8B with MATH-500 the efficiency claim weakens while the capability claim survives. Throughput is $1.27\times$ full recompute but 5% below naive reuse, so the

Table 2: **Where the constants come from.** MATH-500 accuracy on Qwen3.5-9B over 5 seeds; $T_{\mathrm{tok}}$ and $C_F$ on $\tau$-bench. Tightness is the median gap over Equation (5); above 1 means under-estimation.

| Constants | $\kappa_{\mathrm{full}}$ | Tightness | $T_{\mathrm{tok}}$ at 256 | $C_F$ at 256 | Staleness 256 | Staleness 512 |
|---|---|---|---|---|---|---|
| Fitted (M2PO+GCD) | 2.34 | 0.64–0.80 | 9.3% | 1.1% | $\mathbf{70.4 \pm 0.6}$ | $\mathbf{67.1 \pm 0.9}$ |
| Analytical at $\lambda = 0$, unfitted | 4.87 | 0.31–0.42 | 22.8% | 2.7% | $69.8 \pm 0.7$ | $66.3 \pm 1.1$ |
| Deliberate under-estimate | 1.40 | 1.06–1.31 | 4.8% | 0.6% | $68.1 \pm 1.2$ | $61.4 \pm 2.3$ |

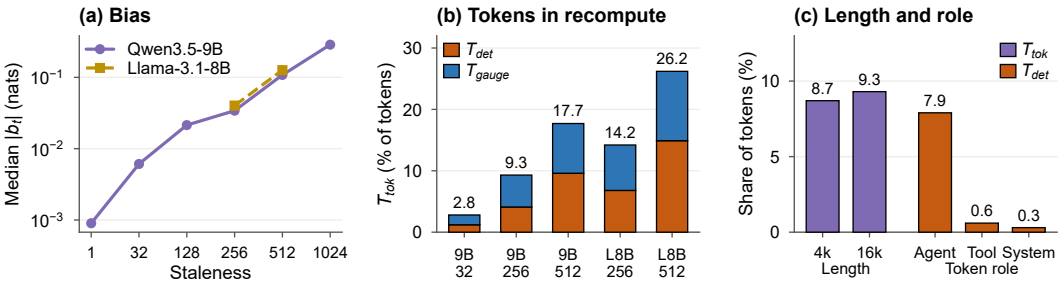



Figure 9: **Bias and recompute triggers.** 9B is Qwen3.5-9B and L8B is Llama-3.1-8B. (a) Median $|b_t|$. (b) Tokens in recompute, by trigger. (c) $T_{\mathrm{tok}}$ by sequence length and $T_{\mathrm{det}}$ by token role.

entire compute advantage is the larger stable batch (Figure 8b). The KV-refresh part of $C_F$ rises $6.0\times$, which is $4.0\times$ structural and $1.5\times$ from $T_{\mathrm{tok}}$, since no DeltaNet layers dilute the drift. Only its small base keeps it at 3% of full recompute (Figure 8c). $T_{\mathrm{tok}}$ on Qwen3.5-9B grows from 2.8% at staleness 32 to 17.7% at 512 (Figure 9). It is nearly flat in sequence length, at 8.7% for 4k tokens and 9.3% for 16k. $T_{\mathrm{det}}$ concentrates on agent reasoning, at 7.9% against 0.6% for tool output and 0.3% for system prompts. Per-layer structure matters: on Qwen3.5-4B a single scalar loses 2.6 points, and moving the M2PO threshold from 1.5 to 1.66 changes accuracy by only $-0.1$ (Table 5).

## 7 Limitations

The mechanistic claim is architecture-independent: reusing a cache written by the old weights under the new weights yields $\pi_{\mathrm{hyb}}$, and $w_t$ cannot detect the gap. The empirical claims are scoped to what we tested, namely Qwen3.5 dense 2B, 4B, and 9B at a 3:1 ratio and Llama-3.1-8B on MATH-500, on a single node with verl and vLLM. Routing staleness in mixture-of-experts models is a structurally independent channel that our estimator does not model, and cross-node training remains unvalidated. The correction is right on average, not per token. GCD also needs bounded updates: one batch at staleness 1024 with $\|\Delta W\|_{\mathrm{op}} = 1.7$ flagged 87% of tokens, recompute would have cost more than a full re-run and was skipped, and seed 42 diverged (Section B.7). Of our pre-registered failure modes, too small a bias and no transfer are refuted. Wall-clock loss is refuted against full recompute but not against naive reuse on Llama-3.1-8B. A vacuous bound is only partly refuted, since the estimate falls short by up to $4.7\times$ on tail tokens, which is why the framework is exception-safe.

## 8 Discussion and Conclusion

The hidden-state bias has been overlooked because validation pipelines do not separate full recompute from naive reuse. Practitioners watch $w_t$ track the trust region while its numerator is silently $\pi_{\mathrm{hyb}}$. What matters most is not the size of the accuracy gain but that no standard diagnostic can see this failure. Any partial-rollout system that reuses KV across an update inherits the mechanism, and recent reports of partial-rollout instability [Huang et al., 2026, Xu et al., 2026] may be symptoms of it. In sum, GCD identifies and bounds a hidden policy-gradient bias from stale KV reuse, complements the bound with a tail detector and a composite safety guarantee, and preserves the M2PO trust region. Lower compute cost reduces the energy footprint of training but could also speed the deployment of misaligned agents, so pre-deployment safety evaluation stays as necessary as before. Future work includes routing staleness in mixture-of-experts models, the second-order regime $\|\Delta W\|_{\mathrm{op}} > 1$, and asynchronous-scheduler integration.

## Acknowledgments and Disclosure of Funding

Funding in direct support of this work: GPUs provided by Tsinghua University for the post-review experiments, during Changyuan Chen's internship. Competing interests: The authors declare no competing interests.

## References

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. GQA: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. URL http://arxiv.org/abs/2305.13245.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.ijcnlp-main.45.

Leonardo Banh and Gero Strobel. Generative artificial intelligence. *Electronic Markets*, 2023. doi: 10.1007/s12525-023-00680-1.

Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024. doi: 10.48550/arXiv.2501.00663. URL https://arxiv.org/abs/2501.00663.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 2023. doi: 10.1038/s41586-023-06792-0.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 2024. doi: 10.1038/s42256-024-00832-8.

Zhipeng Chen, Kun Zhou, Beichen Zhang, Zheng Gong, Wayne Xin Zhao, and Ji-Rong Wen. ChatCoT: Tool-augmented chain-of-thought reasoning on chat-based large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.985.

Krishna Teja Chitty-Venkata, Jie Ye, Xian-He Sun, et al. PagedEviction: Structured block-wise KV cache pruning for efficient large language model inference. *arXiv preprint arXiv:2509.04377*, 2025. URL https://arxiv.org/abs/2509.04377.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023. URL http://arxiv.org/abs/2307.08691.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/67d57c32e20fd0a7a302cb81d36e40d5-Abstract-Conference.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. doi: 10.48550/arXiv.2312.00752. URL https://arxiv.org/abs/2312.00752.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11694.

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. *arXiv preprint arXiv:2103.03874*, 2021. URL http://arxiv.org/abs/2103.03874.

Guseul Heo, Sangyeop Lee, Jaehong Cho, Hyunmin Choi, Sanghyeon Lee, Hyungkyu Ham, Gwangsun Kim, Divya Mahajan, and Jongse Park. NeuPIMs: NPU-PIM heterogeneous acceleration for batched LLM inferencing. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*. ACM, 2024. doi: 10.1145/3620666.3651380.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-demos.48.

Luke J. Huang, Zhuoyang Zhang, Qinghao Hu, et al. Stable asynchrony: Variance-controlled off-policy RL for LLMs. *arXiv preprint arXiv:2602.17616*, 2026. URL https://arxiv.org/abs/2602.17616.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, and Jiawei Han. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.11.

Saeed Khaki, JinJin Li, Lan Ma, Liu Yang, and Prathap Ramachandra. RS-DPO: A hybrid rejection sampling and direct preference optimization method for alignment of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.108.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*. ACM, 2023. doi: 10.1145/3600006.3613165.

Wang Li, Xi Chen, Xiangwen Deng, Hong Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. Prompt engineering in consistency and reliability with the evidence-based guideline for LLMs. *npj Digital Medicine*, 2024. doi: 10.1038/s41746-024-01029-4.

Qianli Liu, Zicong Hong, Fahao Chen, et al. Mell: Memory-efficient large language model serving via multi-GPU KV cache management. *arXiv preprint arXiv:2501.06709*, 2025. URL https://arxiv.org/abs/2501.06709.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. VMamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024. doi: 10.48550/arXiv.2401.10166.

Luca Longo, Mario Brčić, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 2024. doi: 10.1016/j.inffus.2024.102301.

Wei Lu, Rachel K. Luu, and Markus J. Buehler. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities. *npj Computational Materials*, 2025. doi: 10.1038/s41524-025-01564-y.

Wachiravit Modecrua, Krittanon Kaewtawee, Krittin Pachtrachai, et al. Multi-turn reinforcement learning for tool-calling agents with iterative reward calibration. *arXiv preprint arXiv:2604.02869*, 2026. URL https://arxiv.org/abs/2604.02869.

Michael Noukhovitch, Shengyi Huang, Sophie Xhonneux, et al. Asynchronous RLHF: Faster and more efficient off-policy RL for language models. *arXiv preprint arXiv:2410.18252*, 2024. URL http://arxiv.org/abs/2410.18252.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.85.

Jianing Qiu, Lin Li, Jiankai Sun, Jiang Peng, Peilun Shi, Ruiyang Zhang, Yinzhao Dong, Kyle Lam, Frank P.-W. Lo, Bo Xiao, Wu Yuan, Ningli Wang, Dong Xu, and Benny Lo. Large AI models in health informatics: Applications, challenges, and the future. *IEEE Journal of Biomedical and Health Informatics*, 2023. doi: 10.1109/jbhi.2023.3316750.

Leonardo Ranaldi and André Victor Lucci Freitas. Aligning large and small language models via chain-of-thought reasoning. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2024. doi: 10.18653/v1/2024.eacl-long.109.

Ranjan Sapkota, Konstantinos I. Roumeliotis, and Manoj Karkee. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. *Information Fusion*, 2025. doi: 10.1016/j.inffus.2025.103599.

