PRISM: PROBING AND REPAIRING INTERNAL STATES IN LANGUAGE MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Internal monitors reveal changes that aggregate training statistics obscure, but deciding when to intervene requires understanding their consequences. We present PRISM, a framework that connects component-level measurements to counterfactual probes and matched interventions in language model pretraining. Across dense and mixture-of-experts models, these comparisons reveal an intervention gap: suppressing an activation tail can leave training worse, and restoring expert opportunities can coexist with higher observed loss. Conversely, a small aggregate error in optimizer state can precede a catastrophic update. To recover from this failure, we introduce protected-summary recovery, which reconstructs damaged Adam second moments from one independently maintained total per tensor and intact momentum. A constrained reconstruction combines this total with a known moment bound; an equal-information mean reconstruction isolates the role of coordinate allocation. Both rules contain update transients and achieve lower early-horizon negative log-likelihood than moment reset in all five matched execution blocks across two dense settings. PRISM turns internal observations into experimentally testable intervention choices, with compact optimizer-state recovery as a concrete application.

1 INTRODUCTION

Loss, gradient norm, and throughput summarize language model training but obscure individual mechanisms: which experts receive useful training, whether an attention statistic reflects a functional change, or whether a small optimizer error will produce a dangerous update. Mechanism-driven monitors offer a richer view (Huang et al., 2026). The harder question is which observations justify intervention.

Our experiments expose an *intervention gap*. Reducing an activation tail changes the statistic without improving training. Correcting expert imbalance restores training opportunities while yielding slightly higher observed language-modeling loss. Conversely, approximately 0.58% relative RMS error in an Adam second moment can precede a catastrophic update. The consequences depend on which mechanism changes and how it affects subsequent training.

PRISM provides *action-specific qualification*: evaluating an internal signal through the consequences of a candidate intervention. Its workflow combines component-level measurement, counterfactual probes and matched training continuations. Counterfactuals separate changes in representation from changes in function and optimizer response; continuations measure effects on the target mechanism, prediction quality and cost. Global-signal baselines, alternative actions and held-out contexts test where internal measurements add useful information.

The numerical-state study yields *protected-summary recovery*. One independently maintained second-moment total per tensor retains historical scale during native Adam training. After storage damage, that total and surviving momentum constrain reconstruction without a full backup. The experiments separate two questions: whether retained history aids recovery, and how to allocate the reconstructed total across coordinates.

Our contributions are: (1) a six-family measurement system with explicit identity, sampling, timing, coverage and cost; (2) counterfactual and local-response assays separating representation, function and optimization, with matched mechanism and quality outcomes; and (3) protected-summary re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

covery, tested against equal-information reconstruction and lower-information alternatives. Table 1 covers the first two contributions across five mechanisms; Tables 2 and 3 test the third.

2 Related Work

Training diagnostics and internal mechanisms. Mechanism-Driven Monitors links QK/router signals to instability (Huang et al., 2026); program diagnosis recommends fixes (Wardat et al., 2022), while latent-state models describe dynamics (Hu et al., 2023). Prior work studies attention concentration and outliers (Henry et al., 2020; Zhai et al., 2023; Sun et al., 2024), balancing and residual mixing (Wang et al., 2024; Xie et al., 2025). We distinguish matched action qualification from detection and mechanism discovery.

Optimizer state design. Adam/AdamW define the native updates (Kingma & Ba, 2015; Loshchilov & Hutter, 2019). Factored/blockwise states, initialization, momentum-dependent denominators and spike-aware resets alter their adaptation (Shazeer & Stern, 2018; Zhang et al., 2025b; Abuduweili & Liu, 2025; Zhang et al., 2025a; Huang et al., 2025). For matched blocks and gradients, our total divided by block size follows Adam-mini's squared-gradient EMA. We retain this quantity independently alongside native Adam and use it for one-time state reconstruction after damage. Low-bit state zero-point hazards are known (Li et al., 2023).

Checkpointing and fault recovery. ExCP compresses checkpoints, ECRec codes/replicates state, and Checkmate maintains shadow state (Li et al., 2024; Zhang et al., 2023; Bhardwaj et al., 2026). Numerical checksums and attention-specific error correction also precede this work (Tao et al., 2016; Liang et al., 2025). We study approximate reconstruction of damaged second moments from a protected total with intact momentum, weights and clocks.

3 Methods

3.1 Problem Formulation

An internal measurement describes a training state; an intervention changes its future trajectory. PRISM asks what evidence connects the two, taking a signal–action pair in a specified context and horizon as the unit of analysis (Figure 1).

Let $s_t = (\theta_t, o_t, r_t, d_t)$ contain parameters, optimizer state, randomness, and data position; c specifies architecture, precision, schedule, and fault regime. A monitor produces $z_t = M(s_t, x_t)$ on input x_t, with component identity, sampling support and reduction defining the measured object. Appendix A gives these definitions for the six monitor families.

For action a and horizon h, the object of interest is the paired effect

$$\boldsymbol{\tau}_h(a; c) = \mathbf{Y}_{t:t+h}(s_t, a; c) - \mathbf{Y}_{t:t+h}(s_t, \varnothing; c), \quad \mathbf{Y} = (H, Q, C). \tag{1}$$

Here $\varnothing$ denotes no intervention; H measures a mechanism consequence, such as non-finite updates or persistent expert deprivation; Q measures token prediction quality; and C records cost. Matched continuations share the starting state and future data. Reporting these outcomes separately captures both the intended mechanism change and its consequences for training.

3.2 Action-Specific Qualification

The analysis follows three steps: identify the measured component and operation, probe the relationship between its statistic and function, and compare interventions from matched training states. Predictive and held-out tests then assess which observations remain useful across contexts.

Representation and function. A transformation G satisfying $f_{G(s)}(x) = f_s(x)$ provides a controlled test of representation sensitivity: the measured statistic can change while the current function stays fixed. Our LatentMoE probe changes the latent basis while compensating expert input/output maps. Same-input forward agreement and subsequent update KL measure function and

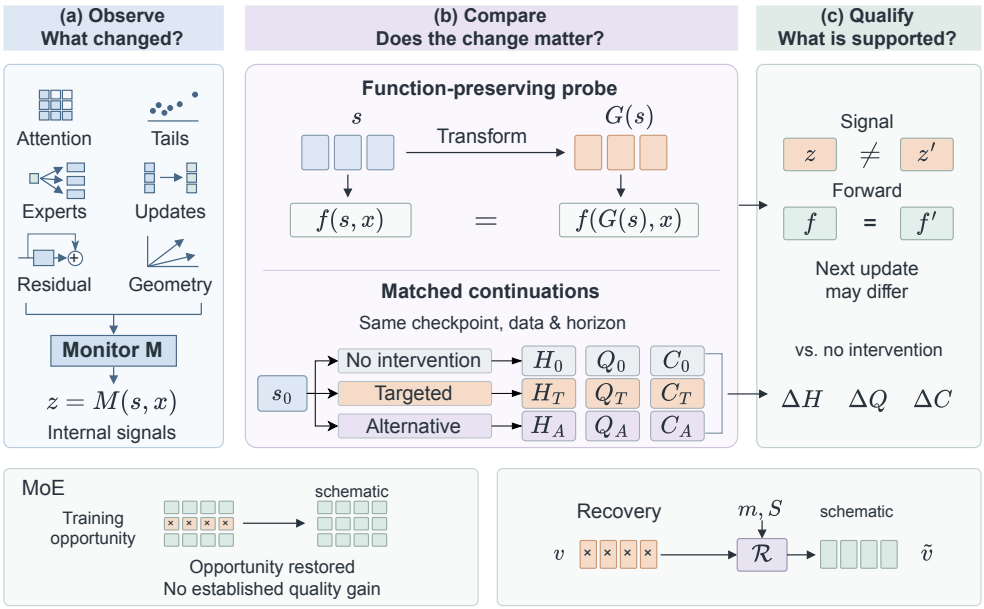


Figure 1: PRISM qualifies signal–action pairs in a specified context and horizon. Counterfactual probes separate representation from current function; matched continuations compare mechanism (H), quality (Q) and cost (C). Arrows denote experimental comparisons used to evaluate signals and interventions.

optimizer response separately. For residual gates, branch write—full layer output minus residual-only transport—measures what the branch adds to the residual stream. Its RMS captures write magnitude, distinct from task necessity or marginal quality contribution (Appendix D).

Target engagement and benefit. We compare targeted, no-intervention and alternative actions using Eq. 1. The targeted action tests the proposed mechanism; alternatives test whether another intervention achieves the same outcome. Measuring both H and Q distinguishes successful tail suppression or router balancing from an improvement in training quality.

Predictive information and transport. Predictors with and without the added measurements test their information value under run-, stage- or action-held-out splits. Global features are held fixed when testing the contribution of module-local signals. Known fault coordinates and simple predictors provide controls for local response analysis; frozen transfer tests reuse beyond the fitted context. These are diagnostic assays, while the repair actions are specified separately.

We summarize the analyses as $\mathcal{Q}(M, a; c, h) = (V, I, L, E, D, T)$: measurement validity, incremental information, mechanism localization, target engagement, action consequences and transport. The entries record the outcomes of the comparisons above. Appendix A gives a worked example.

3.3 Protected-Summary Recovery

Protected-summary recovery reconstructs a damaged second moment from an independently retained total and intact momentum, once before its next application (Figure 2). Ordinary training updates remain native Adam.

Recovery setting. An Adam second moment v_t is damaged after a safe update and before the next application. Parameters, first moment m_t, bias-correction clocks and an independent summary remain intact. Recovery replaces v_t under this one-sided storage-fault assumption.

A scalar memory of scale. For one logical tensor with n entries, initialize $S_0 = \sum_i v_{0,i}$ from a trusted state and maintain

$$S_t = \beta_2 S_{t-1} + (1-\beta_2)\|g_t\|_2^2, \tag{2}$$

using Adam's actual clipped gradient. Summing Adam's coordinate-wise recurrence gives $S_t = \sum_i v_{t,i}$ in exact arithmetic before corruption. Independent maintenance preserves this total along the consumed gradient history even when stored v_t is damaged. Healthy training still uses native Adam. The protected total is in-memory state excluded from the injected fault.

A feasible reconstruction. For Adam's uncorrected moments with a faithful zero-initialized history, or a starting state satisfying the same invariant, and $\beta_1^2 < \beta_2$, weighted Cauchy–Schwarz gives

$$v_{t,i} \geq c\, m_{t,i}^2, \qquad c = \frac{(1-\beta_2)(1-\beta_1^2/\beta_2)}{(1-\beta_1)^2}. \tag{3}$$

At $(\beta_1, \beta_2) = (0.9, 0.95)$, $c = 14/19$, conservatively below the finite-time bound. This known Adam constraint (Ma & Yarats, 2019) supplies coordinate-wise lower bounds $L_i = cm_i^2$, while S fixes the total. The remaining mass is $B = S - \sum_i L_i$. For $B \geq 0$, distributing it uniformly yields

$$\widetilde{v}_i = L_i + B/n. \tag{4}$$

This reconstruction satisfies both constraints in exact arithmetic without a fitted coefficient. The moment invariant also bounds the next adaptive-update coordinate under Adam's recurrence (Appendix G). Uniform allocation maximizes the entropy of the residual mass. Feasibility and update bounds follow algebraically; the experiments evaluate the resulting training behavior.

Separating information from shape. The equal-information control $\widetilde{v}_i = S/n$ retains the same momentum, clocks, weights and total: comparison with Eq. 4 isolates coordinate allocation. Both rules preserve the total; only the constrained rule enforces the moment lower bounds. Clock-preserving reset clears both moments and tests the combined value of retained momentum and the protected total. A momentum-preserving floor and persistent update clipping provide lower-information controls. We evaluate each reconstruction as a fixed rule.

Intervening before damage propagates. An every-update alarm checks positive-support loss at identical sampled coordinates. On a trigger, the implemented scalar path reconstructs all supported tensors before the next apply; validation failure halts execution, while success disables further casts. Detection targets sampled support collapse, whereas reconstruction uses intact m, S across supported coordinates. Appendices A and G specify sampling, thresholds, ownership and FP32 safeguards.

4 Experiments

Design and scope. Central comparisons use 22-layer dense GQA (284M), MLA (378M) and separate LatentMoE (774M total/350M active) settings: 4,096-token sequences, global batch 128, BF16 forward and FP32 master/state where applicable. Dense parents consume 2.84B/3.59B nominal tokens; standard continuations add 0.81B/0.54B, longer tests 2.15B. Additional MHA, hybrid-attention, Muon/AdamW and four-stream cohorts examine complementary mechanisms (Appendix B).

Matched actions share parent, future inputs and settings except for the treatment. Quality is valid-token-weighted NLL on an 8,192-sequence panel; intervals condition on the trained weights and evaluation panel. Appendix C specifies the resampling procedure, replication units, content checks and exclusion of legacy reader-affected endpoints.

Table 1 links signals, consequences and interventions across mechanisms, using matched comparisons within each study. Table 2 fixes the fault, action timing and horizon within each dense setting.

4.1 Diagnosis requires mechanism and function

An FP16 second-moment round-trip leaves 99.936% of sampled entries zero despite only 0.582% relative RMS error and cosine 0.999983 (Fig. 3). Loss at the fault boundary is unchanged. Next-update RMS grows from 1.96×10^{-4} to 1.10×10^{4}, followed by a non-finite forward pass. Both

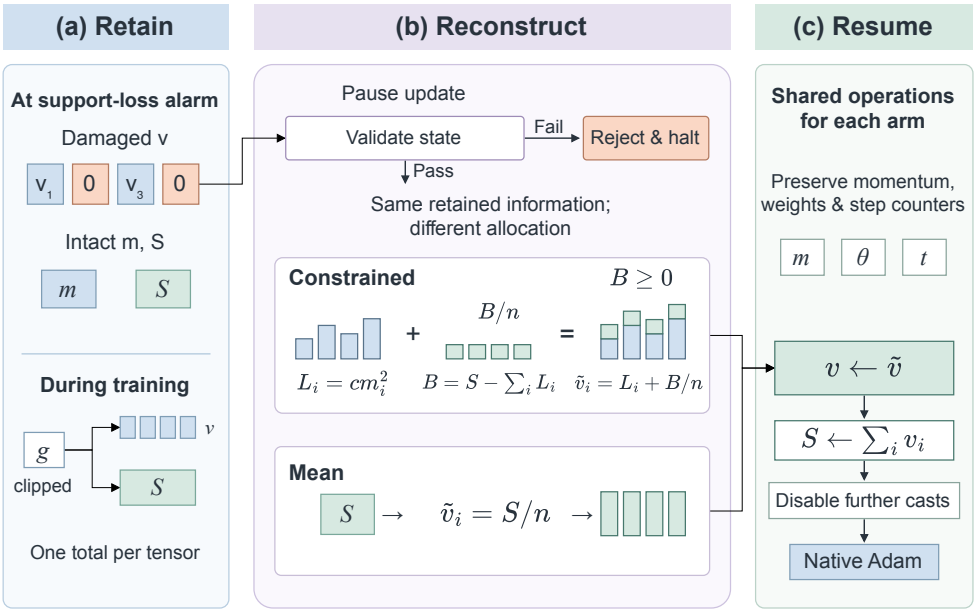


Figure 2: Protected-summary recovery retains one total S per tensor and reconstructs before the next apply, preserving weights, momentum and clocks. The constrained rule enforces the moment bound; the mean control isolates reconstruction shape. Here n counts entries, c is defined in Eq. 3, and $B \geq 0$ denotes exact-arithmetic feasibility.