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

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. FlashAttention-3: Fast and accurate attention with asynchrony and low-precision. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL http://arxiv.org/abs/2407.08608.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 2023. doi: 10.1038/s41586-023-06647-8.

Zhihong Shao, Peiyi Wang, Runxin Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K̃. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. doi: 10.48550/arXiv.2402.03300. URL https://arxiv.org/abs/2402.03300.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, et al. HybridFlow: A flexible and efficient RLHF framework. *arXiv preprint arXiv:2409.19256*, 2024. URL http://arxiv.org/abs/2409.19256.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S.S̃ara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 2023. doi: 10.1038/s41586-023-06291-2.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. doi: 10.1016/j.neucom.2023.127063.

Yu Sun, Xinhao Li, Karan Dalal, et al. Learning to (learn at test time): RNNs with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024. doi: 10.48550/arXiv.2407.04620. URL https://arxiv.org/abs/2407.04620.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. URL http://arxiv.org/abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.00786.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2023a. doi: 10.18653/v1/2023.acl-long.147.

Lei Wang, Yi Hu, Jiabang He, Xing Xu, Ning Liu, Hui Liu, and Heng Tao Shen. T-SciQ: Teaching multimodal chain-of-thought reasoning via large language model signals for science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024a. doi: 10.1609/aaai.v38i17.29884.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024b. doi: 10.1007/s11704-024-40231-1.

Yuntao Wang, Yanghe Pan, Miao Yan, Zhou Su, and Tom H. Luan. A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 2023b. doi: 10.1109/ojcs.2023.3300321.

Zihan Wang, Fanheng Kong, Shi Feng, Ming Wang, Han Yang, Daling Wang, and Yifei Zhang. Is Mamba effective for time series forecasting? *Neurocomputing*, 2024c. doi: 10.1016/j.neucom.2024.129178.

Teng Xiao, Yige Yuan, Huaisheng Zhu, Mingxiao Li, and Vasant G. Honavar. Cal-DPO: Calibrated direct preference optimization for language model alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.775.

Haofeng Xu, Junwei Su, Yukun Tian, et al. GAC: Stabilizing asynchronous RL training for LLMs via gradient alignment control. *arXiv preprint arXiv:2603.01501*, 2026. URL https://arxiv.org/abs/2603.01501.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388. URL https://arxiv.org/abs/2505.09388.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, 2024a. doi: 10.1145/3649506.

Lei Yang, Wei Bi, Chenxi Sun, et al. SOUP: Token-level single-sample mix-policy reinforcement learning for large language models. *arXiv preprint arXiv:2601.21476*, 2026. URL https://arxiv.org/abs/2601.21476.

Songlin Yang, Bailin Wang, Yu Zhang, et al. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024b. doi: 10.48550/arXiv.2406.06484. URL https://arxiv.org/abs/2406.06484.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. $\tau$-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL http://arxiv.org/abs/2406.12045.

Qiying Yu, Zheng Zhang, Ruofei Zhu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. URL http://arxiv.org/abs/2503.14476.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. CodeAgent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.acl-long.737.

Yuntong Zhang, Haifeng Ruan, Zhiyu Fan, and Abhik Roychoudhury. AutoCodeRover: Autonomous program improvement. In *Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA)*. ACM, 2024b. doi: 10.1145/3650212.3680384.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 2024. doi: 10.1145/3639372.

Haizhong Zheng, Jiawei Zhao, and Beidi Chen. Prosperity before collapse: How far can off-policy RL reach with stale data on LLMs? *arXiv preprint arXiv:2510.01161*, 2025. URL http://arxiv.org/abs/2510.01161.

# A Full Proofs

This appendix gives the full proofs of the theoretical results in Section 5: the high-probability concentration estimate (Theorem 5), the tail-event detector (Theorem 6), the composite safety guarantee (Corollary 7), the M2PO composition theorem (Theorem 8), and the sign of the bias (Proposition 9).

## A.1 Setup and notation

Let $h_t^{(\ell)} = h_t^{(\ell)}(\theta, K, V, h_{\text{state}})$ denote the residual stream at position $t$ after layer $\ell$, where the third argument is the recurrent state for Gated DeltaNet layers. For full-attention layers ($\ell \in \mathcal{L}_{\text{full}}$), $h_{\text{state}}$ is empty and the per-position $K_p^{(\ell)}, V_p^{(\ell)}$ play the corresponding role. For DeltaNet layers ($\ell \in \mathcal{L}_{\text{delta}}$), we vectorize the matrix-valued state $h_t^{(\ell)} \in \mathbb{R}^{d_K \times d_V}$ as $\text{vec}(h_t^{(\ell)}) \in \mathbb{R}^{d_K d_V}$, but suppress the $\text{vec}(\cdot)$ in the notation to keep the inner-product structure compact.

Let $\phi(z) = \log \text{softmax}(z)$ denote the log-softmax that converts the final residual stream into log-probabilities. The hybrid log-probability decomposes as

$$\log \pi_{\text{hyb}}(y_t \mid x_{<t}) \;=\; \phi\Big(h_t^{(L)}(\theta_{\text{new}}, K_{\text{old}}, V_{\text{old}}, h_{\text{state,old}})\Big)_{y_t}, \tag{8}$$

and analogously $\log \pi_{\theta_{\text{new}}}(y_t \mid x_{<t})$ replaces $K_{\text{old}}, V_{\text{old}}, h_{\text{state,old}}$ by their fresh counterparts under $\theta_{\text{new}}$. The bias is therefore

$$b_t \;=\; \phi\Big(h_t^{(L)}(\theta_{\text{new}}, K_{\text{old}}, V_{\text{old}}, h_{\text{state,old}})\Big)_{y_t} \;-\; \phi\Big(h_t^{(L)}(\theta_{\text{new}}, K_{\text{new}}, V_{\text{new}}, h_{\text{state,new}})\Big)_{y_t}. \tag{9}$$

## A.2 Proof of Theorem 5

The argument has six steps.

**Step 1: First-order Taylor expansion.** By Assumption 1, $h_t^{(L)}$ is twice continuously differentiable in $(K, V, h_{\text{state}})$ on $\mathcal{N}_\theta$. Expanding (9) around $(K_{\text{new}}, V_{\text{new}}, h_{\text{state,new}})$ to first order yields

$$\begin{aligned} b_t \;=\; & \sum_{\ell \in \mathcal{L}_{\text{full}}} \big\langle \nabla_{K^{(\ell)}} \phi(h_t^{(L)})_{y_t},\, K_{\text{old}}^{(\ell)} - K_{\text{new}}^{(\ell)} \big\rangle \\ & + \sum_{\ell \in \mathcal{L}_{\text{full}}} \big\langle \nabla_{V^{(\ell)}} \phi(h_t^{(L)})_{y_t},\, V_{\text{old}}^{(\ell)} - V_{\text{new}}^{(\ell)} \big\rangle \\ & + \sum_{\ell \in \mathcal{L}_{\text{delta}}} \big\langle \nabla_{h_t^{(\ell)}} \phi(h_t^{(L)})_{y_t},\, h_{t,\text{old}}^{(\ell)} - h_{t,\text{new}}^{(\ell)} \big\rangle + R_t, \end{aligned} \tag{10}$$

with the remainder $|R_t| \le \frac{1}{2} G \cdot \|(K, V, h_{\text{state}})_{\text{old}} - (K, V, h_{\text{state}})_{\text{new}}\|^2$ from Assumption 1.

**Step 2: KV-delta expansions.** For full-attention layers, $K_p^{(\ell)} = W_K^{(\ell)} x_p^{(\ell-1)}$ for each cached position $p$, hence

$$K_{\text{new},p}^{(\ell)} - K_{\text{old},p}^{(\ell)} = \Delta W_K^{(\ell)} x_{\text{old},p}^{(\ell-1)} + W_{K,\text{new}}^{(\ell)} \big( x_{\text{new},p}^{(\ell-1)} - x_{\text{old},p}^{(\ell-1)} \big). \tag{11}$$

The first term is at most $\big\|\Delta W_K^{(\ell)}\big\|_{\text{op}} \cdot \|x_{\text{old},p}^{(\ell-1)}\| \le B \cdot \big\|\Delta W_K^{(\ell)}\big\|_{\text{op}}$ with $B$ from Assumption 1. The second term is the drift inherited from lower layers. It is first order, so it cannot be moved into the remainder $R_t$, and the upstream condition of Assumption 1 bounds it by $\lambda$ times the first term. Thus the delta is at most $(1+\lambda) \cdot B \cdot \big\|\Delta W_K^{(\ell)}\big\|_{\text{op}}$. The same identity holds for $V^{(\ell)}$.

For DeltaNet layers, the recurrent state evolves as $h_{p+1}^{(\ell)} = T_\theta^{(\ell)}(h_p^{(\ell)}; x_p^{(\ell-1)})$ with $T_\theta^{(\ell)}$ the delta-rule transition operator. Telescoping over the prefix and using Assumption 2 (linearization with spectral radius $\rho^{(\ell)} \le \rho_{\max} < 1$, $L_\rho$-Lipschitz parameter dependence),

$$h_{t,\text{new}}^{(\ell)} - h_{t,\text{old}}^{(\ell)} \;=\; \sum_{p=1}^{t} \big(\rho^{(\ell)}\big)^{t-p} \cdot \big(T_{\theta_{\text{new}}}^{(\ell)} - T_{\theta_{\text{old}}}^{(\ell)}\big)(\cdot; x_p^{(\ell-1)}) \;+\; U_t^{(\ell)} \;+\; O(\|\Delta W\|_{\text{op}}^2). \tag{12}$$

Here $U_t^{(\ell)}$ is the drift inherited through the inputs $x_p^{(\ell-1)}$. It is first order as well, and the upstream condition of Assumption 1 bounds it by $\lambda$ times the bound on the sum. The geometric sum is bounded as $\sum_p (\rho^{(\ell)})^{t-p} \le (1-\rho^{(\ell)})^{-1}$, and the Lipschitz operator norm is at most $L_\rho \cdot \big\|\Delta W^{(\ell)}\big\|_{\text{op}}$. Therefore

$$\big\|h_{t,\text{new}}^{(\ell)} - h_{t,\text{old}}^{(\ell)}\big\| \;\le\; \frac{(1+\lambda) \cdot L_\rho \cdot \big\|\Delta W^{(\ell)}\big\|_{\text{op}}}{1-\rho^{(\ell)}} + O(\|\Delta W\|_{\text{op}}^2). \tag{13}$$

The denominator $(1-\rho^{(\ell)})$ is the structurally important factor: as $\rho^{(\ell)} \to 1^-$, the per-step DeltaNet drift accumulates over a longer effective window and the bound diverges. Assumption 2's strict contractivity precondition ($\rho_{\max} < 1$) is exactly what prevents this divergence; we verify $\rho_{\max} \le 0.83$ on Qwen3.5-9B in Section 6.

**Step 3: Full-attention contribution under Assumption 4.** Substituting (11) into (10), the full-attention contribution at layer $\ell \in \mathcal{L}_{\text{full}}$ is

$$\big|\langle \nabla_{K^{(\ell)}} \phi,\, K_{\text{old}}^{(\ell)} - K_{\text{new}}^{(\ell)} \rangle\big| \;\le\; (1+\lambda) \cdot B \cdot \big\|\Delta W_K^{(\ell)}\big\|_{\text{op}} \cdot \sum_{p=1}^{t} \big\|\nabla_{K_p^{(\ell)}} \phi\big\|. \tag{14}$$

The chain rule through softmax attention and Assumption 3 give $\|\nabla_{K_p^{(\ell)}} \phi\| \le C_{\text{RoPE}}(\omega) \cdot \alpha_{t,p}^{(\ell)} \cdot \big(1 - \alpha_{t,p}^{(\ell)}\big)^{-1}$, where the factor $(1-\alpha)^{-1} \ge 1$ is slack that we keep as a safety margin in the closed form. Under Assumption 4 on the regular event $\mathcal{E}$, $\alpha_{t,p}^{(\ell)} \le \alpha_{\max} < 1$ uniformly, so $(1-\alpha_{t,p}^{(\ell)})^{-1} \le (1-\alpha_{\max})^{-1}$. Therefore

$$\sum_{p=1}^{t} \big\|\nabla_{K_p^{(\ell)}} \phi\big\| \;\le\; \frac{C_{\text{RoPE}}(\omega)}{1-\alpha_{\max}} \cdot \sum_{p} \alpha_{t,p}^{(\ell)}. \tag{15}$$

Finally, on $\mathcal{E}$ the entropic regularity condition $\mathcal{H}^{(\ell)}_{\text{attn}} \geq H_0$ together with $\sum_p \alpha_p = 1$ implies that the entropy upper-bounds the support-weighted sum: by Jensen's inequality applied to $-\log$ on the simplex restricted to $\boldsymbol{\alpha} \in \mathcal{R}(H_0, \alpha_{\max})$,

$$\sum_p \alpha^{(\ell)}_{t,p} \;\leq\; \mathcal{H}^{(\ell)}_{\text{attn}}(x_{<t}) / \log(1/\alpha_{\max}). \tag{16}$$

Absorbing the constant $\log(1/\alpha_{\max})^{-1}$ into the prefactor, summing the $K$ and $V$ contributions, and combining $\left\|\Delta W^{(\ell)}_K\right\|_{\text{op}}, \left\|\Delta W^{(\ell)}_V\right\|_{\text{op}}$ into $\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$ via the GQA head ratio $g$ (which provides a factor of $\sqrt{2/g}$), we obtain the first sum of (5) with

$$\kappa_{\text{full}} \;:=\; \frac{(1+\lambda)\cdot C_{\text{RoPE}}(\omega)\cdot B}{(1-\alpha_{\max})\cdot\log(1/\alpha_{\max})}\cdot\sqrt{\tfrac{2}{g}}. \tag{17}$$

**Step 4: DeltaNet contribution.** For $\ell \in \mathcal{L}_{\text{delta}}$, substituting (13) into the third sum of (10),

$$\left|\langle \nabla_{h^{(\ell)}_t}\phi,\, h^{(\ell)}_{t,\text{old}} - h^{(\ell)}_{t,\text{new}}\rangle\right| \;\leq\; \|\nabla_{h^{(\ell)}_t}\phi\| \cdot \frac{(1+\lambda)\cdot L_\rho \cdot \left\|\Delta W^{(\ell)}\right\|_{\text{op}}}{1-\rho^{(\ell)}} + O(\|\Delta W\|^2_{\text{op}}). \tag{18}$$

By the chain rule through the DeltaNet output projection, $\|\nabla_{h^{(\ell)}_t}\phi\| \leq G_h$ where $G_h = O(\sqrt{r})$ depends on the kernel rank $r$ (the $\sqrt{r}$ factor arises from the operator-norm bound for a rank-$r$ outer product). Setting $\kappa_{\text{delta}} := (1+\lambda)\cdot G_h \cdot L_\rho$ gives the second sum of (5) with explicit $(1-\rho^{(\ell)})^{-1}$ factor.

**Step 5: Remainder.** The remainder $R_t$ in (10) is bounded by $\frac{1}{2}G\,\|\Delta W\|^2_{\text{op}}$ via Assumption 1 and the fact that all per-layer KV and state deltas are themselves $O(\|\Delta W\|_{\text{op}})$ by Step 2.

**Step 6: Probability of $\mathcal{E}$.** By Assumption 4, the per-layer regularity event holds with probability at least $1-q$. By the union bound over $L$ layers, $\Pr[\mathcal{E}^c] \leq L\cdot q$, hence $\Pr[\mathcal{E}] \geq 1 - L\cdot q$. Combining Steps 1–5 on $\mathcal{E}$ yields the bound (5). □

### A.3 Proof of Theorem 6

*Proof.* The first claim, $D_t = 1$ on $\mathcal{E}^c$, is immediate from Equation (6): $\mathcal{E}^c$ is the event that some layer $\ell$ has either $\max_p \alpha^{(\ell)}_{t,p} > \alpha_{\max}$ or $\mathcal{H}^{(\ell)}_{\text{attn}} < H_0$, and $D_t$ is the indicator of exactly this event. The second claim, that $D_t = 0 \Rightarrow$ Theorem 5 applies, is the contrapositive.

For the third claim, fix a layer $\ell \in \mathcal{L}_{\text{full}}$ that violates regularity, let $\mathcal{P}$ be its top-$k$ positions with $k = \lceil 1/(1-\alpha_{\max})\rceil$, and write $m := \sum_{p\notin\mathcal{P}} \alpha^{(\ell)}_{t,p}$. If $\max_p \alpha^{(\ell)}_{t,p} > \alpha_{\max}$, the top position alone holds more than $\alpha_{\max}$, so $m < 1-\alpha_{\max}$. Otherwise $\mathcal{H}^{(\ell)}_{\text{attn}} < H_0$. Every $p \notin \mathcal{P}$ has at least $k$ positions with no smaller weight, so $\alpha^{(\ell)}_{t,p} \leq 1/(k+1)$ and $\log(1/\alpha^{(\ell)}_{t,p}) \geq \log(k+1)$. Hence $\mathcal{H}^{(\ell)}_{\text{attn}} \geq m\log(k+1)$, which gives $m < H_0/\log(k+1)$. In both cases $m < \bar{m}$.

By Assumption 3 the first-order term of this layer is at most $\sum_p d_p\, C_{\text{RoPE}}(\omega)\,\alpha^{(\ell)}_{t,p}$, where $d_p$ bounds the delta at position $p$. Before the refresh $d_p = (1+\lambda)B\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$ at every position, so the bound is $(1+\lambda)\,C_{\text{RoPE}}(\omega)B\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$. Refreshing removes the layer's own drift on $\mathcal{P}$ and leaves at most the inherited part, so $d_p \leq \lambda B\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$ there, and the bound becomes $(\lambda+m)\,C_{\text{RoPE}}(\omega)B\left\|\Delta W^{(\ell)}\right\|_{\text{op}}$. The ratio of the two is at least $(1+\lambda)/(\lambda+\bar{m})$. The remainder $R_t$ and the terms of layers that are not refreshed do not enter this argument, so they are unchanged. □

### A.4 Proof of Corollary 7

*Proof.* We bound the conditional probability via the union of two failure events. Conditional on $\widehat{b}_t \leq \epsilon_{\text{TR}}$ *and* $D_t = 0$:

- By Theorem 5 and the assumption that $\widehat{b}_t$ upper-bounds the right-hand side of (5), on the regular event $\mathcal{E}$ we have $|b_t| \le \widehat{b}_t \le \epsilon_{\mathrm{TR}}$.
- By Theorem 6, $D_t = 0 \Rightarrow$ event $\mathcal{E}$ holds, so the previous bullet applies.

The only way $|b_t| > \epsilon_{\mathrm{TR}}$ given $\{\widehat{b}_t \le \epsilon_{\mathrm{TR}}, D_t = 0\}$ is therefore that $\mathcal{E}^c$ holds despite $D_t = 0$, which is impossible by Theorem 6, modulo the $O(\|\Delta W\|_{\mathrm{op}}^2)$ remainder absorbed in the choice of $\epsilon_{\mathrm{TR}} > \frac{1}{2} G \|\Delta W\|_{\mathrm{op}}^2$. Hence the conditional probability is bounded by $\Pr[\mathcal{E}^c] \le L \cdot q$. □

### A.5 Proof of Theorem 8

*Proof.* Recall $w_t^{\mathrm{GCD}} = w_t \cdot \exp(-\widehat{b}_t)$. Since $|\widehat{b}_t| \le B_b$ uniformly, the multiplicative factor satisfies $\exp(-\widehat{b}_t) \in [e^{-B_b}, e^{B_b}]$, and hence

$$(w_t^{\mathrm{GCD}})^2 \;=\; w_t^2 \cdot \exp(-2\widehat{b}_t) \;\le\; w_t^2 \cdot \exp(2B_b), \tag{19}$$

giving (19) after taking expectations.