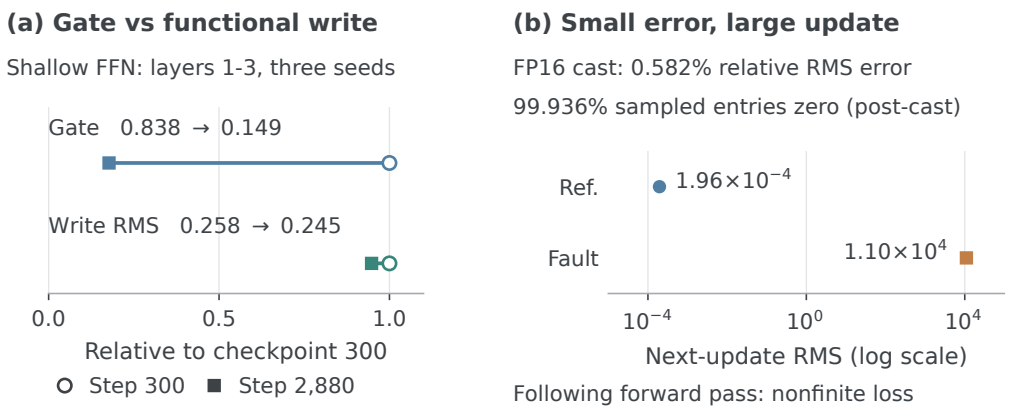


Figure 3: Contrasting diagnostic cases: a large gate decline leaves write magnitude nearly unchanged, whereas a small aggregate state error precedes a catastrophic update. The panels have different contexts and scales.

support loss and the prospective-update guard detect damage before unsafe application. The support measurement locates the failure in second-moment storage, enabling intervention before it propagates to weights. Full-state rollback/replay restore finite training (Appendix F); low-bit zero-point hazards are known (Li et al., 2023).

Table 1C–E separates changes in a statistic, current function and next update. Tail rotation changes p99 at nearly fixed norm, yet native training has the lowest loss area in all six contexts. Deep FFN write grows as its mHC gate falls. Coordinate stress approximately preserves the forward function but changes within-arm next-update KL; rebalancing brings this scalar magnitude closer to native. Together, these cases show why diagnosis must examine the affected computation alongside the monitored statistic. Context-level contrasts and probe definitions appear in Appendix D.

Table 1: **System-level results: from an internal signal to an action-specific conclusion.** A separates a numerical-fault timeline from matched recovery experiments; recovery ΔNLL is constrained minus reset at common $\epsilon = 10^{-8}$. B: terminal deprived identities / cumulative expert-window burden, one parent. C: median action-minus-native changes across six model-stage contexts; area is duration-normalized training loss. D: steps 300–2,880, shallow/deep FFN means over layers 1–3/4–21 and three seeds. E: diagonal condition $\kappa = 100$ (entries 0.1–10); forward KL and within-arm next-update KL measure different responses. Blocks A–E report mechanism-specific measurements and intervention outcomes.

Condition / action	Internal response	Functional / training outcome	Observed outcome
A. Numerical state: FP16 second-moment cast; dense GQA / MLA			
State cast (MLA timeline)	Relative RMS error 0.582%; zero fraction 99.936%	Dangerous next update; subsequent forward nonfinite	Support and update guards detect damage
Protected-summary recovery	Retain total and momentum; reconstruct second moments	ΔNLL vs reset: GQA -0.001399; MLA -0.002232	Lower early-horizon NLL in matched recovery tests
B. Expert opportunity: elevated router LR, balancing off; LatentMoE			
No action	Identities / burden: $9 / 401$	NLL 1.205670	Persistent deprivation baseline
Early targeted balancing	Identities / burden: $0 / 0$	NLL 1.207295	Opportunity restored; slightly higher endpoint NLL
Later global-rule balancing	Identities / burden: $0 / 236$	NLL 1.209668	Timing changes cumulative deprivation
Early learning-rate reduction	Identities / burden: $9 / 442$	NLL 1.237810	Overall LR reduction leaves deprivation
C. Activation tails: norm-preserving rotation; Dense / MoE, three stages			
Tail down	Δp99: -4.15%	Δloss area: +0.000157	Smaller tail; higher loss area
Tail up	Δp99: $+2.55\%$	Δloss area: +0.000076	Larger tail also raises loss area
Random rotation	Δp99: —	Δloss area: +0.000200	Matched-norm direction control
D. Residual transport: natural training; four-stream mHC			
Shallow FFN	Gate: $0.838 \to 0.149$	Write RMS: $0.258 \to 0.245$	Smaller gate; similar write magnitude
Deep FFN	Gate: $1.118 \to 0.584$	Write RMS: $0.342 \to 0.857$	Write magnitude can grow as the gate shrinks
E. Matrix coordinates: controlled diagonal stress; MLA LatentMoE			
Native	Forward KL: 0.000000	Mean next-update KL: 0.031082	Native functional-response reference
$\kappa = 100$	Forward KL: 0.000096	Mean next-update KL: 0.050824	Nearly unchanged forward map; altered optimization response
Rebalanced	Forward KL: 0.000127	Mean next-update KL: 0.031573	Next-update KL magnitude nearer native

Across 325 runs, adding gradient/update features *and global gradient norm* improves final-window loss prediction over context and early loss. This comparison measures their joint contribution. In a separate 72-row MoE assay, internal responses improve later-response prediction within calibrated stages, but the gains disappear under stage holdout. Locally calibrated optimizer-state measurements outperform global/guard inputs; simple same-budget selectors match the proposed selector,

Table 2: **Protected-summary recovery: matched method comparison.** A: common $\epsilon = 10^{-8}$, 1,023 post-action updates. E_{16} sums squared FP32-master updates over the last safe and 15 post-action updates. Damaged arms stop casts. Reset clears both moments but keeps clocks; both summary rules retain momentum, clocks and total. B: paired NLL differences (negative favors first), with 95% fixed-panel, 64-record-block intervals conditional on weights. C: repeated executions sharing parents, RNG, inputs and settings, not independent pretraining seeds.

A. Quality and early update energy

	GQA 284M		MLA 378M	
Continuation	NLL	E_{16}	NLL	E_{16}
Healthy	1.438345	68.19	1.345590	90.69
Stop cast, no reconstruction	1.611137	131692.42	1.472737	98857.57
Reset moments	1.440749	181.60	1.349620	236.36
Protected-total mean	1.439971	40.46	1.344943	53.44
Protected-total constrained	1.439350	36.17	1.347387	50.57

B. Paired quality contrasts in the same settings

	GQA 284M		MLA 378M	
Contrast	ΔNLL	95% interval	ΔNLL	95% interval
Mean – Reset	-0.000778	$[-0.001394, -0.000138]$	-0.004677	$[-0.005860, -0.003664]$
Constrained – Reset	-0.001399	$[-0.002110, -0.000704]$	-0.002232	$[-0.003357, -0.001246]$
Constrained – Mean	-0.000621	$[-0.001029, -0.000214]$	+0.002445	$[+0.001874, +0.003070]$

C. Repeated executions: endpoint NLL

Setting / repeat	Healthy	Reset	Mean	Constrained
GQA / repeat 1	1.438466	1.440604	1.439929	1.438309
GQA / repeat 2	1.437684	1.440203	1.439264	1.438688
MLA / repeat 1	1.346416	1.348829	1.345281	1.346108

and frozen transfer fails (Appendix I). The useful information is therefore specific to the prediction target and calibration context.

4.2 Target engagement and quality are separate outcomes

Load CV summarizes imbalance, whereas expert-identity histories reveal repeated deprivation of both tokens and updates. We compare actions from one parent trained with elevated router LR and balancing disabled. Deprivation is already present at restoration: the internal criterion starts balancing at step 6,688, versus 11,344 under the global rule. Balancing acts across all routed layers. This comparison measures how post-onset intervention timing affects cumulative deprivation.

Early balancing prevents measured post-fork burden accumulation; reducing all learning rates leaves deprivation in place (Table 1B). Early balancing's endpoint NLL is 0.001625 above no action. This small difference occurs alongside NLL variation in healthy zero-action continuations (Appendix C). The clearest effect is restoration of expert opportunities by the existing balancing rule, with quality measured as a separate outcome.

4.3 Recovery from compact trustworthy history

We damage second moments after a safe update and repair before the next application, retaining weights, momentum and clocks for both summary rules. GQA includes a 512-update healthy maintenance prefix; each setting then has 1,023 post-action updates. Table 2 switches all roles to common $\epsilon = 10^{-8}$ after restoring parents trained with 10^{-15}. Table 3A retains the original epsilon.

Retained-history recovery. Both summary rules yield lower observed NLL than reset in all five matched execution blocks: the two original comparisons and three repeats (Table 2). Constrained-minus-reset ranges from -0.002295 to -0.001399 in GQA and from -0.002721 to -0.002232 in MLA. Stopping casts alone leaves much worse quality. For the original comparisons, constrained-

Table 3: **Recovery controls beyond the main comparison.** A: original $\epsilon = 10^{-15}$, same action timing and post-action horizon as Table 2. Two fixed panels evaluate the same weights. B: matched original-epsilon MLA, same first panel; both arms retain momentum and clocks. The floor has no protected total and uses surviving damaged coordinates. E_{16} uses the last-safe-inclusive convention of Table 2.

A. Original epsilon 10^{-15}; two fixed quality panels

	GQA 284M			MLA 378M		
Continuation	NLL 1	NLL 2	E_{16}	NLL 1	NLL 2	E_{16}
Healthy	1.439940	1.449600	70.49	1.348442	1.368485	92.42
Reset moments	1.441770	1.451396	217.46	1.349869	1.371214	281.40
Protected-total mean	1.440863	1.450543	40.63	1.347205	1.368380	53.79
Protected-total constrained	1.439646	1.449905	36.21	1.347240	1.368347	50.91

B. Preserving momentum without a protected total: MLA

Action	NLL	E_{16}
Moment floor, no protected total	1.350431	474.09
Protected-total constrained	1.347240	50.91

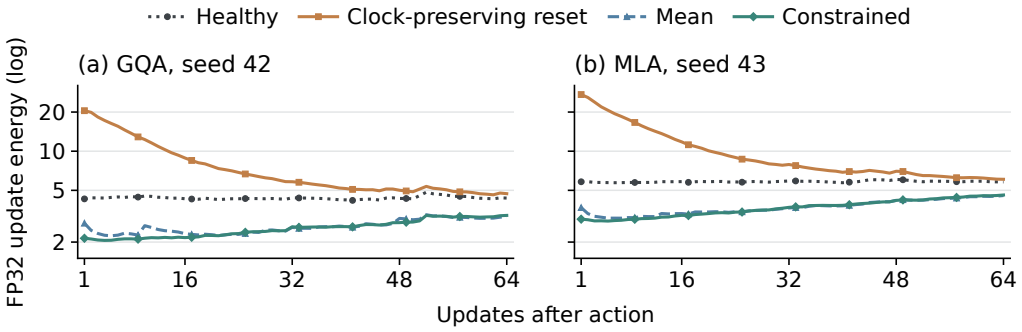


Figure 4: Recovery dynamics at original epsilon: unsmoothed FP32 master-update energy for individual continuations. The 64 post-action updates exclude the last safe update; healthy uses the same boundary. Both scalar rules avoid reset's larger transient.

versus-reset intervals remain below zero at all five tested block sizes; GQA mean-versus-reset includes zero at coarser sizes (Appendix H.4).

Reconstruction shape. At common epsilon, constrained reconstruction yields lower NLL than equal-information mean in every GQA execution but higher NLL in every MLA execution. Both allocations improve early recovery relative to reset, while their ranking depends on the setting. The moment bound gives constrained reconstruction an algebraic property; the shared information, rather than one consistently preferred shape, is the common ingredient in these successful recoveries.

At the original-epsilon MLA endpoint, constrained recovery's NLL is 0.003190 below the momentum-preserving floor, with interval $[-0.004131, -0.002296]$ (Table 3B). It also outperforms persistent update clipping: paired differences are -0.002133 and -0.001383 on two fixed panels, with intervals below zero at all five tested block sizes (Appendix H.4). These momentum-preserving controls extend the comparison beyond reset. Both scalar reconstructions also reduce reset's early update transient (Fig. 4).

Persistence and transfer. An additional MLA parent repeats the improvement over reset, with no resolved difference between reconstruction shapes. In longer runs retaining original $\epsilon = 10^{-15}$, constrained-versus-reset intervals include zero at parent plus 4,096 updates in both settings and panels; MLA mean still beats reset, while constrained loses to mean on one panel (Fig. 5). A direct early–late interaction finds attenuation of GQA's constrained contrast on panel 1 and remains unresolved on panel 2 (Appendix H).

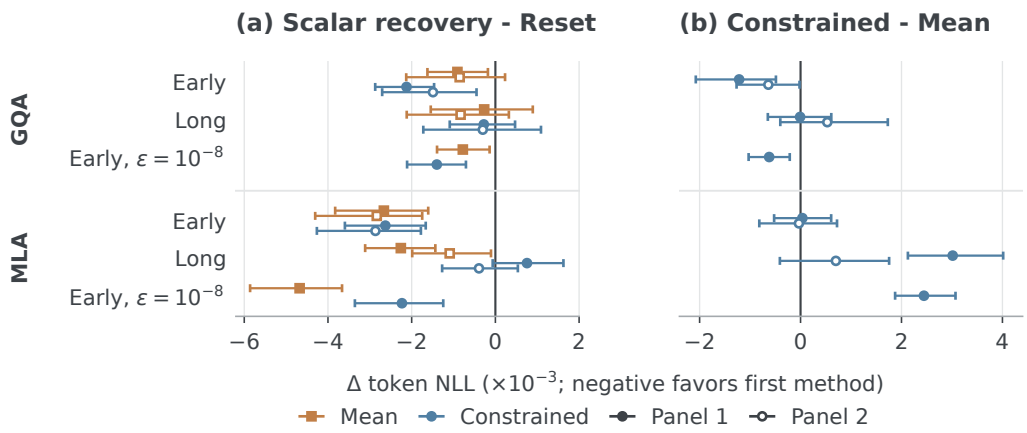


Figure 5: Recovery across controls and horizons. Early: 1,023 post-action updates; long: parent plus 4,096 (3,583/4,095 post-action for GQA/MLA). Unmarked rows use $\epsilon = 10^{-15}$; 10^{-8} rows are separate continuations. Filled/open symbols denote fixed evaluation panels 1/2 of the same endpoint weights; intervals are paired panel uncertainty.

4.4 Cost of observation and retained history

Selected collectors cost less than all-on bundles (Table 4). Summary maintenance adds 3.96%/4.43% in GQA/MLA above a monitored, detector-enabled baseline, with 712/1,064 bytes of totals excluding buffers and detector history. Appendix K separates these measured maintenance costs from the remaining costs of complete fault protection.

Table 4: **Measured monitoring and summary-maintenance costs.** A: seconds/step; selected is the sampled Adam state/guard subset, unavailable for MoE Muon. MLA's cross-revision ratio is descriptive. B: OFF/ON/ON/OFF; the support detector stays enabled in both conditions. Payload counts FP32 totals only. A/B profiles differ (Appendix K).

A. Monitoring scope and end-to-end step time

Setting	Global	Selected	All-on	All-on change
GQA monitors	0.933	1.070	2.001	114.43%
MLA monitors	1.005	1.122	1.809	79.89%
MoE monitors	1.183	—	1.433	21.05%

B. Additional protected-summary maintenance

Setting	OFF (s/step)	ON (s/step)	Added time	Payload
GQA	2.442	2.538	3.96%	712 B
MLA	2.808	2.932	4.43%	1,064 B

5 Conclusion

PRISM connects component-identified measurements, counterfactual probes and matched interventions to separate mechanism restoration from quality benefit. Protected-summary recovery approximately reconstructs Adam second moments from one retained total per tensor and intact momentum under a one-sided storage fault. Both tested rules contain update transients and yield lower early-horizon NLL than reset across two dense settings and their fixed-prefix repeats. PRISM thus provides testable intervention evidence and usable optimizer-state recovery without a full second-moment backup, leaving healthy Adam unchanged.