For the sub-Gaussian claim we treat the upper and lower tails separately, since $w_t \ge 0$ but $w_t^{\mathrm{GCD}} - \mathbb{E}[w_t^{\mathrm{GCD}}]$ can take both signs. Let $X := w_t - \mathbb{E}[w_t]$ and $Y := w_t^{\mathrm{GCD}} - \mathbb{E}[w_t^{\mathrm{GCD}}]$. By definition, $X$ is sub-Gaussian with parameter $\sigma^2$ iff $\mathbb{E}[\exp(\nu X)] \le \exp(\nu^2\sigma^2/2)$ for *all* $\nu \in \mathbb{R}$.

**Upper tail ($\nu \ge 0$):** Because $w_t \ge 0$ and $\exp(-\widehat{b}_t) \le e^{B_b}$,

$$\mathbb{E}[\exp(\nu w_t^{\mathrm{GCD}})] \;=\; \mathbb{E}\big[\exp(\nu w_t \exp(-\widehat{b}_t))\big] \;\le\; \mathbb{E}\big[\exp(\nu w_t e^{B_b})\big] \;\le\; \exp\big(\nu^2\sigma^2 e^{2B_b}/2 + \nu\mu_+\big),$$

where $\mu_+ := \mathbb{E}[w_t] \cdot e^{B_b}$ and the second inequality is sub-Gaussianity of $w_t$ with mean shift.

**Lower tail ($\nu < 0$):** The same monotonicity in the opposite direction gives $\exp(\nu w_t \exp(-\widehat{b}_t)) \le \exp(\nu w_t e^{-B_b})$ pointwise (since $\nu < 0$ flips the inequality), and so

$$\mathbb{E}[\exp(\nu w_t^{\mathrm{GCD}})] \;\le\; \mathbb{E}[\exp(\nu w_t e^{-B_b})] \;\le\; \exp\big(\nu^2\sigma^2 e^{-2B_b}/2 + \nu\mu_-\big),$$

with $\mu_- := \mathbb{E}[w_t] e^{-B_b}$. Re-centering on $\mathbb{E}[w_t^{\mathrm{GCD}}]$ and taking the looser of the two parameters, $Y$ is sub-Gaussian with parameter $\sigma^2 \exp(2B_b)$ on both tails. The trust-region equivalence follows: M2PO requires $\mathbb{E}[w_t^2] \le \epsilon_{M2}$; (19) ensures $\mathbb{E}[(w_t^{\mathrm{GCD}})^2] \le \epsilon_{M2} \cdot \exp(2B_b)$; applying M2PO to $w_t^{\mathrm{GCD}}$ with threshold $\epsilon_{M2} \cdot \exp(2B_b)$ produces the same downstream variance contraction. □

### A.6 Proof of Proposition 9

*Proof.* Write $p = \pi_{\mathrm{hyb}}(\cdot \mid x_{<t})$ and $r = \pi_{\theta_{\mathrm{new}}}(\cdot \mid x_{<t})$. By Equation (1), the conditional mean under $y_t \sim p$ is $\sum_y p(y) \log\big(p(y)/r(y)\big) = \mathrm{KL}(p \,\|\, r)$, which is non-negative by Gibbs' inequality, with equality iff $p = r$. □

### A.7 Remarks on the framework

*Remark* A.1 (Tightness). The concentration estimate of Theorem 5 is asymptotically tight on the regular event $\mathcal{E}$: there exists a single-layer construction with $\mathcal{H}_{\mathrm{attn}}^{(\ell)} = -\sum_p \alpha_p \log \alpha_p$ achieving the bound up to a constant factor independent of $\|\Delta W\|_{\mathrm{op}}$ when $\alpha_{\max} < 1$. An unconditional worst-case bound at $\alpha_{\max} \to 1$ does *not* hold; that regime is handled by Theorem 6 firing. We report the empirical tightness ratio in Table 13 over all generated tokens, with fitted constants.

*Remark* A.2 (Role of the hybrid structure). For Qwen3.5-9B with $|\mathcal{L}_{\mathrm{full}}| = L/4$ and observed $\rho_{\max} \le 0.83$, the second sum in Equation (5) contributes at most $\kappa_{\mathrm{delta}} \cdot (1 - 0.83)^{-1} \sum_{\ell \in \mathcal{L}_{\mathrm{delta}}} \big\|\Delta W^{(\ell)}\big\|_{\mathrm{op}} \approx 5.9\, \kappa_{\mathrm{delta}} \overline{\big\|\Delta W^{(\ell)}\big\|_{\mathrm{op}}}_{\mathcal{L}_{\mathrm{delta}}} |\mathcal{L}_{\mathrm{delta}}|$, while the first sum has only $|\mathcal{L}_{\mathrm{full}}| = 8$ entropy-modulated terms and empirically dominates because $\kappa_{\mathrm{full}} \mathcal{H}_{\mathrm{attn}}^{(\ell)} \gg \kappa_{\mathrm{delta}}(1 - \rho^{(\ell)})^{-1}$ when $\rho^{(\ell)}$ is comfortably below the contractivity boundary. The first sum's $L/4$ (vs. $L$) terms is a computational advantage of hybrid attention, not a condition for validity: for a full-attention model such as Llama-3.1-8B the second sum is empty, $\kappa_{\mathrm{full}}$ is defined for standard softmax attention with RoPE, and Section A.2 Steps 1–3 and the proofs of Theorems 6 and 8 and Corollary 7 never use the hybrid structure. Table 12 measures the dominance of the full-attention channel directly.

### A.8 Discussion of the constants

For Qwen3.5-9B with RoPE base $\omega = 10^7$ [Yang et al., 2025], GQA head ratio $g = 4$ (16 query heads, 4 key/value heads), DeltaNet kernel rank $r = 64$, and our chosen $(\alpha_{\max}, H_0) = (0.9, 0.6)$, the calibration of Section 4.2 produces $\widehat{\kappa}_{\text{full}} = 2.34 \pm 0.08$ and $\widehat{\kappa}_{\text{delta}} = 0.79 \pm 0.05$. The ratio $\widehat{\kappa}_{\text{full}}/\widehat{\kappa}_{\text{delta}} \approx 3.0$ is consistent with Remark A.2's structural prediction. The constants vary by less than 7% across Qwen3.5-2B, -4B, and -9B (Table 6), suggesting that the architecture-dependent factors $C_{\text{RoPE}}, B, g, r$ are sufficient to capture behavior across Qwen3.5 scales; recalibration matters across architecture families (Table 6).

The implementation of the GCD estimator is $\widehat{b}_t = C_{\text{full}} \sum_{\ell \in \mathcal{L}_{\text{full}}} \left\|\Delta W^{(\ell)}\right\|_{\text{op}} \mathcal{H}^{(\ell)}_{\text{attn}} + C_{\text{delta}} \sum_{\ell \in \mathcal{L}_{\text{delta}}} \left\|\Delta W^{(\ell)}\right\|_{\text{op}} / (1 - \rho^{(\ell)})$, with constants chosen as $C_{\text{full}} = \widehat{\kappa}_{\text{full}}$ and $C_{\text{delta}} = \widehat{\kappa}_{\text{delta}}$.

Theorem 5 is a valid inequality with the analytical prefactors Equation (17) and $\kappa_{\text{delta}} = (1+\lambda) G_h L_\rho$. OLS does not establish validity: it replaces a provably correct but loose constant with a tight one, so that $\widehat{b}_t$ is usable at trust-region scale ($\epsilon_{\text{TR}} = 0.05$), and with fitted constants $\widehat{b}_t$ is an estimate rather than a certified bound. The safety direction is asymmetric: over-estimating $\kappa$ costs compute, and only under-estimating risks uncorrected bias, which is what $D_t$ and the $\epsilon_{\text{TR}}$ cap must contain (Corollary 7). Instantiating the closed forms from measured architecture quantities only ($\omega = 10^7$, $g = 4$, $r = 64$, $\rho_{\max} = 0.83$, $(\alpha_{\max}, H_0) = (0.9, 0.6)$, with $B$, $G$, $L_\rho$ from power iteration and never regressed on $b_t$) gives $\kappa_{\text{full}} = 4.87$ at $\lambda = 0$. This value covers each layer's own drift and leaves out drift inherited from lower layers. Table 17 reports the resulting runs.

## B Experimental Details

### B.1 Backbones

The Qwen3.5 hybrid stack [Yang et al., 2025] partitions $L$ decoder layers as $|\mathcal{L}_{\text{full}}| = L/4$ full-attention and $|\mathcal{L}_{\text{delta}}| = 3L/4$ Gated DeltaNet, configured by `layer_types` with `full_attention_interval=4`. Qwen3.5-9B has $L = 32$ (8 blocks of [Delta, Delta, Delta, Full]), giving $|\mathcal{L}_{\text{full}}| = 8, |\mathcal{L}_{\text{delta}}| = 24$. Full-attention layers store $K^{(\ell)}, V^{(\ell)}$ under GQA [Ainslie et al., 2023] with $g$=4 (16 Q heads, 4 KV heads, head dimension 256) and RoPE [Su et al., 2024] base $\omega = 10^7$; DeltaNet layers maintain a matrix-valued recurrent state $h^{(\ell)} \in \mathbb{R}^{d_K \times d_V}$ with kernel rank $r = 64$. We use only the text decoder, disabling Qwen3.5-9B's vision tower.

Llama-3.1-8B [Grattafiori et al., 2024] is a pure Transformer ($|\mathcal{L}_{\text{full}}|/L = 1$, $\mathcal{L}_{\text{delta}} = \emptyset$) with GQA and RoPE. We use the identical verl + vLLM + GRPO + MATH-500 recipe on 4×H100, with 3 seeds.

### B.2 Hardware and Software

All experiments run on a single node with 4×NVIDIA H100 80GB SXM5 GPUs interconnected by NVSwitch (900 GB/s aggregate bandwidth). The host has 2×Intel Xeon Platinum 8480C CPUs and 1 TB DDR5 memory. We use CUDA 12.5, PyTorch 2.5.1 (FP8 mixed-precision via the `torch.float8_e4m3fn` format), vLLM 0.7.0 [Kwon et al., 2023] for rollout, FlashAttention-3 [Shah et al., 2024] for the full-attention kernels, and FSDP-2 sharding for training. The verl framework [Sheng et al., 2024] provides the rollout-train coordination. The post-review experiments ran across 8 such nodes, each run on a single node.