AI USE STATEMENT

We used generative AI tools to assist with literature discovery, experimental design, implementation and debugging, and manuscript polishing.

We did not use generative AI to generate synthetic training or evaluation datasets. The experimental results reported in this paper were obtained from executed experiments and analyses of their outputs.

AI-assisted implementations were tested, numerical claims were checked against experimental artifacts, and references were verified against primary sources.

REFERENCES

Abulikemu Abuduweili and Changliu Liu. Revisiting the initial steps in adaptive gradient descent optimization. In *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 305–322. PMLR, 2025. URL https://proceedings.mlr.press/v280/abuduweili25a.html.

Ankit Bhardwaj, Weiyang Wang, Jeremy Carin, Adam Belay, and Manya Ghobadi. Checkmate: Zero performance overhead model checkpointing via network gradient replication. In *23rd USENIX Symposium on Networked Systems Design and Implementation (NSDI 26)*, pp. 339–358. USENIX Association, 2026. URL https://www.usenix.org/conference/nsdi26/presentation/bhardwaj.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.379. URL https://aclanthology.org/2020.findings-emnlp.379/.

Michael Y. Hu, Angelica Chen, Naomi Saphra, and Kyunghyun Cho. Delays, detours, and forks in the road: Latent state models of training dynamics. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=NE2xXWo0LF. December 2023; corrected author version arXiv:2308.09543v3, January 2024.

Ruixuan Huang, Hantao Huang, Yifan Huang, Ansheng You, Zhenxing Zhang, and Shuai Wang. Mechanism-driven monitors for preemptive detection of LLM training instability. *arXiv preprint arXiv:2606.28116*, 2026. URL https://arxiv.org/abs/2606.28116. Version 2, 27 July 2026.

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

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

Bingrui Li, Jianfei Chen, and Jun Zhu. Memory efficient optimizers with 4-bit states. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/3122aaa22b2fe83f9cead1a696f65ceb-Abstract-Conference.html.

Wenshuo Li, Xinghao Chen, Han Shu, Yehui Tang, and Yunhe Wang. ExCP: Extreme LLM checkpoint compression via weight-momentum joint shrinking. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 27575–27588. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24m.html.

Yuhang Liang, Xinyi Li, Jie Ren, Ang Li, Bo Fang, and Jieyang Chen. ATTNChecker: Highly-optimized fault tolerant attention for large language model training. In *Proceedings of the 30th ACM SIGPLAN Annual Symposium on Principles and Practice of Parallel Programming*, pp. 252–266. ACM, 2025. doi: 10.1145/3710848.3710870. URL https://www.cs.uoregon.edu/Reports/DRP-202505-Liang.pdf.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Jerry Ma and Denis Yarats. Quasi-hyperbolic momentum and Adam for deep learning. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1810.06801. Appendix F; corrected author version 4, 2 May 2019.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604. PMLR, 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2402.17762.

Dingwen Tao, Shuaiwen Leon Song, Sriram Krishnamoorthy, Panruo Wu, Xin Liang, Eddy Z. Zhang, Darren Kerbyson, and Zizhong Chen. New-Sum: A novel online ABFT scheme for general iterative methods. In *Proceedings of the 25th ACM International Symposium on High-Performance Parallel and Distributed Computing*, pp. 43–55. ACM, 2016. doi: 10.1145/2907294.2907306. URL https://doi.org/10.1145/2907294.2907306.

Lean Wang, Huazuo Gao, Chenggang Zhao, Xu Sun, and Damai Dai. Auxiliary-loss-free load balancing strategy for mixture-of-experts. *arXiv preprint arXiv:2408.15664*, 2024. URL https://arxiv.org/abs/2408.15664.

Mohammad Wardat, Breno Dantas Cruz, Wei Le, and Hridesh Rajan. DeepDiagnosis: Automatically diagnosing faults and recommending actionable fixes in deep learning programs. In *Proceedings of the 44th International Conference on Software Engineering*, pp. 561–572. ACM, 2022. doi: 10.1145/3510003.3510071. URL https://lab-design.github.io/papers/ICSE-22c/.

Mitchell Wortsman, Peter J. Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D. Co-Reyes, Izzeddin Gur, Abhishek Kumar, Roman Novak, Jeffrey Pennington, Jascha Sohl-Dickstein, Kelvin Xu, Jaehoon Lee, Justin Gilmer, and Simon Kornblith. Small-scale proxies for large-scale Transformer training instabilities. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d848cb2c84f0bba7f1f73cf232734c40-Abstract-Conference.html.

Zhenda Xie, Yixuan Wei, Huanqi Cao, Chenggang Zhao, Chengqi Deng, Jiashi Li, Damai Dai, Huazuo Gao, Jiang Chang, Kuai Yu, Liang Zhao, Shangyan Zhou, Zhean Xu, Zhengyan Zhang, Wangding Zeng, Shengding Hu, Yuqing Wang, Jingyang Yuan, Lean Wang, and Wenfeng Liang. mHC: Manifold-constrained hyper-connections. *arXiv preprint arXiv:2512.24880*, 2025. URL https://arxiv.org/abs/2512.24880. First posted 31 December 2025; version 2, 5 January 2026.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhai23a.html.

Huishuai Zhang, Bohan Wang, and Luoxin Chen. AdamS: Momentum itself can be a normalizer for LLM pretraining and post-training. In *Proceedings of the 2025 Conference on Empirical*

Methods in Natural Language Processing, pp. 10719–10738. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.543. URL https://aclanthology.org/2025.emnlp-main.543/.

Tianyu Zhang, Kaige Liu, Jack Kosaian, Juncheng Yang, and Rashmi Vinayak. Efficient fault tolerance for recommendation model training via erasure coding. *Proceedings of the VLDB Endowment*, 16(11):3137–3150, 2023. doi: 10.14778/3611479.3611514. URL https://www.cs.cmu.edu/~juncheny/publication/vldb23-dlrm.pdf.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P. Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/45ae878717399e6f62d57c65f052cd46-Abstract-Conference.html.

A Measurement definitions and collection procedures

PRISM associates each observation with a model component, sampling support, reduction rule and optimizer timing. The definitions below distinguish local from pooled statistics, layer averages from extrema, and BF16-visible parameter motion from FP32 master-weight motion.

Table 5: An internal signal is qualified for a use, not assigned a context-free health label.

Qualification	Required comparison
Validity	Native reference, coverage, ownership and no-op behavior
Information	Explicit baseline and feature additions on unseen outcomes
Localization	Mechanism identity and discriminating probes
Engagement	Intended target changes under the declared action
Consequences	Targeted/no-op/alternative actions, quality and cost
Transport	Paired controls in a new context or horizon

Worked qualification. For second-moment support loss and protected-summary repair at 1,023 post-action updates, validity (V) comes from same-coordinate state checks; information (I) compares detection with the prospective-update guard, which detects the same event; localization (L) uses isolated second-moment damage; engagement (E) measures finite updates; consequences (D) comprise contained transients and lower NLL than reset in fixed-prefix repeats; and transport (T) compares two dense settings, reconstruction shapes and horizons. The MoE record instead follows deprived expert identities, burden removed by balancing and the accompanying NLL contrast (Table 30).

A.1 Common reductions

For a finite vector $x \in \mathbb{R}^n$, define

$$\mathrm{RMS}(x) = \sqrt{n^{-1}\sum_i x_i^2}, \qquad \mu_{|x|} = n^{-1}\sum_i |x_i|,$$
$$q_p(x) = \inf\{z : n^{-1}\sum_i \mathbf{1}[|x_i| \le z] \ge p\}. \tag{5}$$

Here q_p denotes the inverse empirical CDF. Native activation keys use the supports and estimators in Table 6; comparisons preserve both. For a reference x and an altered value x', relative RMS error is $\|x' - x\|_2/\|x\|_2$ on nonzero references; cosine is $\langle x, x'\rangle/(\|x\|_2\|x'\|_2)$ when both norms are positive. Empty, zero-reference and nonfinite cases receive explicit validity flags. Controller-specific denominator floors are specified with each controller.

The subsequent export path takes an unweighted mean across ranks reporting a position `abs_p99` or position `rms` key; it takes MAX for position `abs_max` and `channel_p99`. Thus global-

Table 6: Activation keys and their rank-local reduction semantics. Position denotes a declared hook site. MAX-aggregated local statistics are not pooled distribution estimates; the final rank reducer is specified below.

Key	Object and local estimator	Aggregation contract
Position `abs_p99`	Absolute entries of the local activation tensor; `paddle.quantile` with linear interpolation	Maximum across observations accumulated for that layer; global key is the maximum across active layers.
`channel_p99`	Per-channel maxima over token/batch positions; sorted channel values at zero-based index $\mathrm{round}(0.99(n-1))$	Per-channel maxima are first synchronized across configured tensor/context parallel groups; scalar key is MAX-aggregated.
Position `rms`, `abs_max`	RMS or absolute maximum over the local tensor	MAX-aggregated over collected observations and active layers, not an RMS over their union.

prefixed position p99/RMS keys average rank-local layer maxima. These are not pooled-entry quantiles or pooled RMS, even though each local accumulator uses MAX. A matched comparison must preserve both reduction stages.

Table 7: Representative measurements used in the evidence. Collection scope and backend availability accompany each record.

Family	Measured object and calculation	Interpretation boundary
Attention	Projected Q, K RMS; valid-logit standard deviation; attention probability concentration/entropy	Mask, normalization, head grouping and sampled rows define the comparison.
Activation	$\mu_{\lvert x\rvert}, \mathrm{RMS}(x), q_p(x), \max\lvert x\rvert$; tail support above a declared threshold	A large value or lower quantile is not itself a damage endpoint.
MoE	Per-expert assignments, load CV, router scores/biases; expert gradient and update shares	Preserve expert identities, routed/shared status and window history.
Gradient/update	Gradient norms, parameter norms, realized $\Delta\theta$, relative update and accumulated update energy	Distinguish optimizer proposals from actual parameter changes and master from visible weights.
Residual transport	Gates, residual mixing and actual branch-write RMS	A gate mean is not the magnitude of the resulting branch contribution.
Matrix coordinates	Factor RMS, factor imbalance, selected singular-value diagnostics and composite-map changes	A coordinate-dependent quantity need not indicate a change in the composite function.
Numerical extension	Cast error, positive-support loss, denominator scales and prospective updates	Small aggregate error can hide near-zero denominator damage.

A.2 Attention and activations

For a sampled query/head with valid key set $\mathcal{K}_i$, logits are the implemented scaled dot products, $z_{ij} = q_i^\top k_j/\sqrt{d}$, including any architecture-specific projection or normalization. We compute probability entropy only over valid keys:

$$p_{ij} = \frac{\exp z_{ij}}{\sum_{k\in\mathcal{K}_i}\exp z_{ik}}, \qquad H_i = -\sum_{j\in\mathcal{K}_i} p_{ij}\log p_{ij}. \tag{6}$$

SWA and causal masks change $\mathcal{K}_i$, so entropy comparisons account for valid-key support. Projected Q/K RMS measures vector magnitude; QK normalization is a method that changes the forward computation.

Activation statistics refer to declared sites such as FFN/MoE branch outputs. Tail support can be quantile-defined or relative to a scale such as $\mu_{|x|}$: a quantile fixes the selected fraction, whereas a relative threshold lets that fraction vary. Sampled quantiles can miss sparse extreme values.

A.3 Expert opportunity and parameter motion

Let $A_{\ell e,w}$ count assignments to expert e in layer ℓ during window w, and let $\Delta\theta_{\ell e,t}$ be its measured parameter change. Load CV is the population standard deviation of the expert counts divided by their mean. The opportunity analysis uses the joint shares

$$a_{\ell e,w} = \frac{A_{\ell e,w}}{\sum_j A_{\ell j,w}}, \qquad u_{\ell e,w} = \frac{\sum_{t\in w} \|\Delta\theta_{\ell e,t}\|_2^2}{\sum_j \sum_{t\in w} \|\Delta\theta_{\ell j,t}\|_2^2}. \tag{7}$$

The same-run MoE case measures changes between BF16-visible parameters, converted to FP32 for differencing and accumulation. Its energy therefore includes BF16 rounding effects. The recent Adam reconstruction experiments instead use owned FP32-master changes. Neither quantity is the squared distance between the first and last parameter vectors: cancellation between updates makes that a different statistic. The relative update is $\sqrt{\sum \|\Delta\theta\|_2^2 / \max(\sum \|\theta_{\text{before}}\|_2^2, 10^{-30})}$ for the controller that uses it. A low update-energy share alone can reflect optimizer memory, decay, or quantization; combining it with assignment identity supplies a more specific opportunity test.

A.4 Residual write magnitude and coordinate geometry

For an n_s-stream residual layer (Xie et al., 2025), take $X_\ell \in \mathbb{R}^{n_s\times d}$, column vectors $h_{\text{pre},\ell}, h_{\text{post},\ell} \in \mathbb{R}^{n_s}$ and $H_{\text{res},\ell} \in \mathbb{R}^{n_s\times n_s}$. A per-token schematic form is

$$X_{\ell+1} = H_{\text{res},\ell}^\top X_\ell + h_{\text{post},\ell} F_\ell(h_{\text{pre},\ell}^\top X_\ell). \tag{8}$$

The implementation's stream/batch axes determine the contractions. We compute realized branch write as the complete layer output minus residual-only transport on the same input. Its RMS measures the joint magnitude of the branch output and post-gate, whereas the mean of the four gate entries measures the coefficient alone. Task necessity requires a separate intervention. Mixing-matrix gains and pre-branch activations help locate possible compensation.

For a linear composite $C = AB$, a nonsingular diagonal coordinate change gives $A' = AD$, $B' = D^{-1}B$, preserving C in exact arithmetic. Any intervening operation must also respect or be transformed with this equivalence. Factor norms and singular spectra measure representation; composite-map and same-input output discrepancies measure function. Reciprocal projected-Q/K scaling similarly preserves their product under its forward assumptions. Optimizer response is tested separately because momentum, decay, normalization and finite precision can change under these transformations.

A.5 Numerical support and timing

The Adam support-loss alarm retains up to 128 deterministically evenly spaced samples per valid local tensor slice from the preceding completed boundary. A slice is eligible with at least 32 previously positive samples. For those samples, it computes the fraction now equal to zero; the maximum eligible fraction across slices/ranks triggers at ≥ 0.90. Negative/nonfinite values follow a separate invalid-state path, and summary recovery rejects invalid moments. The reference is the previous completed boundary. Fault validation separately checks the injected cast and resulting damage.

The alarm detects sampled support collapse. It can miss damage outside sampled coordinates, positive-valued shrinkage or subthreshold support loss. One qualifying slice initiates reconstruction of every supported tensor from intact momentum and total. Detection coverage and reconstruction coverage are therefore evaluated separately.

The controller checks this abrupt fault every optimizer update. The generic displayed monitors have independent cadences: in the same-run MoE continuations, the six-family bundle records at 6,689, 6,705, through 13,361, giving 418 observations per family/rank. Extra MoE observations at 7,560 and 10,080 give 420 for that family. Opportunity statistics are accumulated every update and finalized every 128 updates; the global trend is tested at absolute steps divisible by 16, out of phase

with the dashboard. Heavy matrix spectra are collected only at 6,689 in these runs. The abrupt-fault controller and dashboard thus operate at different time resolutions.

A.6 Implementation validation and availability

Validation covers monitor-off versus monitor-on same-input behavior, intended layer/expert coverage, owned parameter slices, logical offsets and exact replay where required. Historical QK/activation and four-stream checks pass 18 of 18 same-input on/off parity cases; the mHC coverage test reaches all 22 intended layers. These tests cover the named collectors and execution paths. Statistics deduplicate shared parameter aliases, exclude tensor padding and distinguish replayed forwards from training observations.