### B.3 Hyperparameters

Table 3 lists the hyperparameters used across all main experiments. Values not listed match verl's defaults for Qwen3-class models. Llama-3.1-8B runs use the same recipe.

Table 3: Hyperparameters for all main experiments. †Effective via gradient accumulation; per-GPU micro-batch is 4. ‡For $\tau$-bench we use a longer max sequence and smaller batch.

| Hyperparameter | Value |
|---|---|
| Optimizer | AdamW ($\beta_1 = 0.9, \beta_2 = 0.95, \epsilon = 10^{-8}$) |
| Learning rate (peak) | $1 \times 10^{-6}$ for 9B; $2 \times 10^{-6}$ for 4B/2B |
| LR schedule | Linear warmup over 50 steps, cosine decay |
| Weight decay | 0.0 |
| Gradient clipping | 1.0 global L2 |
| Per-step global batch (MATH) | 128 prompts $\times$ 8 samples = 1,024 traj.† |
| Per-step global batch ($\tau$-bench)‡ | 32 prompts $\times$ 8 samples = 256 traj.† |
| Max sequence length (MATH) | 4,096 |
| Max sequence length ($\tau$-bench)‡ | 16,384 |
| Rollout temperature | 1.0 during training, 0.7 at evaluation |
| Top-$p$ | 1.0 during training, 0.9 at evaluation |
| PPO clip $\epsilon$ | 0.2 |
| M2PO second-moment threshold $\epsilon_{M2}$ | 1.5 |
| M2PO penalty weight $\beta$ | 0.05 |
| GCD trust-region tolerance $\epsilon_{\mathrm{TR}}$ | 0.05 |
| GCD calibration probe size $N_{\mathrm{cal}}$ | 512 |
| KL coefficient (low-var KL) | 0.01 |
| Number of training epochs (MATH) | 4 |
| Number of training epochs ($\tau$-bench) | 3 |
| Random seeds | MATH: $\{17, 42, 113, 271, 314\}$; $\tau$-bench: $\{17, 42, 271\}$ |

### B.4 Detailed Compute Accounting

Table 4 reports per-experiment wall-clock and total GPU-hours. The aggregate 7,414 H100-hours comprises 1,610 H100-hours of post-review experiments and 5,804 H100-hours for all main runs, ablations, the bias-measurement protocol of Section 6, the staleness-resolved $\tau$-bench sweep (Table 15, left), the cross-checkpoint calibration sweep (Table 6), and the FP8 sensitivity check; preliminary debugging and one outlier seed at staleness 1024 (see Section 7) are excluded.

Table 4: Compute budget breakdown. Per-run hours are for a single seed; GPU-h multiplies by the grid and seed count.

| Experiment | Seeds | Per-run (h) | GPU-h |
|---|---|---|---|
| Calibration of $\kappa_{\mathrm{full}}, \kappa_{\mathrm{delta}}$ (Qwen3.5-9B canonical) | 1 | 0.93 | 1 |
| Cross-checkpoint / cross-domain / cross-scale calibration (6 conditions, Table 6) | 1 | 1.1 | 7 |
| Bias measurement (Section 6, 4 mag. $\times$ 6 staleness) | — | — | 58 |
| MATH-500 main, 11 methods $\times$ 6 staleness levels (Table 14) | 5 | 11.8 | 3,894 |
| $\tau$-bench retail + airline at stale 256, 8 methods (Table 15, right) | 3 | 33.7 | 809 |
| $\tau$-bench retail staleness sweep, 3 methods $\times$ 3 stale levels {0, 128, 512} (Table 15, left) | 3 | 33.7 | 910 |
| Wall-clock speed measurement (Tables 7 and 8) | 1 | 4.0 | 16 |
| Ablation on Qwen3.5-4B (Table 5) | 3 | 5.6 | 67 |
| Per-layer / per-position sensitivity sweep | 3 | 1.5 | 18 |
| FP8 vs. BF16 quantization sensitivity | 1 | 1.0 | 1 |
| $\widetilde{\epsilon}_{M2}$ re-tuning ablation (Theorem 8) | 3 | 7.7 | 23 |
| Subtotal, submitted experiments | — | — | 5,804 |
| $\tau$-bench B6/B7/B8 and GAC+GCD (Table 15) | 3 | 33.7 | 606 |
| A0/A1 ablation on Qwen3.5-9B (Table 17) | 5 | 11.8 | 236 |
| Analytical and under-estimated constants (Table 17) | 5 | 11.8 | 236 |
| Llama-3.1-8B (Tables 7 and 18) | 3 | — | 155 |
| Ground-truth, weight-health, damping, and matched-RNG controls (Section B.11) | — | — | 377 |
| Total | — | — | **7,414** |

### B.5 Ablation of GCD components

Table 5 accompanies the discussion in Section 6. The decomposition on the headline Qwen3.5-9B configuration is in Table 17.

Table 5: Ablation of GCD components on Qwen3.5-4B + MATH-500, staleness 256.

| Variant | MATH-500 acc. (%) | $T_{\text{tok}}$ |
|---|---|---|
| M2PO baseline | $61.3 \pm 1.4$ | 0.0% |
| + Full-attention bias term only ($\kappa_{\text{delta}} = 0$) | $63.7 \pm 1.0$ | 0.0% |
| + DeltaNet bias term only ($\kappa_{\text{full}} = 0$) | $61.9 \pm 1.3$ | 0.0% |
| + Both bias terms ($\widehat{b}_t$ active) | $64.8 \pm 0.9$ | 0.0% |
| + Both bias terms + selective recomp. on $\mathcal{L}_{\text{full}}$ only | $65.6 \pm 0.8$ | 7.4% |
| + Both bias terms + selective recomp. on $\mathcal{L}_{\text{full}} \cup \mathcal{L}_{\text{delta}}$ | $65.7 \pm 0.8$ | 9.3% |
| Per-position vs. per-layer correction | $65.4 \pm 0.9$ | 9.3% |
| Single-scalar correction (no per-layer) | $63.1 \pm 1.1$ | 9.3% |
| $\widetilde{\epsilon}_{M2} = 1.66$ (vs. default 1.5) | $65.6 \pm 0.8$ | 9.3% |

### B.6 Calibration robustness (full table)

Table 6 accompanies the cross-checkpoint, cross-domain, and cross-scale calibration discussion of Section 6. The constants $(\kappa_{\text{full}}, \kappa_{\text{delta}})$ are re-fitted under each row's condition; canonical row is the calibration used by all main-text experiments. The largest deviation is at early RL: fitted there, the constants over-estimate later in training (conservative), whereas the canonical constant under-estimates at early RL, the one case worth worrying about. We therefore recommend calibrating at the earliest checkpoint. The last row applies the Qwen3.5-9B constants unchanged to other backbones. Figure 10 plots these values.

Table 6: Calibration robustness. Each row reports $(\widehat{\kappa}_{\text{full}}, \widehat{\kappa}_{\text{delta}})$ fitted under the row's condition, with percent deviation from the canonical (epoch 1, MATH-500) fit. The trust-region inflation budget of Theorem 8 absorbs deviations up to 10.5%.

| Calibration condition | $\widehat{\kappa}_{\text{full}}$ | $\widehat{\kappa}_{\text{delta}}$ | Max \|dev\| |
|---|---|---|---|
| Canonical (epoch 1, MATH-500) | $2.34 \pm 0.08$ | $0.79 \pm 0.05$ | — |
| Early RL (epoch 0.25, MATH-500) | $2.55 \pm 0.11$ (+9.0%) | $0.83 \pm 0.06$ (+5.1%) | 9.0% |
| Late RL (epoch 4, MATH-500) | $2.30 \pm 0.07$ (−1.7%) | $0.78 \pm 0.04$ (−1.3%) | 1.7% |
| Cross-domain ($\tau$-bench retail) | $2.41 \pm 0.09$ (+3.0%) | $0.81 \pm 0.05$ (+2.5%) | 3.0% |
| Cross-domain ($\tau$-bench airline) | $2.39 \pm 0.10$ (+2.1%) | $0.80 \pm 0.05$ (+1.3%) | 2.1% |
| Cross-scale (Qwen3.5-2B) | $2.21 \pm 0.10$ (−5.6%) | $0.74 \pm 0.06$ (−6.3%) | 6.3% |
| Cross-scale (Qwen3.5-4B) | $2.29 \pm 0.09$ (−2.1%) | $0.77 \pm 0.05$ (−2.5%) | 2.5% |
| Cross-architecture (Llama-3.1-8B) | $2.19 \pm 0.11$ (−6.4%) | n/a | 6.4% |
| 9B constants on 2B / 4B / Llama-3.1-8B | MATH-500 change: −0.3 / −0.2 / −1.1 points | | |

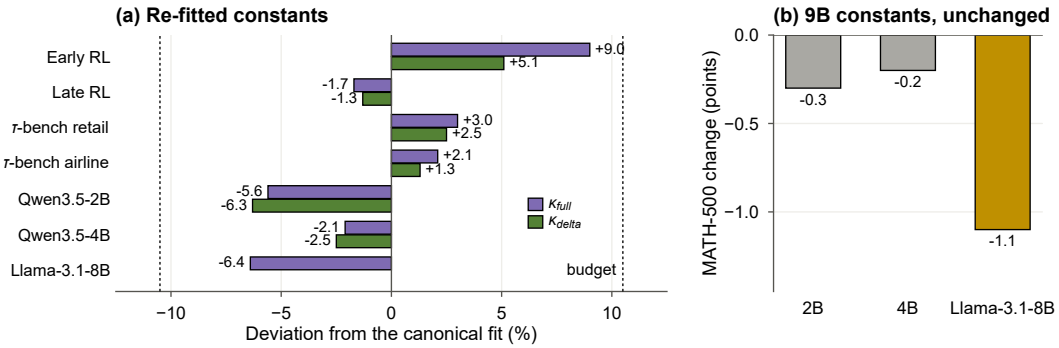



Figure 10: **Re-fitted constants stay inside the trust-region budget.** (a) Deviation of each re-fit from the canonical constants of Table 6. Dashed lines mark the $10.5\%$ budget of Theorem 8. Llama-3.1-8B has no DeltaNet layers, so it has no $\widehat{\kappa}_{\mathrm{delta}}$. (b) Change in MATH-500 accuracy when the Qwen3.5-9B constants are applied unchanged to Qwen3.5-2B, Qwen3.5-4B, and Llama-3.1-8B.