Research collectors extend the reusable monitoring package. Adam-specific state and prospective-update collectors are marked unavailable in the tested MoE Muon cost profile. Model-specific all-on profiles enable only applicable families, including mHC collectors for four-stream residual models. Availability and component identity accompany the scalar measurements.

B Architectures, training settings, and matched budgets

Table 8: Experimental settings for the evaluated models. Repair methods and baselines are compared in matched continuations within each model setting. Dense counts are verified unpadded parameter inventories; the MoE total/active counts are configuration-level estimates.

Setting	Dense GQA	Dense MLA	GQA LatentMoE
Parameters	284,332,544	378,434,816	774.064M total / 350.178M active estimate
Layers / width	22 / 768	22 / 768	12 / 512
Attention	12 Q / 3 KV heads, head width 64	12 heads; Q/KV ranks 768/512; NoPE/RoPE widths 64/64; V width 128	8 Q / 2 KV heads, head width 64
FFN / experts	FFN width 1920	FFN width 1920	Latent/expert width 896; 24 routed, top-8; one shared; first layer dense
Optimizer	AdamW	AdamW	Muon on eligible matrices; Adam support path
Main seed	42	43; additional parent 42	42
Parent step	5,424	6,840	6,688
Standard suffix	1,536 updates	1,024 updates	6,688 updates
Devices / microbatch / accumulation	4 / 16 / 2	4 / 8 / 4	8 / 16 / 1; expert parallel 8

All headline settings use sequence length 4,096, global batch 128, a vocabulary of 201,216, tied embeddings, BF16 forward execution and FP32 master/state where applicable. The nominal learning rate is 6×10^{-4}, weight decay 0.1, and maximum gradient norm 1. Adam betas are $(0.9, 0.95)$. The dense MLA setting enables its configured attention gate/merger. Dense GQA retains a warmup–stable–decay schedule with 543 warmup updates, stable boundary 9,763 and total 10,848; MLA uses 684, 12,312 and 13,680. Continuations preserve these clocks rather than restarting warmup. Original Adam epsilon is 10^{-15}; the common-epsilon experiment changes every continuation, including healthy and damaged controls, to 10^{-8} after restoring its parent.

The MoE setting uses Muon momentum 0.95, scale 0.2, five configured quintic Newton–Schulz steps, and the native QKV/head and FFN splitting. Embedding, bias, output-head and router parameters use the excluded/support path. The router has sigmoid scores, top-8 selection, and zero auxiliary/z-loss coefficients. Its elevated router learning rate and correction action are specified in Appendix E; these are not Adam-state recovery runs.

Table 9: Nominal processed tokens for the headline settings. Every update processes $128 \times 4096 = 524{,}288$ tokens. Replayed or repeated content still counts toward compute, not corpus diversity.

Segment	Updates	Nominal tokens
GQA parent	5,424	2,843,738,112
MLA parent	6,840	3,586,129,920
MoE parent	6,688	3,506,438,144
Standard MLA suffix	1,024	536,870,912
Standard GQA suffix	1,536	805,306,368
Long dense suffix	4,096	2,147,483,648
MoE suffix	6,688	3,506,438,144

The standard GQA recovery suffix contains 512 healthy updates before the fault window. Damage and repair follow safe update 5,937; the endpoint is 6,960. For MLA they follow safe update 6,841, with endpoint 7,864. Each stated 1,024-update recovery window includes the last safe update and 1,023 post-action updates. Long continuations end at parent plus 4,096: 9,520 for GQA and 10,936 for MLA. They introduce no second repair. Full-state rollback experiments have separate endpoints, 10,848 and 13,680, and should not be substituted for the compact-summary suffixes.

Additional cohorts. Activation Tail V1 uses a 22-layer width-768 dense MLA and a 12-layer width-512 MLA-MoE with 24 routed experts/top-8, seed 42, Muon, global batch 128 and sequence length 4,096. Each of six model-stage contexts has four 512-update arms. The mHC functional analysis uses a 22-layer width-768 GQA with four streams, Muon, seeds 42/314159/271828, and checkpoints 300/1,440/2,880. Matrix-coordinate continuations use a separate 12-layer width-512 MLA LatentMoE at checkpoint 300, seed 42, latent width 896 and the same 24/top-8 expert pattern. These are not the 774M GQA MoE recovery model.

Data provenance. Training uses an internal pretraining corpus and its accompanying tokenizer. Within each paired recovery comparison, runs start from the same checkpoint and use the same subsequent training data to control for input differences.

C Quality evaluation and statistical analysis

For sequence j, let N_j be the valid next-token target count and ℓ_j its mean negative log-likelihood. Our token-weighted endpoint is

$$Q = \frac{\sum_j N_j \ell_j}{\sum_j N_j}, \qquad \Delta Q_{A-B} = \frac{\sum_j N_j(\ell_{A,j} - \ell_{B,j})}{\sum_j N_j}. \tag{9}$$

Negative ΔQ favors A. The corrected panel contains 8,192 sequences and 33,554,432 valid targets per role. Matching checks data identifier, token/content identity, physical provenance and valid-token count. Panel 2 contains 8,184 distinct contents among 8,192 records; its eight duplicate-content records are retained. It shares no reconstructed sequence content with panel 1. Exclusion against the entire historical training corpus has not been verified.

C.1 Paired intervals

Unless specified otherwise, we group records by source file, file epoch and part, sort physical offsets, and build blocks of at most 64 adjacent records within each group. There are 129 blocks in panel 1 and 130 in panel 2. We resample blocks 2,000 times with RNG seed 42, retaining token weighting, and report pointwise percentile 95% intervals. Appendix H.4 varies block size for the principal recovery contrasts. The intervals condition on the trained weights and evaluation panel and do not measure training-seed variation. Both panels evaluate the same trained checkpoint.

Blocks preserve local physical adjacency, and the recovery sensitivity analysis increases block size up to 1,024 records. The resampling units are physical record blocks rather than independently identified documents.

The additional MLA parent supplies a second trained prefix. Prefix histories differ, while suffix inputs and evaluation are shared; effects are reported separately by parent.

We use three complementary checks. Same-input collector parity tests whether a collector perturbs its checked computation. Native reload checks that evaluation of a saved checkpoint is reproduced exactly. Matched training continuations align inputs and settings. Healthy armed/no-action trajectories still exhibit nonzero differences, with execution nondeterminism and other run-level effects not separately identified.

C.2 Reader audit and exclusions

A legacy reader produced repeated evaluation content. We exclude endpoint-NLL rankings from the affected cohorts in Table 10. Training telemetry, same-input functional comparisons and mechanism measurements are evaluated separately from these invalid quality endpoints.

Table 10: Excluded quality endpoints and retained mechanism measurements. Counts are records or experimental arms.

Cohort	Audited count	Retained evidence, where independently valid
Metric qualification	150	Training-window prediction and recorded internal signals
Activation Tail V2	198	Actual tail engagement and training trajectories
MoE exact-fork	144 (108 evaluated)	Routing and expert-identity mechanisms
mHC matched-mode	36	Gate response and actual branch writes
Matrix-coordinate probes	9 (6 evaluated)	Same-input function and next-update changes
Data-switch study	11	Gradient/update transients and finite continuation

The numerical-state and MoE recovery tables use corrected evaluations of their named checkpoints. Reader-affected endpoint rankings for activation thresholds, mHC gate amplification, matrix rebalancing and data-state reset are excluded. An earlier large-model SWA-mask error also invalidates its entropy/probability comparisons; mask-independent magnitude measurements are assessed separately.

D Counterfactuals separating a statistic from its function

Table 11: Counterfactual evidence across mechanisms. A: 24 continuations; Δarea is action minus native duration-normalized training-loss area (loss units), with descriptive ranges across six contexts, not intervals. B: four-stream means over three seeds; shallow/deep are layers 1–3/4–21. C: controlled LatentMoE coordinate scaling, diagonal condition $\kappa = 100$ (entries 0.1–10); current and next-update functional KL are distinct probes, not quality endpoints.

A. Tail intervention: changing the statistic need not improve training

Action	Median Δp99	Median Δarea	Min Δarea	Max Δarea
Tail down	-4.15%	+0.000157	+0.000073	+0.000841
Tail up	$+2.55\%$	+0.000076	+0.000004	+0.000304
Random rotation	—	+0.000200	+0.000093	+0.000632

B. Residual transport: post-gate magnitude is not write magnitude

Component	Gate, 300	Gate, 2,880	Write RMS, 300	Write RMS, 2,880
Shallow attention	0.882	0.162	0.323	0.215
Shallow FFN	0.838	0.149	0.258	0.245
Deep attention	0.875	0.649	0.218	0.589
Deep FFN	1.118	0.584	0.342	0.857

C. Coordinate stress: approximately preserved function, changed update

	Native	$\kappa = 100$	Rebalanced
Current-forward KL	0.000000	0.000096	0.000127
Mean next-update KL	0.031082	0.050824	0.031573

D.1 ACTIVATION-TAIL DIRECTION AT FIXED VECTOR NORM

In each model-stage context, the activation assay selects the layer with the largest recorded absolute FFN/MoE-output p99. Selection uses absolute scale rather than p99 divided by mean activation. For a selected width-d activation vector f, the tail contains the largest $k = \min(d-1, \max(1, \lceil 0.01d \rceil))$ absolute entries. Split those entries into f_T and the remaining entries into f_R, with disjoint support. Write $a = \|f_T\|_2$, $b = \|f_R\|_2$, and unit directions $u = f_T/a$, $v = f_R/b$ when both norms are nonzero. The norm-preserving rotation is

$$f(\theta) = (a\cos\theta - b\sin\theta)u + (a\sin\theta + b\cos\theta)v. \tag{10}$$

For a small positive angle in the qualified regime, this transfers squared norm out of the selected tail; reversing the sign transfers norm into it. The requested maximum angle is 0.05 radians, capped separately for each vector:

$$\phi = \operatorname{atan2}(b, a), \qquad \delta = \min\{0.05,\ \max(0, 0.9\min(\phi, \pi/2 - \phi))\}. \tag{11}$$

Tail-down uses $\theta = \delta$ and tail-up uses $\theta = -\delta$ only when $a, b, \|f\|_2 > 10^{-12}$; otherwise the vector is unchanged. The random control rotates by the same δ toward a deterministically seeded unit direction orthogonal to f. The native arm leaves f unchanged. Thus the intervention is vector-wise, not an edit to the global p99 or an exact 1% of the full activation tensor. The identity $\|f(\theta)\|_2^2 = a^2 + b^2$ holds in exact arithmetic; BF16 execution has maximum reported relative norm mismatch 0.259%.

The forward hook returns the transformed tensor to ordinary autograd, not a detached output or straight-through estimator. Gradients flow through the FP32 norms, normalized directions, trigonometric expressions and active branches of the angle cap; discrete top-k indices and validity predicates have no gradient. The random direction is a deterministic sinusoidal candidate projected orthogonally to the current f and normalized, with a paired-coordinate rotation fallback for degeneracy. Its seed depends on experiment seed, optimizer step, layer and hook-call index; it is recomputed per call without consuming the training RNG. The orthogonalization remains differentiable in f. Detached statistics are for logging only. Consequently, this probes the full forward/backward intervention, not solely forward activation magnitude.

There are 24 continuations: two models, three stages and four arms, each 512 updates. The cross-context median p99 response is approximately -4.15% for tail-down and $+2.55\%$ for tail-up. Nevertheless, native training has the lowest normalized training-loss area in all six contexts. The area is the trapezoidal integral over the 512 equally spaced loss observations divided by the integration span. Median increases relative to native are 0.000157 (down), 0.000076 (up), and 0.000200 (random); observed context ranges are $[0.000073, 0.000841]$, $[0.000004, 0.000304]$, and $[0.000093, 0.000632]$, respectively. These ranges are not confidence intervals.

At approximately fixed norm, suppressing the selected tail worsens these continuations. Later thresholds at 4, 6 and 8 times mean absolute activation select approximately 0.478%, 0.0365% and 0.00785% of entries, respectively; their legacy endpoint rankings are excluded because of the reader error. Elevated learning rates produce more extreme tails together with other optimization changes, requiring separate controls for tail-specific effects.

Table 12: Activation-tail contexts and action-minus-native normalized training-loss area, in 10^{-4} units. Layers are zero-based. Each context contains four matched 512-update continuations, seed 42. Positive values indicate higher loss area. Each contrast is a single matched continuation pair.

Context	Checkpoint	Layer	Tail-down	Tail-up	Random
Dense early	300	0	3.143	0.553	6.321
Dense middle	1740	21	0.734	0.047	0.927
Dense late	2880	21	0.785	0.037	1.196
MoE early	300	0	8.410	3.038	6.014
MoE middle	1740	11	1.642	1.386	1.873
MoE late	2880	11	1.507	0.961	2.122

Table 12 identifies the six parent/layer choices frozen before comparing endpoints. Continuations retain the original schedule's 3,450-update duration and 172-update warmup; these clocks are not restarted at the fork.

D.2 GATE VALUES AND ACTUAL BRANCH WRITES

Table 13: Four-stream residual observations from checkpoints 300 to 2,880. Values are means over three seeds and the indicated layer groups: shallow, layers 1–3; deep, layers 4–21.

Component	Early gate	Late gate	Early write RMS	Late write RMS
Shallow attention	0.882113	0.161677	0.323428	0.215318
Shallow FFN	0.838306	0.148553	0.258403	0.244848
Deep attention	0.875	0.649	0.217975	0.588569
Deep FFN	1.118	0.584	0.341970	0.857170

Nine checkpoint/seed replays cover 396 layer-component points on 128 fixed input sequences each. Those points share inputs, seeds and checkpoints. Table 13 separates post-gate amplitude from realized write magnitude. Under external scaling, native gate outputs adapt in the opposite direction: for a $2\times$ arm the effective attention-write ratio moves from about 1.228 to 1.063 and the native gate to about $0.85\times$. Pre-branch input decreases and residual mixing gains remain near 1.00–1.02 in the examined shallow cases. The measured compensation is therefore gate adaptation, rather than amplification of those other quantities. Quality endpoints from the legacy reader are excluded as described in Appendix C.

D.3 EQUIVALENT FORWARD MAPS CAN HAVE DIFFERENT NEXT UPDATES

The identity $AB = (AD)(D^{-1}B)$ illustrates coordinate freedom, but the LatentMoE assay must transform the intervening nonlinear experts too. In row-vector convention, let P, Q be the latent input/output projections and U_e, V_e the read/write weights of routed expert e. For a positive diagonal D, the implemented transformation is

$$P' = PD, \qquad U'_e = D^{-1}U_e, \qquad V'_e = V_eD, \qquad Q' = D^{-1}Q. \tag{12}$$

Applicable latent-input and expert-output biases also scale by D; read biases and final output bias are unchanged. Transforming these quantities jointly preserves expert preactivations and the composite map when routing is unchanged and no intervening normalization breaks the coordinate change. The tested path meets these conditions.

Let $d_{\text{in}}, d_{\text{out}}$ denote the model widths and let r_e, w_e be the read-output and write-input widths of the routed experts. Native coordinate energies are

$$a_i = \frac{1}{d_{\text{in}}} \sum_j P_{ji}^2 + \frac{\sum_e \sum_{j=1}^{w_e} (V_e)_{ji}^2}{\sum_e w_e}, \tag{13}$$

$$b_i = \frac{1}{d_{\text{out}}} \sum_j Q_{ij}^2 + \frac{\sum_e \sum_{j=1}^{r_e} (U_e)_{ij}^2}{\sum_e r_e}. \tag{14}$$

Expert sums and denominators are reduced over the expert group, counting each shared projection once. With a_i, b_i floored at 10^{-30}, rebalancing uses $D_{ii} = (b_i/a_i)^{1/4}$, equating the transformed energies apart from floor and arithmetic effects. The stress diagonal is deterministically shuffled and log-spaced from $\kappa^{-1/2}$ to $\kappa^{1/2}$. Thus $\kappa = 100$ is a diagonal condition ratio with entries from 0.1 to 10. All trajectory arms load checkpoint-300 weights and identically reset optimizer state, LR schedule and trainer step. We evaluate the resulting optimizer response.

From checkpoint 300, native, $\kappa = 100$ gauge and rebalanced-gauge continuations have mean next-update functional KL values 0.031082, 0.050824 and 0.031573, and composite-update relative RMS values 0.004458, 0.013422 and 0.004451. The KL column averages eight update probes over 512 updates (pre-update offsets 0,64,...,448). Each measures $\mathrm{KL}(p_{\text{arm,before}} \| p_{\text{arm,after}})$ relative to the arm's own pre-update distribution. Current-forward KL instead measures $\mathrm{KL}(p_{\text{native replay}} \| p_{\text{initialized arm}})$ after temporarily undoing the coordinate transform for the reference replay.

Both probes replay one captured training microbatch per rank with the same inputs and RNG, selecting up to 32 evenly spaced positions with valid labels. Full-vocabulary KL uses FP32 natural

logarithms, a token mean and an unweighted rank mean over eight ranks; the second accumulation microbatch is excluded. Current-forward KL under the gauge is small but nonzero (0.000096). Rebalancing brings scalar update magnitudes closer to native under controlled stress; update-vector recovery and natural-training benefits were not evaluated by these probes. Natural Matrix Health trajectories remain finite at ratios about 18.1 and at a maximum/minimum incoming coordinate-energy ratio of 559.6. This latter statistic measures coordinate energies, separately from spectra, which were collected only at the first step.

E EXPERT-OPPORTUNITY RECOVERY: COMPLETE PROTOCOL

E.1 CONSTRUCTION AND INTERNAL TRIGGER

The affected MoE parent uses router learning rate 0.01 instead of the healthy 0.0006, with bias correction disabled. This is a training-configuration intervention, distinct from directly overwriting activations or optimizer states. At step 6,688, four arms continue from the same affected parent: no action, internal-triggered balancing, global-triggered balancing, and internal-triggered overall learning-rate reduction. Two further arms continue a separate healthy parent, with the controller disarmed or armed. Each continuation ends at 13,376.

For 24 routed experts and 128-update windows, define

$$b_{\ell e,w} = \mathbf{1}[a_{\ell e,w} < 1/240 \ \wedge \ u_{\ell e,w} < 1/240], \qquad p_{\ell e,w} = b_{\ell e,w-1} b_{\ell e,w}. \tag{15}$$

The threshold is one tenth of equal share, with strict inequalities. The internal trigger is any nonzero p for the same layer/expert in consecutive windows. The parent restores windows 6,401–6,528 and 6,529–6,656 plus 32 partial-window updates. Five identities already meet the condition, so action starts at the restored boundary 6,688, after deprivation has begun.

The reported post-fork burden is $\sum_{\ell,e,w} p_{\ell e,w}$ over 50 adjacent pairs of completely post-fork windows. The first complete window ends at 6,912 and the current window of the first eligible pair ends at 7,040. There are 51 complete windows; the final incomplete 64 updates and inherited trigger windows are excluded. The unit is expert-window exposure.

E.2 GLOBAL BASELINE

The global controller observes training loss, log global gradient norm and log relative update each completed step. It floors logarithm inputs at FP32's smallest positive normal value. Checks require absolute step > 925 (669 warmup plus 256 history updates) and divisibility by 16. It fits each channel on the previous 256 observations, excluding the current one. For $i = -256, \ldots, -1$, the design row is $(1, i/256)$. Three weighted least-squares iterations start with unit weights. Each iteration computes residuals, scale

$$s = \max\{1.4826 \operatorname{median} |r - \operatorname{median}(r)|, s_{\min}\}, \quad w_i = \min\{1, 1.345 s / \max(|r_i|, s_{\min})\}. \tag{16}$$

The final signed score is (current value minus fitted intercept)/scale. Calibration uses the healthy prefix through 6,688. For channel y, its floor is the maximum of the 0.01 quantile of positive $|y_t - y_{t-1}|$ and $\epsilon_{\mathrm{FP32}} \max(\operatorname{median} |y|, 1)$. Frozen floors for loss, log-gradient and log-update/weight are 0.0016861891746520995, 0.0008965283059714069 and 0.00017280808930753606. A check is positive when any score exceeds 3; two consecutive positive checks trigger action, even if different channels cross. A normal check resets the streak. Nonfinite inputs separately force a trigger. Threshold calibration counts an event after two positives and rearms after two normal checks. It chose the first threshold from $\{3, 4, 5, 6, 8, 12\}$ with at most one event per 1,000 covered updates; threshold 3 had one event across 5,776 updates. Runtime restores the affected parent's own rolling history and streak.

The affected branch's loss score is 3.2594 at step 11,328, followed by update/weight score 3.5517 at 11,344. The latter completes the streak and activates balancing. This controller uses loss, gradient and update history; abrupt Adam-state corruption is evaluated against the separate prospective-update guard.

Table 14: Expert-opportunity recovery from the same affected parent. Burden sums deprived experts over adjacent complete post-fork window pairs. Lower NLL indicates better prediction quality.

Action	Final deprived experts	Persistent burden	NLL
No action	9	401	1.205670
Early targeted balancing	0	0	1.207295
Later global-rule balancing	0	236	1.209668
Early learning-rate reduction	9	442	1.237810

E.3 Actions and results

Balancing enables the existing per-step bias rule (Wang et al., 2024)

$$b_{\ell e} \leftarrow b_{\ell e} + \gamma \, \mathrm{sign}(\bar{A}_\ell - A_{\ell e}), \qquad \gamma = 0.001, \tag{17}$$

on all 11 routed layers, using expert usage aggregated across the applicable training groups. The callback runs every optimizer end and then clears its counters. Scores plus FP32 biases select experts; selected output weights use the original unbiased sigmoid scores before native normalization/scaling. The action switches the correction rate from zero to 0.001 and leaves it active. The alternative action multiplies the scheduler's base/current LR by 0.06 at step 6,688, from 0.0006 to 0.000036. The router multiplier remains $16\frac{2}{3}$, so its effective LR falls from 0.01 to 0.0006 at that boundary. Subsequent decay follows the unchanged schedule clock with the reduced base. Both router and non-router learning rates change.

Table 15: MoE quality contrasts, action minus reference. Brackets are conditional paired-panel 95% intervals. Mechanism endpoints are in Table 14.

Comparison	ΔNLL	Lower	Upper
Early balance – no action	+0.001625	+0.000882	+0.002305
Late balance – no action	+0.003998	+0.003140	+0.004841
LR reduction – no action	+0.032140	+0.027750	+0.036003
Early – late balance	-0.002373	-0.002840	-0.001907
Healthy armed – healthy no-op	-0.001558	-0.002557	-0.000738

Early balancing eliminates measured burden, late balancing reduces it to 236 versus 401 without action, and overall LR reduction leaves 442. Both healthy continuations have zero interventions and zero burden. Table 15 reports quality contrasts alongside healthy armed/no-op variation. The early-balancing endpoint reproduces all 8,192 sample NLLs on native reload.

F Numerical-state diagnosis and full-state recovery

F.1 A small aggregate error with a large update consequence

The diagnostic fault rounds valid Adam second-moment entries through FP16 and back to native FP32 storage after a safe update. In one paired seed-42 MLA timeline, the recorded zero fraction is 0.999363, relative RMS error 0.005824, and cosine 0.999983. At step 6,841 both losses equal 2.56517029. At 6,842 the fault/reference losses are still 2.64240527/2.64244199, while recorded update RMS is 11,022.47 versus 0.000195625. The next forward becomes nonfinite. A small normwise storage error thus precedes a large optimizer error. The normwise error and zero fraction describe different aspects of denominator damage; relative error in $S = \sum_i v_i$ is a separate quantity. This diagnostic uses its recorded update-measurement pipeline, while the recovery experiments measure owned FP32-master changes.

One reason for the amplification is the nonlinear denominator. For the next Adam recurrence, write $r_i = \sqrt{v_i^+/C_2}$ and $u_i = m_i^+/[C_1(r_i+\epsilon)]$, using the next-state and clock definitions in Equation 28. Holding the next gradient, first moment and clocks fixed,

$$\frac{\partial u_i}{\partial v_i} = -\frac{\beta_2 m_i^+}{2C_1C_2 r_i (r_i+\epsilon)^2}, \qquad r_i > 0. \tag{18}$$

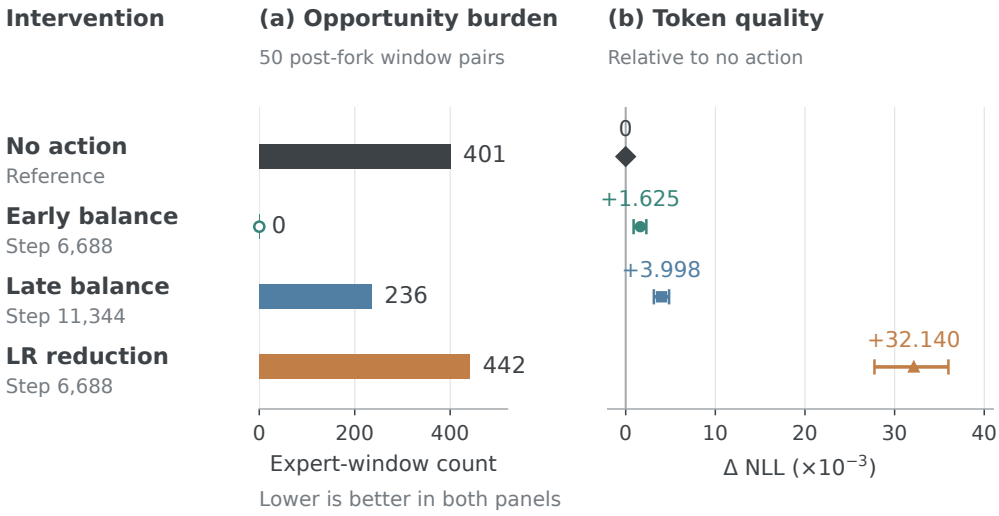


Figure 6: Expert opportunity and token quality. Earlier balancing prevents post-fork opportunity burden; later application of the same correction leaves accumulated burden. Both corrected continuations have higher observed endpoint NLL than no action. Each point corresponds to one continuation.

This sensitivity explains why average state error need not track update error. Zero-point hazards are also discussed in low-bit optimizer work (Li et al., 2023).

F.2 Protection opportunity and full rollback

Both the internal support alarm and the prospective-update guard act before the first dangerous parameter application. With equal full-state rollback/replay permission, both recover finite continuation. The global guard inspects the proposed update before applying it. Offline decimation of this timeline leaves a pre-dangerous-apply opportunity in one of k phases at interval k, illustrating how sparse monitoring can miss an abrupt event.

The guard's score is

$$G = \sqrt{\frac{\sum_{i\in\mathcal{C}}(\eta_{\text{guard}} d_i)^2}{\max(\sum_{i\in\mathcal{C}}\theta_i^2, 10^{-30})}}, \tag{19}$$

where $\mathcal{C}$ contains collector-covered parameter entries, η_{guard} is the common prospective scheduler LR, and d_i is the bias-corrected prospective Adam direction including applicable decay. The score uses the common LR without parameter-specific multipliers. It checks every update before application and triggers for $G > 0.004$ or a nonfinite proposal. The collector unscales available matching-size master/ordinary gradients and reads moments, stored beta powers, epsilon and visible weights. It operates before native gradient clipping, unlike the summary's post-clip interception; its coverage is logged. Both alarms receive the same full-state rollback and data-replay permission in Table 16.

Both alarms support recovery with full-state rollback/replay. Healthy armed controls exhibit small differences from healthy no-action continuations, contextualizing endpoint contrasts. Restoring FP32 storage stops further rounding; recovering values already lost requires reconstruction or rollback.

Table 16: Corrected-panel full-state recovery. These endpoints differ from the compact-summary experiments. Deltas are recovery minus healthy; intervals are conditional on the trained states.

Setting/action	NLL	ΔNLL	95% interval
MLA healthy	1.186995	–	–
MLA internal recovery	1.186936	-0.000060	$[-0.000574, 0.000454]$
MLA global recovery	1.186852	-0.000143	$[-0.000613, 0.000316]$
GQA healthy	1.291452	–	–
GQA internal recovery	1.291011	-0.000440	$[-0.000762, -0.000130]$
GQA global recovery	1.290281	-0.001171	$[-0.001570, -0.000812]$

F.3 Useful negative controls

In a separate 1,024-update MLA comparison, healthy/full-reset/full-reset-with-rewarm NLLs are 1.348198, 1.350841 and 1.353003. These earlier full resets zero both moments and *restart bias-correction clocks* by setting the stored powers to β_1, β_2, unlike the clock-preserving R comparator used for protected-summary recovery. The action follows the qualified support-loss boundary. Rewarm multiplies the original prospective scheduled LR at the jth subsequent apply by $[1-\cos(\pi j/150)]/2$ for $j = 1, \ldots, 150$, then by 1; it does not restart the trainer's LR schedule. Rewarm reduces an early transient yet worsens the panel endpoint relative to full reset by 0.002161 with interval $[0.001309, 0.003163]$. This local reset/rewarm control is not a full reproduction of SPAM (Huang et al., 2025). Another large-epsilon configuration has NLL 1.277193 without "restoration" versus 1.346148 after reducing epsilon, a difference of 0.068955 $[0.065593, 0.072504]$. That configuration did not establish a quality fault worth repairing. A deviation from the nominal setting is therefore insufficient grounds for intervention.

G Protected-summary recovery: derivation and implementation

G.1 The moment constraint and its assumptions

For a coordinate of ordinary Adam, let the consumed post-clip gradient be g_t and

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t, \qquad v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2. \tag{20}$$

Assume constant $0 \le \beta_1 < 1$, $0 < \beta_2 < 1$ and $r = \beta_1^2/\beta_2 < 1$. The following algebra holds for arbitrary consumed gradients. The tight zero-initialized bound is a known Adam result (Ma & Yarats, 2019, Appendix F); its derivation specifies the history assumptions used in reconstruction.

Proposition 1 (Moment feasibility under a known Adam bound)**.** *For $m_0 = v_0 = 0$ and $t \ge 1$,*

$$v_t \ge c_t m_t^2 \ge c_\infty m_t^2, \qquad c_t = \frac{c_\infty}{1-r^t}, \quad c_\infty = \frac{(1-\beta_2)(1-r)}{(1-\beta_1)^2}. \tag{21}$$

The weaker inequality $v \ge c_\infty m^2$ is preserved by subsequent same-beta recurrences whenever it holds at the starting state.

Proof. Expand the zero-initialized recurrence and apply weighted Cauchy–Schwarz:

$$m_t^2 = \left[\sum_{j=0}^{t-1} (1-\beta_1)\beta_1^j g_{t-j}\right]^2 \tag{22}$$

$$\le \left[\frac{(1-\beta_1)^2}{1-\beta_2}\sum_{j=0}^{t-1} r^j\right]\left[\sum_{j=0}^{t-1}(1-\beta_2)\beta_2^j g_{t-j}^2\right] = \frac{v_t}{c_t}. \tag{23}$$

Equality is attained by gradients proportional to $(\beta_1/\beta_2)^j$ over this finite history (with the natural zero-beta convention). For a general feasible starting state,

$$(\beta_1 m + (1-\beta_1) g)^2 \le \left[\frac{\beta_1^2}{\beta_2 c_\infty} + \frac{(1-\beta_1)^2}{1-\beta_2}\right][\beta_2 c_\infty m^2 + (1-\beta_2) g^2] \tag{24}$$

$$\le [\beta_2 v + (1-\beta_2) g^2]/c_\infty, \tag{25}$$

because the first bracket equals $1/c_\infty$. This proves invariance. □

At the experimental betas, $r = 81/95$, $c_\infty = 14/19$, and $c_t = (14/19)/[1 - (81/95)^t]$. For example, $c_1 = 5$, $c_2 = 475/176$ and $c_{16} \approx 0.799192$. We use the conservative analytic constant c_∞. A faithful checkpoint retains its accumulated moment history; resetting moments while retaining mature bias clocks separates moment age from clock age. Feasibility is checked from the state, since a nonzero m with zero v violates the invariant.