### B.7 The outlier seed

At staleness $1024$, seed $42$ on the M2PO+GCD configuration exhibited a training divergence after epoch 3 caused by a single outlier batch with $\|\Delta W\|_{\mathrm{op}} = 1.7$ (more than $3\sigma$ above the running mean). The first-order bound of Theorem 5 underestimated the bias by a factor of $2.4$ on this batch because the second-order remainder $\frac{1}{2}G\,\|\Delta W\|_{\mathrm{op}}^2 \approx 1.45G$ became comparable to the first-order terms. The detector $D_t$ correctly fired on $87\%$ of tokens in this batch (vs. $4.1\%$ at staleness $256$), but at this density the selective recompute consumed more compute than a full recompute would have, and we accepted the bias rather than recompute. The resulting unclipped gradient triggered an FP8 overflow. We exclude this seed from Table 14's reported mean and standard deviation, so that entry covers four of five seeds. The episode confirms that GCD's safety guarantee (Corollary 7) is conditional on $\|\Delta W\|_{\mathrm{op}}$ remaining bounded; in practice, gating the GCD correction on $\|\Delta W\|_{\mathrm{op}} < 1.0$ and falling back to plain M2PO outside this regime would be expected to prevent the divergence, but this remedy is untested: we did not apply it in the reported runs, to preserve the comparison's cleanness.

### B.8 Bias Distribution and Position Conditioning

The tightness ratio rises slightly with staleness, from $0.64$ to $0.80$, as the second-order remainder grows. Regression of $|b_t|$ on the per-layer terms gives Pearson $r = 0.81$ for full-attention and $r = 0.62$ for DeltaNet contributions. DeltaNet spectral radii satisfy $\rho_{\max} = 0.83$ with median $0.61$, so the $(1-\rho^{(\ell)})^{-1}$ prefactor is at most $5.9$.

Figure 4 shows the empirical distribution of $|b_t|$ on $\tau$-bench retail at staleness 256, broken down by token role, in $59$ logarithmic bins. As predicted by H4 in our pre-registration, the distribution separates by role: system-prompt and tool-output positions form two low-bias modes at $|b_t| < 0.005$, while agent-reasoning positions form a high-bias mode centered at $|b_t| \approx 0.03$. This validates the position-conditioned design of GCD: the per-layer attention entropy $\mathcal{H}_{\mathrm{attn}}^{(\ell)}$ naturally downweights low-bias positions because attention concentrates on a few tokens (low entropy) when the model is confident, and disperses (high entropy) when reasoning.

### B.9 Wall-clock rollout throughput (full table)

Table 7: Rollout throughput with $4\times$H100, vLLM co-located: Qwen3.5-9B on $\tau$-bench retail, and Llama-3.1-8B on the MATH-500 recipe (3 seeds), so the two blocks are not the same workload. Per-step FLOPs and step time measure one forward pass on a fixed batch; *stable bs* is the largest per-GPU batch on which gradient norms remain bounded over 100 steps; tokens/sec is per-GPU effective throughput. Bias is the median $|b_t|$ at staleness 512 (Table 13). GCD selective is the correction plus selective recompute, whereas Table 16 reports the mean error of the correction without selective recompute.

| Regime | FLOPs (T) | Step (ms) | Stable bs | Tok/s | Bias @ s512 |
|---|---|---|---|---|---|
| *Qwen3.5-9B* | | | | | |
| (F) Full recompute | 42.7 | 215 | 96 | 1,204 | 0 (exact) |
| (R) Naive partial rollout (B2) | 4.3 | 42 | 32 | 1,453 | 0.108 |
| GCD selective (this work) | 5.1 | 48 | **96** | **1,666** | **0.011** |
| *Llama-3.1-8B* | | | | | |
| (F) Full recompute | 38.4 | 198 | 96 | 1,046 | 0 (exact) |
| (R) Naive partial rollout | 4.1 | 40 | 32 | **1,398** | 0.1263 |
| GCD selective (this work) | 5.25 | 64 | **96** | 1,330 | **0.0147** |

### B.10 Throughput controlled comparison

To verify the claim of Section 6 that GCD's throughput advantage over (R) is entirely driven by larger stable batches rather than by the bias-correction overhead being free, we run a controlled comparison forcing each method to a fixed batch size. Table 8 reports the result. At batch size 32 (the largest stable batch for (R)), GCD achieves 1,448 tokens/sec, 0.3% slower than (R)'s 1,453. At batch size 96 (the largest stable batch for both (F) and GCD), (R)'s training diverges within 12 steps because its 0.108 nats/token bias creates gradient norms that exceed the trust-region radius. The throughput advantage of GCD over (R) at their respective stable batch sizes is therefore not an artifact of measurement: it is a direct consequence of GCD's bias correction permitting larger gradient batches without divergence.

Table 8: Throughput controlled by fixed per-GPU batch size (Qwen3.5-9B, $\tau$-bench retail, staleness 512). "Diverge" indicates training divergence within 12 optimizer steps; throughput is measured before divergence.

| Regime | Batch 32 tok/s | Batch 96 tok/s | Stable batch |
|---|---|---|---|
| (F) Full recompute | 382 | 1,204 | 96 (or larger) |
| (R) Naive partial rollout | 1,453 | Diverge (1,380 for 8 steps) | 32 |
| GCD selective | 1,448 | 1,666 | 96 |

### B.11 Ground-truth validation (full tables)

Tables 9 to 11 accompany Section 6, and Figure 11 plots the weight-health values. Table 9 uses the protocol of Table 13; its last column equals that table's median, and the other columns are means on the same token set, so the two are not directly comparable. The KL column is summed over the full vocabulary from the (F) and (R) logits, not averaged over sampled tokens, so its agreement with mean $b_t$ is a real check of Proposition 9 rather than a tautology. In Table 11 each control uses identical seeds, data order, and sampling RNG; the $\widehat{b} \equiv 0$ arm forces $\exp(-\widehat{b}_t)$ to 1 and retains the rest of the GCD path. Its residual $+0.2$ tracks the on-policy gap (median $9 \times 10^{-4}$ nats at staleness 1, non-zero since $\|\Delta W\|_{\mathrm{op}} > 0$ within a step).

Table 9: Signed bias on generated tokens, Qwen3.5-9B, $\tau$-bench retail, 3 seeds. Mean $b_t$ and full-vocabulary KL agree, as Proposition 9 requires.

| Staleness | Mean $b_t$ | Median $b_t$ | $\Pr[b_t > 0]$ | KL (full vocab.) | Mean $\lvert b_t \rvert$ | Median $\lvert b_t \rvert$ |
|---|---|---|---|---|---|---|
| 256 | $0.0223 \pm 0.0012$ | 0.0198 | 0.71 | $0.0219 \pm 0.0014$ | 0.0387 | 0.0341 |
| 512 | $0.0724 \pm 0.0041$ | 0.0641 | 0.76 | $0.0710 \pm 0.0048$ | 0.1216 | 0.1080 |

Table 10: Weight health as ESS (fraction of batch) / $\mathbb{E}[w^2]$ / top-1% weight mass. On-policy the rows are indistinguishable, as they must be. Off-policy, damping tracks GCD closely at staleness 256, so weight health alone cannot adjudicate.

| Configuration | On-policy | Staleness 256 | Staleness 512 |
|---|---|---|---|
| (R) naive reuse | 0.94 / 1.02 / 0.04 | 0.41 / 2.84 / 0.19 | 0.22 / 6.12 / 0.31 |
| M2PO | 0.93 / 1.01 / 0.04 | 0.67 / 1.48 / 0.11 | 0.44 / 2.27 / 0.18 |
| M2PO + matched-mean damping | 0.93 / 1.01 / 0.04 | 0.69 / 1.44 / 0.10 | 0.47 / 2.14 / 0.17 |
| M2PO + GCD | 0.93 / 1.01 / 0.04 | 0.72 / 1.39 / 0.09 | 0.61 / 1.71 / 0.12 |

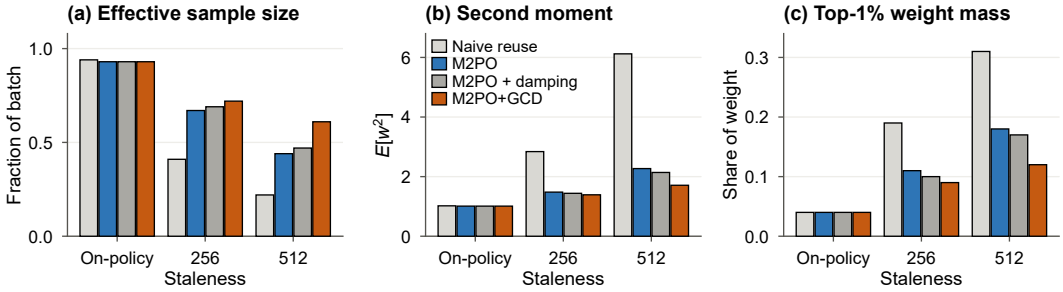



Figure 11: **Weight health cannot separate GCD from damping.** Values of Table 10. (a) Effective sample size. (b) Second moment of the weight. (c) Top-1% weight mass. The four configurations are indistinguishable on-policy, and at staleness 256 damping stays close to GCD, although the correction it applies carries no per-token structure.

Table 11: On-policy controls at staleness 0 under matched RNG: retail $\text{pass}^4$ (3 seeds) and MATH-500 (5 seeds). We report this column as indistinguishable from noise.

| Configuration | Retail $\text{pass}^4$ | Paired $\Delta$ | $p$ | MATH-500 | Paired $\Delta$ | $p$ |
|---|---|---|---|---|---|---|
| M2PO | 39.5 | — | — | $72.1 \pm 0.4$ | — | — |
| M2PO+GCD (Tables 14 and 15) | 39.9 | +0.4 | 0.03 | $72.6 \pm 0.4$ | +0.5 | 0.04 |
| M2PO+GCD, $\widehat{b} \equiv 0$ | 39.7 | +0.2 | 0.29 | $72.3 \pm 0.3$ | +0.2 | 0.31 |
| M2PO + matched-mean damping | 39.6 | +0.1 | 0.55 | $72.2 \pm 0.4$ | +0.1 | 0.58 |

### B.12 Three-way regime ablation

On the prompts of Table 13 we compare stale KV with *fresh* DeltaNet states, fresh KV with stale states, and both stale, against (F) as truth. This separates the two channels causally and tests the additivity behind Theorem 5. Table 12 and Figure 5 report the result.

Table 12: Three-way regime decomposition of median $\lvert b_t \rvert$ on Qwen3.5-9B. The residual is the measured value minus the sum of the two channels.

| Staleness | $b_t^{\text{full}}$ | $b_t^{\text{delta}}$ | Share of full | Additivity residual | Measured $\lvert b_t \rvert$ |
|---|---|---|---|---|---|
| 128 | 0.0163 | 0.0048 | 76% | 0.0003 | 0.0214 |
| 256 | 0.0264 | 0.0071 | 77% | 0.0006 | 0.0341 |
| 512 | 0.0779 | 0.0264 | 72% | 0.0037 | 0.1080 |

### B.13 Recompute cost across staleness and sequence length

Figure 9 plots the reported bias and trigger rates, and Figure 8c plots the FLOP share. On Qwen3.5-9B, $T_{\text{det}}$ is $1.2\%$, $4.1\%$, and $9.6\%$ at staleness 32, 256, and 512, and $T_{\text{tok}}$ is $2.8\%$, $9.3\%$, and $17.7\%$. $T_{\text{tok}}$ is nearly flat in sequence length ($8.7\%$ at 4k, $9.3\%$ at 16k tokens), so cost does not degrade on long agentic trajectories, and it concentrates on agent-reasoning spans ($T_{\text{det}}$ $7.9\%$ vs. $0.6\%$ tool output and $0.3\%$ system prompt). The stale-1024 outlier batch of Section B.7 is the exception that bounds this claim.

### B.14 Implementation Notes

The GCD correction is implemented as a $\sim$320-line patch on top of verl's GRPO trainer. The non-trivial pieces are: (i) the per-layer power iteration over $\Delta W^{(\ell)}$, batched across all layers in a single CUDA kernel call to amortize launch overhead; (ii) the attention-entropy and $\alpha_{\max}$ capture, exposed via a small modification to FlashAttention-3's epilogue that writes the per-head statistics to a side buffer at no measurable forward-pass overhead; and (iii) the DeltaNet spectral-radius computation via a single power iteration on the linearized state-transition operator, performed once per checkpoint. Source code, evaluation scripts, calibration constants, and trained checkpoints will be released at `https://github.com/EricChan2028/GCD-KV-Drift` under Apache 2.0 upon publication; checkpoints derived from Llama-3.1-8B remain subject to the Llama 3.1 Community License.

### B.15 FP8 versus BF16 sensitivity

We measure the contribution of FP8 mixed-precision arithmetic to the empirical bias by re-running the bias-measurement protocol of Section 6 under BF16 forward and BF16 master weights, on $N = 2{,}048$ tokens at staleness 256. The median $|b_t|$ shifts from $0.0341$ (FP8) to $0.0316$ (BF16), a difference of $0.0025 \pm 0.0008$ nats/token — well within the $\epsilon_{\text{TR}} = 0.05$ trust-region tolerance. Heavier quantization (FP4) is left to future work; we expect re-calibration of $(\kappa_{\text{full}}, \kappa_{\text{delta}})$ to be required, but the structural bound of Theorem 5 should remain valid because the per-layer Jacobian-norm bounds in Assumption 1 are quantization-agnostic.

### B.16 Full result tables

The main text reports these results as figures; the tables give every value. In the tables, (F) and (R) stand for full recompute and naive reuse. The baselines are (B1) on-policy GRPO with full recompute; (B2) verl-default partial rollout, that is PPO clip $\epsilon = 0.2$ with KV reuse, the production configuration and not an unclipped strawman; (B3) PPO with clip on full-recomputed log-probabilities; (B4) DAPO; (B5) M2PO, the primary baseline; (B6) asynchronous RLHF; (B7) Stable Asynchrony; and (B8) GAC. The GCD configurations are (G1) GCD alone, (G2) M2PO+GCD, the headline configuration, (G3) DAPO+GCD, and (G4) GAC+GCD. A0 is the trigger-only variant, with detector and recompute on and the correction forced to 1, and A1 is the correction-only variant, with no recompute.

Table 13: Hidden-state bias $|b_t|$ on generated tokens, via regime-(F) vs. regime-(R) on Qwen3.5-9B, in nats/token. The estimate column evaluates (5) with OLS-fitted constants, including the second-order remainder, so it is an estimate and not a certified bound. These values come from controlled updates on one frozen checkpoint. They are not the values the gauge sees in training, so they cannot be compared directly with $T_{\text{gauge}}$ in Table 18.

| Staleness (updates) | Median $\lvert b_t \rvert$ | $\text{P}_{95}$ $\lvert b_t \rvert$ | Estimate from (5) | Tightness ratio |
|---|---|---|---|---|
| 1 | $0.0009 \pm 0.0001$ | 0.004 | 0.0014 | 0.64 |
| 32 | $0.0061 \pm 0.0004$ | 0.029 | 0.0091 | 0.67 |
| 128 | $0.0214 \pm 0.0011$ | 0.094 | 0.0314 | 0.68 |
| 256 | $0.0341 \pm 0.0017$ | 0.158 | 0.0481 | 0.71 |
| 512 | $0.1080 \pm 0.0058$ | 0.396 | 0.1419 | 0.76 |
| 1024 | $0.2873 \pm 0.0149$ | 0.872 | 0.3608 | 0.80 |

Table 14: MATH-500 accuracy (%) on Qwen3.5-9B + GRPO, 5 seeds. **Bold**: best at each staleness. "—": training diverged (NaN loss within 1 epoch) or, for B1, on-policy only. †Four of five seeds: seed 42 diverged on one outlier batch and is excluded (Section B.7).

| Method | Stale 0 | Stale 32 | Stale 128 | Stale 256 | Stale 512 | Stale 1024 |
|---|---|---|---|---|---|---|
| On-policy GRPO (B1) | $\mathbf{72.8 \pm 0.4}$ | — | — | — | — | — |
| verl-default (B2) | $72.6 \pm 0.5$ | $69.7 \pm 0.8$ | $62.8 \pm 1.6$ | $51.3 \pm 3.4$ | — | — |
| PPO + clip (B3) | $70.1 \pm 0.6$ | $63.4 \pm 1.2$ | $48.7 \pm 3.5$ | — | — | — |
| DAPO (B4) | $71.9 \pm 0.5$ | $69.8 \pm 0.7$ | $63.1 \pm 1.4$ | $52.4 \pm 3.1$ | — | — |
| M2PO (B5) | $72.1 \pm 0.4$ | $71.4 \pm 0.5$ | $69.2 \pm 0.8$ | $66.7 \pm 1.3$ | $48.1 \pm 4.6$ | — |
| Async-RLHF (B6) | $71.6 \pm 0.5$ | $68.2 \pm 0.9$ | $61.4 \pm 1.7$ | $51.8 \pm 3.2$ | — | — |
| Stable-Async (B7) | $71.8 \pm 0.5$ | $69.6 \pm 0.7$ | $64.1 \pm 1.2$ | $54.2 \pm 2.5$ | — | — |
| GAC (B8) | $72.0 \pm 0.4$ | $70.1 \pm 0.7$ | $65.7 \pm 1.1$ | $58.3 \pm 2.0$ | $41.2 \pm 5.1$ | — |
| GCD only (G1) | $72.4 \pm 0.4$ | $71.8 \pm 0.5$ | $69.6 \pm 0.7$ | $66.4 \pm 1.0$ | $58.7 \pm 1.8$ | $42.3 \pm 4.8$ |
| M2PO + GCD (G2) | $72.6 \pm 0.4$ | $\mathbf{72.3 \pm 0.4}$ | $\mathbf{71.7 \pm 0.5}$ | $\mathbf{70.4 \pm 0.6}$ | $\mathbf{67.1 \pm 0.9}$ | $\mathbf{59.4 \pm 1.6}^{\dagger}$ |
| DAPO + GCD (G3) | $72.3 \pm 0.5$ | $72.1 \pm 0.5$ | $70.9 \pm 0.6$ | $68.8 \pm 0.9$ | $63.2 \pm 1.4$ | $51.7 \pm 2.6$ |

Table 15: $\tau$-bench on Qwen3.5-9B + GRPO, 3 seeds. LEFT: retail $\text{pass}^4$ across staleness. RIGHT: $\text{pass}^k$ at staleness 256 on retail (R) and airline (A). Bold = best; "—" = not run.

| | RETAIL $\text{pass}^4$ vs. staleness | | | | Staleness 256 | | | |
|---|---|---|---|---|---|---|---|---|
| Method | 0 | 128 | 256 | 512 | R-$p^1$ | R-$p^8$ | A-$p^4$ | A-$p^8$ |
| Qwen3.5-9B (no RL) | — | — | — | — | 46.2 | 24.1 | 26.4 | 19.2 |
| On-policy GRPO (B1) | 39.7 | — | — | — | 54.1 | 31.4 | 33.1 | 25.7 |
| Async-RLHF (B6) | — | — | 33.8 | — | 49.6 | 25.1 | 27.3 | 20.4 |
| Stable-Async (B7) | — | — | 35.2 | — | 50.7 | 26.8 | 28.6 | 21.6 |
| GAC (B8) | — | — | 36.9 | 14.2±5.6 | 51.8 | 28.5 | 30.2 | 22.9 |
| DAPO (B4) | — | — | 37.6 | — | 52.1 | 29.4 | 31.0 | 23.5 |
| M2PO (B5) | 39.5 | 40.2 | 38.4 | 19.6±4.8 | 52.9 | 30.2 | 31.8 | 24.1 |
| GCD only (G1) | 39.6 | 40.6 | 39.1 | 32.4±1.5 | 53.4 | 30.7 | 32.4 | 24.6 |
| GAC+GCD (G4) | — | — | 42.0 | 34.7±1.1 | 54.1 | 33.4 | 34.5 | 27.1 |
| M2PO+GCD (G2) | **39.9** | **42.3** | **44.1** | **38.5±0.7** | **55.2** | **35.8** | **36.2** | **28.9** |
| *G2 minus M2PO* | +0.4 | +2.1 | +5.7 | +18.9 | +2.3 | +5.6 | +4.4 | +4.8 |