G.2 Protected total and feasible set

Initialize $S_0 = \sum_i v_{0,i}$ from trusted storage. The independent recurrence in Equation 2 equals the intact tensor sum in exact arithmetic. Distributed accumulation uses disjoint valid owned slices whose union is the logical tensor; replicated copies and padding are excluded. After storage damage, the recurrence retains the intact total along the actual consumed gradient stream. If bad parameter updates have already changed that stream, it is no longer the total of a counterfactual undamaged-model trajectory.

At a completed boundary set $L_i = c_\infty m_i^2$ and $B = S - \sum_i L_i$. The relaxed feasible set is

$$\mathcal{F}(m, S) = \{x \in \mathbb{R}^n : x_i \geq L_i,\ \sum_i x_i = S\}. \tag{26}$$

It is nonempty exactly when $B \geq 0$. It contains the intact states under the stated assumptions, but is not a characterization of all realizable finite clipped-gradient histories. For $B > 0$, write $x = L + Bp$ and choose the permutation-symmetric distribution

$$p^* = \arg \max_{p_i \geq 0, \sum_i p_i = 1} -\sum_i p_i \log p_i = \mathbf{1}/n. \tag{27}$$

This is maximum entropy of the *normalized residual mass*, not of x/S. Maximizing the latter would generally yield $x_i = \max(L_i, \tau)$ for a suitable τ, a different operation. For $L = (0, 1)$ and $S = 3$, uniform residual gives $(1, 2)$ while full-state entropy is maximized by $(1.5, 1.5)$. At $B = 0$, the reconstruction is L; at $S = 0$, feasibility requires $m = 0$.

The residual $v - c_\infty m^2$ is algebraic: the EMAs use different weights, so it is distinct from gradient variance. Reconstruction allocates this residual independently of the damaged tensor's coordinates. Uniform residual and $\max(L_i, \tau)$ both satisfy the feasible set and subsequent update bound; only the former is evaluated here. The equal-information mean S/n can violate a coordinate lower bound, yet yields favorable MLA quality. Feasibility supplies a constraint on reconstruction, while quality distinguishes its allocations empirically.

G.3 Adaptive-update bound

Let m^+, v^+ be the outputs of the next recurrence in Equation 20. For stored beta powers $0 \leq p_j < 1$ at that apply, define $C_j = 1 - p_j > 0$. With effective parameter LR $\eta \geq 0$, $\epsilon > 0$ and applicable decay λ (zero on decay-excluded parameters), the evaluated update convention is

$$u_i = \frac{m_i^+/C_1}{\sqrt{v_i^+/C_2} + \epsilon}, \qquad \Delta\theta_i = -\eta u_i - \eta\lambda\theta_i. \tag{28}$$

If $v_i^+ \geq c_\infty (m_i^+)^2$, then for positive v_i^+,

$$|u_i| \leq \frac{\sqrt{C_2}}{C_1\sqrt{c_\infty}} \frac{\sqrt{v_i^+}}{\sqrt{v_i^+} + \epsilon\sqrt{C_2}} \leq K = \frac{\sqrt{C_2}}{C_1\sqrt{c_\infty}}. \tag{29}$$

If $v_i^+ = 0$, feasibility forces $m_i^+ = 0$ and positive native epsilon gives $u_i = 0$. For a tensor with common LR/decay, $\|\Delta\theta\|_2 \leq \eta\sqrt{n}K + \eta|\lambda|\|\theta\|_2$. Mature clocks give $K \approx \sqrt{19/14} = 1.164965$ for the unscaled adaptive coordinate. This bound is in exact arithmetic; FP32 rounding introduces an additional residual.

The optimizer stores beta powers for the next apply. Immediately after completed update t, they correspond to β_j^{t+1} in exact arithmetic. A probe of the just-completed state and a proposed next update must therefore use different clock positions. Epsilon in Equation 28 is outside the bias-corrected square root; changing to uncorrected moments scales that epsilon by $\sqrt{C_2}$. Omitting this distinction can invalidate a purportedly exact response reconstruction.

G.4 Executable procedure and safeguards

The implemented scalar recovery path proceeds as follows:

1. Restore trusted weights, master weights, moments, clocks, RNG, data cursor and schedule. Resolve each logical tensor's valid owned slices; reject overlaps, gaps or unsupported ownership. Initialize its total from trusted v.
2. On each ordinary Adam apply, observe the actual dense gradient after aggregation, unscaling and clipping. Accumulate squared gradient only over owned valid coordinates, then complete the independent total recurrence once per tensor and step.
3. At the completed boundary, apply the declared fault in its experimental arm and evaluate the support alarm. The summary is excluded from the cast. One qualifying slice starts synchronous reconstruction across the entire supported tensor inventory, not just the sampled or alarming coordinates.
4. Form candidates from retained m, S. Require a completed boundary, finite nonnegative total, finite m, L, finite nonnegative damaged v, and $\sum_i L_i \le S(1 + 10^{-5})$. Collectively reject detected invalid inputs before modifying any tensor. The consistency test does not authenticate the trusted inputs.
5. Set $B = \max(0, S - \sum_i L_i)$ and write $L + B/n$ to valid coordinates, leaving padding, weights, first moments and clocks intact. Rebase the summary; on success disable further casts and continue ordinary training with no further repair. Rejection raises an exception that stops this execution path before the next apply, with no implicit reset or rollback fallback.

Within-tolerance negative residuals can produce a total slightly above S; exact conservation is an algebraic property, while native agreement is numerical. Collective precommit validation is not crash-atomic distributed commit. Save/reload validation checks the repaired native endpoint; the helper does not implement serialization of independently protected summaries for recovery from a damaged checkpoint after process restart.

Native tests cover active clipping, same-gradient no-op behavior, damage isolation, state preservation, invalid input rejection, ownership slices and rebasing. Before-damage summary/native relative error is qualified below 10^{-5}; formal execution retains coverage checks without paying for a full intact-v comparison every step. These tests do not establish collateral safety for localized faults, where inventory-wide reconstruction can replace intact second moments as well as damaged ones. Supported assumptions exclude sparse/lazy updates, arbitrary changing beta schedules, AMSGrad, shared corruption of gradients and summary, and unqualified skipped-update lifecycles.

G.5 Information, shape, and exploratory extensions

Table 17 separates recovery strategy from reconstruction shape. Clock-preserving reset clears both moments, so comparison with summary recovery tests retention of m, S jointly. The moment floor retains m, clocks and surviving damaged coordinates, using $v_i \leftarrow \max\{v_i^{\mathrm{dam}}, c_\infty m_i^2\}$ without a protected total. This control preserves momentum while changing both scale information and reconstruction. Mean versus constrained holds retained information fixed and tests coordinate allocation.

Richer variants retain row/column totals. For $S > 0$ and $F_{ij} = R_i C_j / S$, the floor variant $\max(F, L)$ generally increases mass, whereas $L + (B/S)F$ preserves total but not each original marginal. At a feasible $S = 0$ boundary the reconstructed state is zero. The rank-one construction has an established Adafactor precedent (Shazeer & Stern, 2018).

An exploratory extension replaces uniform residual with a product of the most recently consumed squared-gradient marginals. For a matrix, define $R_i^g = \sum_j g_{ij}^2$, $C_j^g = \sum_i g_{ij}^2$, $G = \sum_{ij} g_{ij}^2$, and,

Table 17: Information retained or used to construct post-action moments. Live repairs require intact weights and act before the next apply. Validation also inspects damaged storage. Full rollback restores saved state and replays training.

Action	Retained information	Reconstructed or reset state
Clock-preserving reset	Weights and clocks	Clears m and v
Moment-floor repair	Weights, clocks, m and surviving damaged v; no protected S	$v_i \leftarrow \max\{v_i^{\mathrm{dam}}, c_\infty m_i^2\}$
Protected-total mean	Weights, clocks, m and S	$v_i = S/n$
Protected-total constrained	Same as mean	$v_i = c_\infty m_i^2 + (S - c_\infty \|m\|_2^2)/n$
Full-state rollback	Saved weights, moments, clocks, schedule, RNG and data position	Restores a previous state and replays

when $G > 0$, $q_{ij} = (R_i^g/G)(C_j^g/G)$ and $\widetilde{v} = L + Bq$. Zero-gradient tensors and vectors retain uniform residual allocation. This requires additional stored information and is not the scalar-only method in the main table. The raw squared-gradient prior failed the reference checks on all 96 development snapshots. A factored lagged prior reduced local response error from about 0.5455 to 0.4114 on 24 correlated adjacent snapshot pairs; eight native-execution and eight exact-reload tests passed. The measured endpoint is local response error.

H Recovery comparisons across information, epsilon and horizon

Tables 2–3 summarize the recovery comparisons. This appendix supplies both quality panels, execution-level contrasts, interval sensitivity and longer horizons. C denotes constrained reconstruction, M the same-total mean, and R clock-preserving reset of both moments. C/M versus R tests recovery with total and momentum retained; C versus M isolates shape; floor and clipping compare lower-information strategies in original-epsilon MLA. Each contrast preserves its parent, future inputs, endpoint and panel. Brackets are the paired-panel 95% intervals defined in Appendix C; effects are action minus reference.

Table 18: Recovery across settings, quality panels and update transients. A: original epsilon; $E_{16} = \sum_{j=0}^{15} \|\Delta\theta_j\|_2^2$ sums FP32 master-update energy over the last safe update ($j = 0$) and 15 post-action updates. B: a separate common-epsilon comparison. All repairs stop casts; reset clears both moments but retains clocks; mean and constrained retain the same total and momentum. Paired uncertainty and longer horizons are in Fig. 5.

	GQA 284M			MLA 378M		
Continuation	NLL 1	NLL 2	E_{16}	NLL 1	NLL 2	E_{16}
A. Original epsilon 10^{-15}; two fixed quality panels						
Healthy	1.439940	1.449600	70.49	1.348442	1.368485	92.42
Reset moments	1.441770	1.451396	217.46	1.349869	1.371214	281.40
Protected-total mean	1.440863	1.450543	40.63	1.347205	1.368380	53.79
Protected-total constrained	1.439646	1.449905	36.21	1.347240	1.368347	50.91
B. Common epsilon 10^{-8}; one fixed quality panel						
	GQA NLL		E_{16}	MLA NLL		E_{16}
Healthy	1.438345		68.19	1.345590		90.69
Stop cast, no reconstruction	1.611137		131692.42	1.472737		98857.57
Reset moments	1.440749		181.60	1.349620		236.36
Protected-total mean	1.439971		40.46	1.344943		53.44
Protected-total constrained	1.439350		36.17	1.347387		50.57

H.1 A control that retains momentum without the protected total

The original-epsilon MLA comparison in Table 19 uses seed 43, parent 6,840, endpoint 7,864, the same cached continuation inputs and the same corrected evaluation panel. Both actions occur

after safe update 6,841 and preserve first moments and clocks. Constrained recovery improves NLL over the moment-floor control by 0.003190, with C–floor interval $[-0.004131, -0.002296]$. This comparison favors summary-based reconstruction over a momentum-preserving alternative; the control definitions are given in Appendix G.

Table 19: Matched MLA recovery at original $\epsilon = 10^{-15}$, retaining momentum and clocks in both arms. Energy sums squared updates over the first reported 16-update window. The comparison uses one early parent and the first quality panel.

Action	NLL	Actual-update energy sum
Moment floor, no protected total	1.350431	474.090014
Protected-total constrained	1.347240	50.912982

H.2 Shape and horizon comparisons

Table 20: Early recovery at original $\epsilon = 10^{-15}$. Both panels evaluate the same retained states.

Setting	Contrast	Panel 1	Panel 2
GQA	C–R	$-0.002124\ [-0.002871, -0.001465]$	$-0.001491\ [-0.002705, -0.000451]$
GQA	C–M	$-0.001217\ [-0.002076, -0.000488]$	$-0.000638\ [-0.001268, -0.000020]$
MLA	C–R	$-0.002629\ [-0.003596, -0.001664]$	$-0.002866\ [-0.004265, -0.001780]$
MLA	C–M	$+0.000036\ [-0.000524, 0.000607]$	$-0.000032\ [-0.000817, 0.000722]$

For the additional seed-42 MLA parent, M–R is $-0.002973\ [-0.004515, -0.001680]$, C–R is $-0.003189\ [-0.004508, -0.002112]$, and C–M is $-0.000216\ [-0.000892, 0.000412]$. Both summary strategies improve on reset, while the difference between shapes is unresolved.

Table 21: Early actual-update energy under original epsilon. Entries sum squared owned FP32-master changes over the first reported 16-update window. The healthy row anchors the transient.

Action	GQA	MLA
Healthy	70.485121	92.416363
Clock-preserving reset	217.455646	281.397269
Protected-total mean	40.629104	53.786496
Protected-total constrained	36.209782	50.912982

Both summary rules beat reset in the original matched early common-epsilon comparisons. Constrained reconstruction remains worse than healthy in these two executions, and the C–M ranking reverses between settings. The retained total fixes scale while leaving coordinate allocation underdetermined; which state features explain the preferred allocation remains open.

H.3 Repeated executions at fixed training prefixes

We repeat the healthy, reset, mean and constrained continuations twice for GQA and once for MLA. Each repetition restores the same respective weights, moments, clocks, RNG state and data cursor, then consumes the same future inputs. The replication unit is a new execution at a fixed pretraining prefix. All twelve new continuations complete their specified horizons, and endpoint reloads reproduce their per-record evaluation results. Table 2C gives endpoint NLLs; Table 23 gives every original and repeated paired effect.

GQA uses seed 42's step-5,424 parent and continues to step 6,960, with damage and immediate action after safe update 5,937. MLA uses seed 43's step-6,840 parent, acts after safe update 6,841 and continues to step 7,864. Thus both have 1,023 strictly post-action updates; GQA additionally includes 512 healthy maintenance updates before the last safe update. All roles switch to $\epsilon = 10^{-8}$ on restoration. On four GPUs, global batch is 128 and sequence length 4,096; per-GPU microbatch/accumulation are 16/2 for GQA and 8/4 for MLA. LR is 6×10^{-4} under the retained

Table 22: Common-epsilon contrasts corresponding to Table 18B. All roles change to 10^{-8} after restoring parents trained with 10^{-15}.

Contrast	GQA	MLA
M–R	-0.000778 $[-0.001394, -0.000138]$	-0.004677 $[-0.005860, -0.003664]$
C–R	-0.001399 $[-0.002110, -0.000704]$	-0.002232 $[-0.003357, -0.001246]$
C–M	-0.000621 $[-0.001029, -0.000214]$	$+0.002445$ $[0.001874, 0.003070]$
C–healthy	$+0.001005$ $[0.000402, 0.001763]$	$+0.001798$ $[0.000740, 0.002900]$

schedule, betas are $(0.9, 0.95)$, weight decay is 0.1 and global gradient clipping is 1.0. BF16 forward computation uses FP32 master weights and optimizer states. The fault is one FP32–FP16–FP32 second-moment cast, with subsequent casts disabled. Each endpoint uses the same 8,192-record, 33,554,432-target quality panel.

Table 23: All fixed-prefix execution contrasts at common $\epsilon = 10^{-8}$, in 10^{-3} token NLL. M is mean reconstruction, C constrained reconstruction and R moment reset. Negative favors the first action. Original rows correspond to Table 2A; repeats to part C. Rows within each setting share the pretraining parent and future data.