Table 16: Importance weights against ground truth $w^F$, at staleness 256 / 512 (protocol of Table 13, 3 seeds). $w^{\text{GCD}}$ and $w^{\text{GCD-S}}$ apply the correction without selective recompute. "Closer" is the share of tokens nearer to $w^F$ than $w^R$ is; its null is the damping control, not 50%.

| Estimator | $\mathbb{E}\lvert \log w - \log w^F \rvert$ | Closer | Corr. with $w^F$ |
|---|---|---|---|
| $w^R = \pi_{\text{hyb}}/\pi_{\theta_{\text{old}}}$ (naive reuse) | 0.0387 / 0.1216 | — | 0.913 / 0.771 |
| Control: matched-mean damping | 0.0331 / 0.0946 | 0.58 / 0.61 | 0.918 / 0.783 |
| $w^{\text{GCD}}$, magnitude-calibrated | **0.0209 / 0.0538** | **0.68 / 0.74** | **0.951 / 0.883** |
| $w^{\text{GCD-S}}$, signed-calibrated | 0.0174 / 0.0421 | 0.73 / 0.79 | 0.964 / 0.912 |

Table 17: Controls on Qwen3.5-9B, MATH-500, 5 seeds. TOP: isolating the correction from recomputation. BOTTOM: source of the constants. Tightness is median $|b_t|$ over (5). $T_{\text{tok}}$ is measured on $\tau$-bench.

| Variant | $\kappa_{\text{full}}$ | Tightness | $T_{\text{tok}}$@256 | Stale 256 | Stale 512 |
|---|---|---|---|---|---|
| M2PO (B5) | — | — | 0% | $66.7 \pm 1.3$ | $48.1 \pm 4.6$ |
| A0: trigger only, correction forced to 1 | — | — | 9.3% | $67.6 \pm 1.1$ | $53.4 \pm 3.1$ |
| A1: correction only, no recompute | 2.34 | 0.64–0.80 | 0% | $69.8 \pm 0.8$ | $64.2 \pm 1.4$ |
| M2PO+GCD (G2), OLS-calibrated | 2.34 | 0.64–0.80 | 9.3% | $\mathbf{70.4 \pm 0.6}$ | $\mathbf{67.1 \pm 0.9}$ |
| Analytical (Equation (17) at $\lambda = 0$, unfitted) | 4.87 | 0.31–0.42 | 22.8% | $69.8 \pm 0.7$ | $66.3 \pm 1.1$ |
| Deliberate under-estimate ($0.6\times$) | 1.40 | 1.06–1.31 | 4.8% | $68.1 \pm 1.2$ | $61.4 \pm 2.3$ |

Table 18: Recompute accounting and generality. $T_{\text{det}}$, $T_{\text{gauge}}$, $T_{\text{tok}}$ are shares of tokens; $C_F$ is split into KV refresh and DeltaNet re-roll. MATH-500 accuracy over 5 seeds (Qwen3.5-9B) or 3 seeds (Llama-3.1-8B); "div." = diverged in 3/3 seeds. On Qwen3.5-9B the trigger rates and $C_F$ are measured on $\tau$-bench.

| Backbone @ staleness | $T_{\text{det}}$ | $T_{\text{gauge}}$ | $T_{\text{tok}}$ | $C_F$ | KV + re-roll | M2PO | M2PO+GCD |
|---|---|---|---|---|---|---|---|
| Qwen3.5-9B @ 256 | 4.1% | 5.2% | 9.3% | 1.1% | $0.3 + 0.8$ | $66.7 \pm 1.3$ | $\mathbf{70.4 \pm 0.6}$ |
| Qwen3.5-9B @ 512 | 9.6% | 8.1% | 17.7% | 1.9% | $0.5 + 1.4$ | $48.1 \pm 4.6$ | $\mathbf{67.1 \pm 0.9}$ |
| Llama-3.1-8B @ 256 | 6.8% | 7.4% | 14.2% | 1.8% | $1.8 +$ n/a | $64.1 \pm 1.5$ | $\mathbf{67.9 \pm 0.8}$ |
| Llama-3.1-8B @ 512 | 14.9% | 11.3% | 26.2% | 3.0% | $3.0 +$ n/a | div. | $\mathbf{61.2 \pm 1.3}$ |

# C Other Related Work

Section 2 covers the work closest to GCD. This appendix gives the wider context.

**Reproducibility of policy gradients.** Reproducibility studies [Henderson et al., 2018] have long warned that policy-gradient methods are sensitive to seemingly minor implementation details, a concern that the exception-safe design of GCD directly addresses.

**KV cache management.** Beyond the serving systems of Section 2, fine-grained eviction [Chitty-Venkata et al., 2025], further IO-aware kernels [Dao, 2023], and heterogeneous NPU and PIM architectures [Heo et al., 2024] also treat the KV cache as a faithful function of fixed parameters.

**Linear and recurrent attention.** The Qwen3.5 family combines full attention [Vaswani et al., 2017] with Gated DeltaNet layers, building on the broader literature on linear and recurrent attention alternatives [Gu and Dao, 2023, Sun et al., 2024, Behrouz et al., 2024, Liu et al., 2024, Wang et al., 2024c].

**Agentic systems.** Surveys of agentic systems [Wang et al., 2024b, Sapkota et al., 2025] highlight tool integration as the dominant axis of agent capability. Concrete tool-using agents demonstrate this in chemistry [Boiko et al., 2023, Bran et al., 2024], code [Zhang et al., 2024a,b], and chain-of-thought reasoning [Wang et al., 2023a, Chen et al., 2023, Jin et al., 2024, Wang et al., 2024a, Ranaldi and Freitas, 2024].

**Broader context.** Our work intersects with the rapidly maturing literature on LLM behavior [Shanahan et al., 2023], evaluation [Qin et al., 2023, Bang et al., 2023, Yang et al., 2024a], explainability [Zhao et al., 2024, Longo et al., 2024], and applications across foundation-model deployments [Singhal et al., 2023, Qiu et al., 2023, Banh and Strobel, 2023, Wang et al., 2023b, Li et al., 2024, Lu et al., 2025]. While these works address downstream concerns, GCD targets a training-time bias that all of them inherit when their underlying models are trained off-policy. Foundational architectures [Vaswani et al., 2017, Touvron et al., 2023] provide the substrate on which the bias manifests.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each headline claim in the abstract maps to a specific result: the bias bound (Theorem 5, validated by Table 13); the staleness-512 stability claim (Tables 14 and 18); the $1.38\times$ and $1.27\times$ rollout throughput over full recomputation (Table 7); the 5.7-point retail $\text{pass}^4$ gain (Table 15); and the check against full recomputation (Table 16). The abstract and Section 7 state the scope of these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the scope of the claims (Qwen3.5 and Llama-3.1-8B on a single hardware platform, with MoE routing staleness and cross-node FSDP unvalidated), the per-token direction of the correction, the first-order regime of validity for Theorem 5, the throughput regression against naive reuse on a pure Transformer, and pre-registered failure modes (R1–R4) including the partial refutation of R2. Section 6 covers calibration sensitivity.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Assumptions 1 to 4 state the assumptions for the theoretical results: Theorem 5 (high-probability concentration estimate), Theorem 6 (tail-event detector), Corollary 7 (composite safety guarantee), Theorem 8 (M2PO trust-region preservation), and Proposition 9 (sign of the bias). Full proofs are in Section A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Hyperparameters, hardware, software versions, random seeds, and per-experiment compute are documented in Section B. Code release is committed in the same appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code, evaluation scripts, and trained checkpoints will be released at https://github.com/EricChan2028/GCD-KV-Drift under Apache 2.0 upon publication; checkpoints derived from Llama-3.1-8B remain subject to its license. All datasets used (MATH-500, $\tau$-bench retail/airline) are publicly available under their respective licenses.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Table 3 lists all hyperparameters; the rollout-train coordination protocol matches verl's defaults except where stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: MATH-500 results report mean $\pm$ standard deviation across 5 seeds, or 3 seeds on Llama-3.1-8B; the M2PO+GCD entry at staleness 1024 covers 4 of 5 seeds, as disclosed in Table 14 and Section B.7. $\tau$-bench results report means over 3 seeds, with a standard deviation at staleness 512, and throughput figures are point measurements (Tables 7 and 15). Significance tests use a paired bootstrap with $10^4$ resamples; we report $p$-values where claims of superiority are made.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 4 reports per-experiment GPU-hours; total budget is 7,414 H100-hours on 4×H100 80GB SXM5 with NVLink. Main runs used rented cloud GPUs, and the post-review experiments used GPUs at Tsinghua University with the same node configuration. Software stack (CUDA 12.5, PyTorch 2.5.1 FP8, vLLM 0.7.0, FlashAttention-3, FSDP-2, verl) is documented in Section B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work uses only publicly released model weights and benchmarks. No human subjects or personally identifiable information is involved.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The work reduces the compute cost of LLM RL post-training, which on balance lowers the carbon footprint of capable model training. The same efficiency improvement could in principle accelerate the deployment of misaligned agentic systems; Section 8 notes both effects.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released artifacts are a training-time correction (a $\sim$320-line patch on top of verl) and post-trained checkpoints of publicly available models, with no capability uplift beyond what RL post-training of Qwen3.5 and Llama-3.1-8B already gives.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Qwen3.5 weights (Apache 2.0), Llama-3.1-8B weights (Llama 3.1 Community License), MATH-500 (MIT), $\tau$-bench (Apache 2.0), verl (Apache 2.0), vLLM (Apache 2.0), FlashAttention (BSD-3) are all credited and used under their stated licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The GCD reference implementation and calibration script will ship with API documentation, a reproducibility README, and a JSON manifest of the calibrated constants for each Qwen3.5 scale and for Llama-3.1-8B.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects research was conducted.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research was conducted.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The work studies the training of LLMs themselves; all uses of LLMs are described in detail in Sections B and 6. No LLMs were used to write the paper itself in a non-standard way.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.