Setting / execution	M–R	C–R	C–M
GQA / original	-0.778	-1.399	-0.621
GQA / repeat 1	-0.675	-2.295	-1.620
GQA / repeat 2	-0.940	-1.515	-0.576
MLA / original	-4.677	-2.232	+2.445
MLA / repeat 1	-3.547	-2.721	+0.826

Both summary strategies outperform reset in all five execution blocks. The shape ranking also repeats: constrained is better in the three GQA executions, whereas mean is better in both MLA executions. Healthy controls vary across executions, and recovered quality can remain above or below the corresponding healthy loss.

H.4 Dependence sensitivity and persistent-clipping comparison

We assess the original six common-epsilon contrasts and two original-epsilon MLA clipping contrasts at five predeclared block sizes, $B \in \{64, 128, 256, 512, 1024\}$. These are the original endpoint weights, not the repeated executions above. The estimand is the token-weighted paired NLL difference. Blocks contain at most B adjacent evaluated records within a physical file/epoch/part, with the shorter tail retained at each boundary. We resample blocks with replacement for 2,000 draws and report the 2.5th and 97.5th percentiles. Physical slice positions are ordered numerically. This probes sensitivity to local dependence assumptions; physical files are not assumed to be independent text sources.

Table 24: Dependence sensitivity: all 40 pointwise 95% paired-block intervals, in 10^{-3} token NLL. The six GQA/MLA rows use the original common-epsilon executions; clipping rows use original-epsilon MLA and the two panels of the same endpoint weights. C/M/R are as in Table 23. No intervals pool epsilon settings, repeated executions or panels.

Contrast	$B = 64$	$B = 128$	$B = 256$	$B = 512$	$B = 1024$
GQA M–R	$[-1.394, -0.138]$	$[-1.485, +0.024]$	$[-1.607, +0.193]$	$[-1.756, +0.227]$	$[-1.682, +0.183]$
GQA C–R	$[-2.110, -0.704]$	$[-2.233, -0.528]$	$[-2.398, -0.352]$	$[-2.531, -0.361]$	$[-2.322, -0.488]$
GQA C–M	$[-1.029, -0.214]$	$[-1.115, -0.141]$	$[-1.095, -0.137]$	$[-1.101, -0.151]$	$[-1.045, -0.192]$
MLA M–R	$[-5.860, -3.664]$	$[-6.055, -3.500]$	$[-6.322, -3.339]$	$[-6.546, -3.195]$	$[-7.074, -2.997]$
MLA C–R	$[-3.357, -1.246]$	$[-3.617, -0.984]$	$[-3.965, -0.702]$	$[-3.882, -0.582]$	$[-4.014, -0.752]$
MLA C–M	$[+1.874, +3.070]$	$[+1.709, +3.237]$	$[+1.557, +3.487]$	$[+1.452, +3.587]$	$[+1.286, +3.820]$
C–Clip, panel 1	$[-3.014, -1.222]$	$[-3.326, -0.974]$	$[-3.629, -0.769]$	$[-3.812, -0.595]$	$[-3.848, -0.623]$
C–Clip, panel 2	$[-2.585, -0.410]$	$[-2.583, -0.331]$	$[-2.568, -0.435]$	$[-2.618, -0.327]$	$[-3.015, -0.183]$

Constrained-minus-reset remains below zero at every tested block size in both settings, as does MLA mean-minus-reset. GQA mean-minus-reset includes zero at all sizes above 64. The C–M ordering remains opposite between settings. Panel 1 has 129/65/34/18/9 blocks and panel 2 has 130/67/35/19/11; the coarsest analyses therefore have few resampling units. These are sensitivity analyses of the fixed-weight, pointwise intervals defined in Appendix C.

Table 25: Constrained recovery versus persistent adaptive-update clipping, MLA at original $\epsilon = 10^{-15}$. Both panels evaluate the same endpoint weights. Differences and 64-record-block intervals are in token NLL, without the 10^{-3} scaling used in Table 24.

Panel	Constrained	Clipping	C–Clip	95% interval
Panel 1	1.347240	1.349373	-0.002133	$[-0.003014, -0.001222]$
Panel 2	1.368347	1.369731	-0.001383	$[-0.002585, -0.000410]$

Both actions follow safe update 6,841 from the same seed-43 MLA parent and continue to step 7,864. Persistent clipping bounds every valid adaptive-update coordinate for all 1,023 post-action updates; constrained recovery reconstructs second moments once. Both panel differences favor constrained recovery at every tested block size. The second panel evaluates the same trained weights at the next 8,192-record offset, retaining its 8,184 unique input/target pairs within the original panel estimand (Appendix C).

H.5 Longer-horizon persistence

Table 26: Persistence after 4,096 updates from the original parent, $\epsilon = 10^{-15}$, no new action. These extend Table 18A, not its common-epsilon panel B.

Setting	Panel	M–R	C–R	C–M
GQA	1	-0.000268	-0.000277	-0.000009
		$[-0.001544, 0.000894]$	$[-0.001090, 0.000470]$	$[-0.000648, 0.000609]$
GQA	2	-0.000831	-0.000301	$+0.000530$
		$[-0.002118, 0.000322]$	$[-0.001722, 0.001089]$	$[-0.000401, 0.001729]$
MLA	1	-0.002259	$+0.000755$	$+0.003014$
		$[-0.003111, -0.001435]$	$[-0.000064, 0.001626]$	$[0.002127, 0.004015]$
MLA	2	-0.001092	-0.000392	$+0.000700$
		$[-0.001987, -0.000105]$	$[-0.001272, 0.000537]$	$[-0.000410, 0.001755]$

Mean recovery retains an MLA advantage over reset on both panels. GQA's early separation is unresolved later, and constrained MLA reconstruction is worse than mean on panel 1. At the later endpoints, C–healthy is also adverse for GQA panel 2 ($+0.001491$ $[0.000752, 0.002342]$) and MLA panel 1 ($+0.001478$ $[0.000402, 0.002591]$).

Testing attenuation directly. To test whether the constrained-versus-reset effect changes over time, we compare the paired four-state contrast

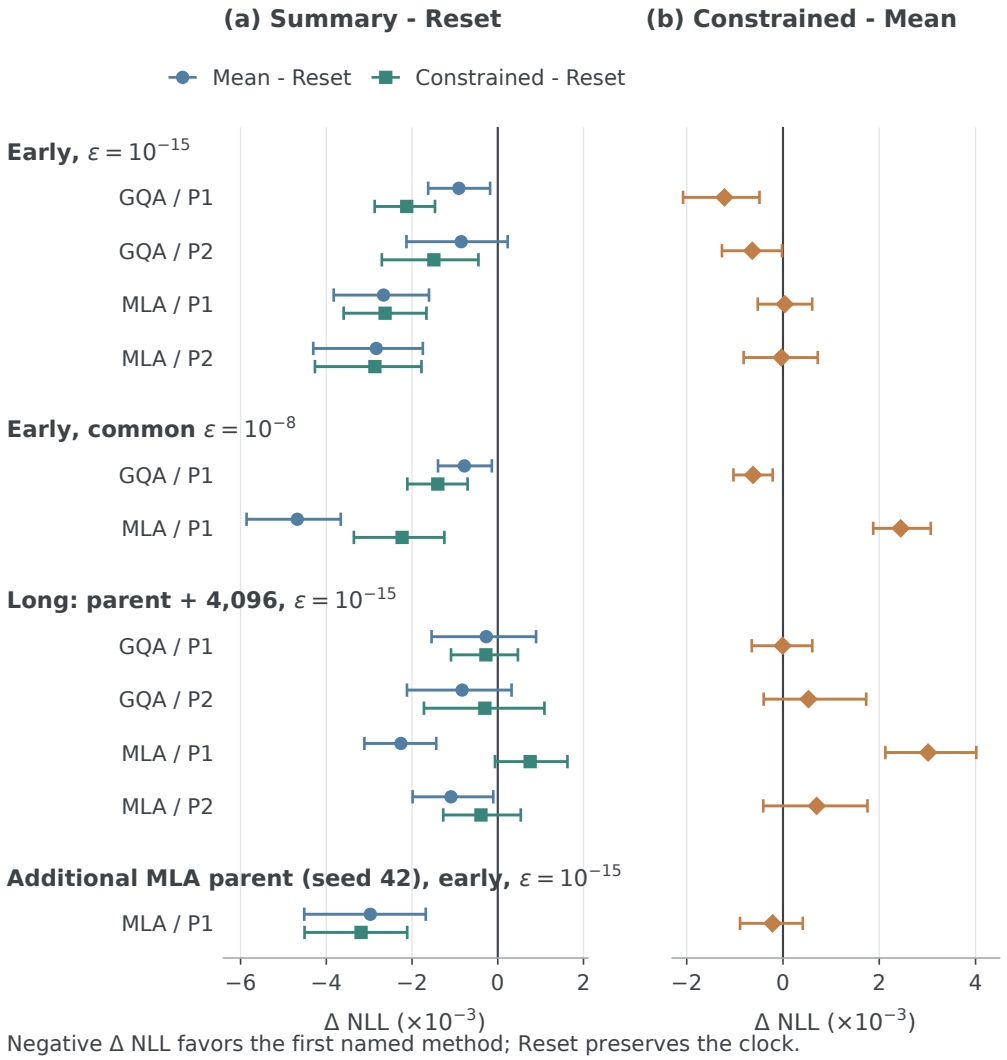
$$A_{C-R} = (Q_C - Q_R)_{\text{late}} - (Q_C - Q_R)_{\text{early}}. \tag{30}$$

Positive values mean a loss of the constrained strategy's advantage over reset. In original-epsilon GQA, $A_{C-R} = +0.001847$ on panel 1, with interval $[+0.001150, +0.002524]$; panel 2 gives $+0.001190$ with interval $[-0.000448, +0.003328]$. These exploratory contrasts find attenuation on panel 1 while leaving its magnitude unresolved on panel 2.

Earlier matched MLA development controls give panel NLLs 1.349373 for update clipping, 1.350431 for the moment floor, and 1.349713 for clipping followed by the floor; the last does not resolve an advantage over clipping. Rich-summary additive/factor variants reach 1.347841/1.347966 in their own matched setting. These exploratory results are reported within their respective settings and motivated the scalar-only comparisons, matched controls, additional parent and longer-horizon tests.

The clipping comparator uses the same original-epsilon MLA parent and endpoint. Each valid owned adaptive coordinate u_i in Equation 28 is clipped to $[-K, K]$, with the native pre-apply clock

and K from Equation 29; weight decay is not clipped. Native moments and clocks still advance. Persistent clipping checks all 1,023 post-alarm updates. The clipping-plus-floor arm instead clips only the first post-alarm update, then floors v_i at $(14/19)m_i^2$ after its weight update and synchronization, preserving the just-written weights and disabling subsequent clipping. This differs from the immediate moment-floor arm, which acts at the preceding safe boundary.



Negative Δ NLL favors the first named method; Reset preserves the clock.
Early: 1,024-update window includes last safe update.
Common ε: 10^{-8} after restoring 10^{-15} parents; a separate branch.
Long: 4,096 updates from parent, not from repair; no new action.
Primary parents: GQA seed 42, MLA seed 43. P1/P2 are evaluation panels.
95% fixed-panel intervals, not training-seed uncertainty.

Figure 7: Recovery across settings and horizons. Early summary strategies improve over reset in the common-epsilon comparisons. Relative performance of mean and constrained reconstruction changes with the setting and evaluation horizon.

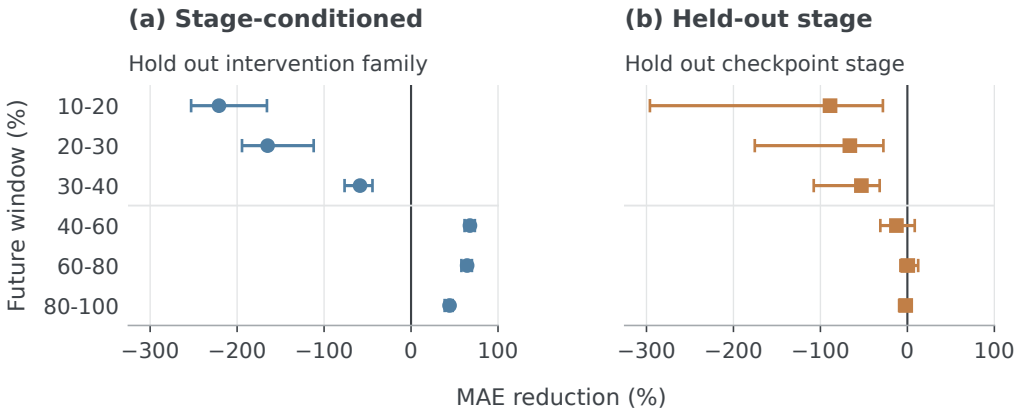


Figure 8: Prediction across horizons and stages. Positive MAE reduction favors internal responses added to context, arm code and early paired loss. The target is future treatment-minus-native mean loss; windows are continuation fractions. Intervals resample six checkpoint cells from three seed lineages.

I Incremental prediction and local response analysis

I.1 Predictive qualification

Let B be the declared baseline features and M the added measurement block, all measured before the target window. For matched held-out splits and a specified fitting protocol, define empirical predictive gain

$$I_M(h) = 1 - \frac{\mathcal{R}(\widehat{f}_{B,M}, Y_h)}{\mathcal{R}(\widehat{f}_B, Y_h)}, \tag{31}$$

when the denominator is positive. This quantity measures relative predictive-risk reduction. Attribution to module-local information requires the same global features in both fits. The fitting protocol specifies preprocessing, feature selection and the grouping of related trajectories.

The two assays below use different fitting protocols. The first fits transformations on outer-training data before inner regularization selection. The second nests its prediction pipeline but fixes an outcome-free coverage schema using the full cohort. These preprocessing choices are specified with each assay.

The historical core pool contains 325 runs within a larger, overlapping 507-sample archive, not 325 independent contexts. Its baseline combines configuration/context metadata with early-loss mean, slope and change; it is not the stronger prospective-update guard used in the local assay below. Using the first 20% of training, gradient/update features reduce final-50-step mean training-loss MAE from 0.048533 to 0.041157, a 15.2% relative reduction. The absolute MAE improvement is 0.007376 with group-bootstrap interval $[0.002407, 0.012789]$. For final-window standard deviation, MAE changes from 0.001808 to 0.001443, an absolute gain 0.000365 $[0.000187, 0.000556]$. This result uses five outer folds grouped by configuration context, fitting preprocessing/PCA only within the outer training fold; 4,000 outer-group bootstrap draws estimate absolute improvement. Whole-curve AUC and maximum-jump targets include the observed prefix and do not establish unseen-future prediction.

Historical learner and feature specification. The 325 configurations all use seed 42 and start from scratch: 210 layout and 115 intervention runs. Their 81 groups use the literal tuple (setting, duration, Full-attention count, campaign, design family, layout/context name); deterministic five-fold GroupKFold keeps these groups intact. Targets are mean and sample standard deviation of the final 50 logged losses. Point MAE pools held-out rows; the bootstrap resamples the 81 groups with multiplicity and recomputes the paired absolute-error difference without refitting.

The numeric baseline consists of early-loss mean, slope and change; setting index, duration, Full-attention count, planned steps and layer count; layout count and normalized-index mean, standard deviation, first/last index and gap standard deviation; intervention dose and changed value. Categorical inputs are campaign, design family, layout/context, intervention branch and changed key. Numeric inputs use median imputation and standardization; categorical inputs use one-hot encoding with unknown categories ignored. This baseline excludes global gradient norm. The augmented block includes both `grad_update/` signals and `training/global_grad_norm` signals; its gain cannot be attributed exclusively to module-local measurements beyond a strong global baseline.

For each signal the prefix ends at $\max(3, \lfloor 0.2T \rfloor)$; features are the finite-observation mean, OLS slope against step/T, and $(x_{\text{last}} - x_{\text{first}})/(|x_{\text{first}}| + |x_{\text{last}}| + 10^{-12})$, requiring three observations. Layered collectors contribute layer mean, maximum, population standard deviation and normalized-depth slope, plus Full/SWA means where applicable. Direct global update keys are retained. Monitor observations at steps zero/one, nonfinite-fraction fields, timing and size fields are excluded; loss and global-gradient records are collected before that step filter. Internal columns must occur in at least 70% of outer-training rows. They are median-imputed, standardized, clipped to $[-10, 10]$, and constant columns removed before PCA with at most eight components, also bounded by feature count and training rows minus one. PCA uses the library's automatic solver and random state 20260724; it applies only to this block.

Both fits use Ridge with $\alpha \in \{0.01, 0.1, 1, 10, 100\}$, selected separately by pooled inner out-of-fold MAE with up to four GroupKFold splits and first-grid tie breaking. Feature transformations are fitted once per outer-training fold, not separately inside each inner split; outer test rows are excluded. There is no target standardization.

The prospective MoE cohort has 72 action rows: four intervention families (attention geometry, activation tail, update history and router balancing), three seeds (42, 314159, 271828), two starting checkpoints (300, 1,440) and three arms per family/context. Each arm has horizon 512 and a matched native baseline. The observation prefix is relative updates 1–51, the floor of 10% of the horizon; the available MoE observations in that prefix occur at 1, 17, 33 and 49. For relative progress p, the target is a paired future response,

$$Y_{(a,b]} = \text{mean}_{a<p\le b}(\ell_{\text{treatment,p}} - \ell_{\text{native,p}}), \tag{32}$$

not absolute future loss. The baseline includes architecture/stage indicators, signed arm code, its magnitude, a random-control indicator and early paired loss difference; internal features are paired prefix responses. Four held-out-family folds give stage-conditioned MAE reductions of -220.85%, -164.92% and -58.67% for the early future 10–20%, 20–30% and 30–40% windows: internal features worsen prediction there. Reductions become 67.7%, 64.5% and 44.3% for the 40–60%, 60–80% and 80–100% windows. The corresponding late context-clustered intervals are $[61.5, 73.3]\%$, $[57.9, 70.0]\%$ and $[38.6, 47.9]\%$. Holding out training stage instead gives approximately -12.6%, $+0.5\%$ and -1.9% in those late windows, with intervals crossing zero. This predictor knows the declared arm code, unlike the measurement-based local inversion below. An informative within-stage measurement does not automatically transport across stages or horizons. Reader-affected held-out-NLL targets are excluded.

Holding out seed instead of stage gives late-window reductions of 46.93%, 51.80% and 53.89%, with context-clustered intervals $[6.61, 58.22]\%$, $[26.82, 62.01]\%$ and $[31.92, 60.73]\%$, respectively. These three folds reanalyze the same 72-row cohort. The contrast with stage holdout shows that predictive transfer depends on which aspect of context changes.

MoE learner, schema and lineage. Each seed's two checkpoints belong to one trained trajectory: six checkpoint cells span three seed lineages. The reported 2,000-draw intervals (RNG seed 20260805) concatenate the 12 actions from each resampled cell and recompute relative MAE gain from fixed out-of-fold predictions. They condition on this cohort and omit cross-stage dependence within a lineage; the whole-lineage sensitivity below examines that dependence.

The six baseline columns are a constant MoE-setting indicator, checkpoint-1,440 indicator, signed arm code d, $|d|$, the indicator $d = 0$, and mean early treatment-minus-native loss. Here $d = -1, +1, 0$ codes the two directional arms and random control, not physical dose magnitude: flatten/sharpen/random attention, tail-down/tail-up/random tail, reverse/native/random update direction, or correction-off/correction-2×/shuffled correction. It contains no family or seed one-hot encoding, loss slope or prospective-update guard.

Whole-lineage sensitivity of saved predictions. Keeping both checkpoints together, family-held-out relative MAE reductions range across the three seeds from 62.85–74.33%, 60.44–70.55%, and 41.67–48.92% in the three late windows. Omitting each entire lineage from scoring gives ranges of 64.65–70.42%, 61.38–66.64%, and 41.99–45.68%. The corresponding stage-held-out omission ranges are -18.31 to -2.53%, -3.80 to 2.84%, and -3.65 to 1.19%. Thus the observed contrast between family and stage transport is not driven solely by one lineage's scoring weight. These ranges reaggregate fixed predictions under lineage-specific scoring; the original fits and full-cohort schema remain unchanged. This sensitivity analysis covers family and stage holdouts.

The augmented input adds 38 prefix responses. For each rank-zero global key, compute $2(\overline{x}_a - \overline{x}_0)/(|\overline{x}_a| + |\overline{x}_0| + 10^{-12})$ from finite prefix means, not a mean of stepwise relative differences. The keys are assignment-load and gate-mass CV, normalized entropy, KL to uniform, maximum/minimum fractions and max/min ratio (six each); bias/affinity Jaccard; expert-bias and expert-output-norm maximum, mean, minimum and standard deviation (four each); probability entropy and router entropy; router-input absolute maximum, absolute p99 and RMS; router-logit absolute maximum, mean and standard deviation; router-margin mean, minimum, p01 and p10; score-sum maximum, mean and minimum; shared-expert norm; and shared/routed ratio. Keys are sorted lexicographically after at least 80% coverage and nonconstant-range screening on the full 72-row cohort, before the outer split. This schema screen uses coverage and variability, without targets.

The predictor pipeline median-imputes with missingness indicators, standardizes, clips to $[-10, 10]$, applies full-SVD PCA to the entire input with $\min(16, \text{raw feature count})$ components, and fits Ridge. Thus baseline/augmented inputs have six/44 raw columns and six/16 components. The same five-alpha grid is selected by GridSearchCV's mean inner-fold negative MAE, with preprocessing refitted inside each inner split. Outer LeaveOneGroupOut uses family, seed or checkpoint (four, three or two folds). Corresponding inner groups are (setting, seed, checkpoint), (family, setting, checkpoint), or (family, setting, seed), using four GroupKFold splits. Neither PCA dimension nor alpha is chosen on outer-test targets.

Table 27: Absolute errors for late-window MoE prediction. Baseline and augmented MAEs are in 10^{-3} paired training-loss units; gain and its interval are percentages of relative MAE reduction. Both predictors are refit under each holdout. Intervals resample six seed/checkpoint cells from three lineages; rows reuse the same 72 actions.

Held out	Future window	Baseline MAE	+ Internal MAE	Gain [95% interval]
Family	40–60%	2.139620	0.690723	67.72 [61.54, 73.34]
Family	60–80%	2.764009	0.981357	64.50 [57.91, 69.99]
Family	80–100%	2.228460	1.241886	44.27 [38.63, 47.88]
Seed	40–60%	1.016165	0.539264	46.93 [6.61, 58.22]
Seed	60–80%	1.217723	0.586982	51.80 [26.82, 62.01]
Seed	80–100%	1.346531	0.620833	53.89 [31.92, 60.73]
Stage	40–60%	1.152055	1.297210	−12.60 [−31.11, 8.45]
Stage	60–80%	1.416891	1.410273	0.47 [−8.00, 12.42]
Stage	80–100%	1.542913	1.572013	−1.89 [−9.49, 3.43]

I.2 Local response assay

From each of two MLA parents, we collect ten 512-update forks: two center repeats, four axis perturbations and four withheld combinations. The coordinates are $m \leftarrow e^a m$ and $v \leftarrow e^{2b} v$, with axis doses $\pm 1/32$ chosen for local derivative estimation. The prediction target is future update energy over updates 2–16.

The parents are seed 42 and seed 43 at step 6,840, each continued with restored full state and RNG to 7,352. Centers are $(0, 0)$; calibration axes are $(\pm h, 0), (0, \pm h)$; withheld combinations are all four $(\pm h, \pm h)$, $h = 1/32$. Optional groups are activation, QK and sampled optimizer state, with budgets of one, two or three groups added to global/guard inputs. At the headline budget one, every selector chooses state: seven scalar inputs including the anchor, versus three for state-only. Group-count equality does not imply equal scalar dimension or runtime cost.

Table 28: Predictive augmentation and its transport boundary. A: first 20% of 325 runs, five context-grouped outer folds; context plus early loss is augmented with gradient/update features including global gradient norm. B: first 10% of 72 MoE action rows, baseline includes context, arm code and early paired loss; each holdout refits both predictors. Positive relative MAE reduction favors the augmentation, not necessarily module-local information beyond a strong global baseline. B's columns reuse the same cohort; intervals and adverse early windows are in Fig. 8.

A. Predicting final-window training outcomes			
Target	Baseline MAE	+ Grad/update MAE	Reduction
Mean loss, final 50 steps	0.048533	0.041157	15.2%
Loss standard deviation	0.001808	0.001443	20.2%
B. Future paired-response MAE reduction under different holdouts			
Future window	Family held out	Seed held out	Stage held out
40–60%	+67.7%	+46.9%	−12.6%
60–80%	+64.5%	+51.8%	+0.5%
80–100%	+44.3%	+53.9%	−1.9%

The headline assay uses information available at the first update to predict strictly subsequent energy at updates 2–16. Global inputs are first-batch loss, global gradient norm and LR; the guard is Equation 19; sampled state inputs are pre-apply momentum RMS, second-moment mean and zero fraction. Optional internal groups use forward QK/activation observations, excluding post-update measurements from this cutoff. The state-only comparator uses those sampled state features alone, not a full-state oracle or global-plus-state set.

Write $z = (a, b)$ and $h = 1/32$. Let $\overline{M} = (M^{(1)} + M^{(2)})/2$ average the two center repeats, and let J and K_{raw} be measurement and outcome responses from central differences on the axes. For the selected features define $\delta\widetilde{M} = W(M - \overline{M})$. The local coordinate estimate and predicted consequence change are

$$\widehat{z} = ((WJ)^\top WJ + \rho I)^{-1}(WJ)^\top \delta\widetilde{M}, \qquad \widehat{\delta Y} = K_{\mathrm{raw}}\widehat{z}. \tag{33}$$

The default replay scale, standardized center difference and weighting are

$$s_i = \max\{|M_i^{(1)} - M_i^{(2)}|,\ \epsilon_{\mathrm{FP32}} \max(|\overline{M}_i|, h \max_j |J_{ij}|)\}, \tag{34}$$

$$r_i = (M_i^{(1)} - M_i^{(2)})/s_i, \qquad W = [(1-\alpha)I + \alpha rr^\top]^{-1/2}\,\mathrm{diag}(1/s), \quad \alpha = 0.1. \tag{35}$$

Identically zero rows use $s_i = 1$. The rank-one weighting is recomputed for each selected subset. Regularization is $\rho = 0.01\lambda_{\max}((W_{\mathrm{all}}J)^\top W_{\mathrm{all}}J)$, replacing a zero spectral scale by 1, and remains fixed across subsets. For center consequence Y_0, the diagnostic uses $K = K_{\mathrm{raw}}/\max(|Y_0|, h\max_j |(K_{\mathrm{raw}})_j|, \epsilon_{\mathrm{tiny}})$ and examines

$$\sup_{z \neq 0} \frac{\|Kz\|_2^2}{\|WJz\|_2^2 + \rho\|z\|_2^2}. \tag{36}$$

This classical quadratic quotient measures poorly observed local directions under the specified coordinates, scales and regularization. The weighting uses the two center repeats as a local replay scale. Group budgets count measurement families; collector runtime is measured separately.

Greedy consequence selection adds groups to a fixed global/guard anchor to minimize this worst-direction quotient. Variance, correlation and axis-only leave-one-out alternatives use the same candidate groups and group budgets. None fits or selects using the four withheld combinations. Sensitivity checks use regularization strengths $0.001, 0.01, 0.1$, shrinkage $0, 0.1$, and a scale-only alternative. The scale-only alternative gives selected-feature MAE about 0.1948, compared with the default 0.2610.

The known-dose comparator receives the perturbation coordinates that the measurement-based model attempts to infer. It tests how much prediction error comes from inferring those coordinates. Simple same-budget feature selection matches the more elaborate selector, and frozen transfer fails. The 15,008 prediction records are repeated measurements from the described forks. The assay evaluates local prediction, rather than online action selection or terminal quality.

Frozen transfer reuses the earlier centering and response fit together. Its failure can therefore reflect a shifted baseline, a changed response map, or both.

Table 29: Local response controls at a newly calibrated MLA parent. The target is $\sum_{j=2}^{16}\sum_i(\Delta\theta_{j,i})^2$ over valid owned FP32-master entries. First-update inputs predict four withheld combinations. Known-dose receives the perturbation coordinates; frozen transfer reuses an earlier calibration.

Information / predictor	MAE
Global signals plus strong guard	2.456880
Optimizer-state measurements	0.262499
Selected / all-on measured features	0.260955
Known intervention coordinates, Taylor predictor	0.073141
Frozen state predictor from previous parent	approximately 25.4

I.3 Why update motion need not mean new learning

For an optimizer response $U(g;s)$ at fixed state, the exact decomposition

$$U(g;s) = U(0;s) + [U(g;s) - U(0;s)] \tag{37}$$

separates a zero-current-gradient baseline from the current-gradient change. Squared norms include their cross term, and decay contributes to $U(0;s)$ when present in U. Native checks cover 384 Muon and 128 Adam snapshots. This decomposition interprets actual expert motion; incremental predictive value over gradient/momentum features was not established by those probes.

J Fault coverage and informative non-successes

Table 30: Signal–action comparisons across the study: measurements, mechanism responses and intervention outcomes.

Case	Measurement and comparison	Observed outcome
Activation tails	Tail changes at nearly fixed norm	Native training has the lowest loss area in all six contexts.
Gates / coordinates	Coefficients, writes and next-update KL separated	Gate and write magnitudes evolve differently; rebalancing brings next-update KL magnitude closer to native.
Expert opportunity	Persistent identity and targeted exposure recovery	Burden removed; higher observed NLL; timing tested on an affected parent.
State corruption	Support alarm before dangerous apply	Both support and global guards enable full-state recovery.
Summary recovery	Intact-state contract and matched restoration	Lower early NLL than reset; reconstruction rankings vary with setting and horizon.
Prediction / response	Incremental future-window or local information	Locally calibrated prediction improves; cross-stage/parent transfer fails.

Table 31 groups the case studies by training-configuration changes, controlled mechanism injuries and observed deviations. Each row summarizes the resulting trajectory and evaluated actions.

The data-switch cohort uses one MLA parent with seven continuations in 11 executions. Switching to code, Chinese or mathematics inputs produces first-16-window gradient multipliers approximately 4.51, 9.93 and 2.07, and first-update multipliers 1.35, 1.29 and 1.15. All inspected 2,048-update continuations remain finite. Quality endpoints are excluded because of the legacy reader error (Appendix C).

Table 31: Representative fault constructions, observed consequences and evaluated actions.

Construction / assay	Observed consequence	Observed intervention outcome
FP16 second-moment round-trip	Support loss, dangerous next update; small normwise error	Full rollback works; compact summaries improve over reset in specified contexts.
Elevated router LR, no balancing	Repeatedly underserved expert identities	Bias correction recovers opportunity with higher observed NLL; tested overall LR reduction fails that target.
Large training LR	Activation tails and degraded training can co-occur	Tail-intervention outcomes are reported separately in Table 12.
BF16 router computation	About 1–1.5% top-k changes in tested probes	Measured endpoint: routing assignments.
Attention output/gradient mantissa reduction	About 2.37% relative RMS error in one tested output probe	Measured endpoint: attention output error.
Increased second-moment decay	No onset in the tested $\beta_2 = 0.999$ continuation	No fault-triggered action in this continuation.
Data-source switch	Gradient/update transients, finite 2,048-step continuations	Finite continuations; endpoint quality excluded after reader audit.
Reciprocal matrix-coordinate stress	Present function nearly preserved, next-update KL changes	Rebalancing returns next-update KL magnitude toward native.
Post-gate scaling	Branch response and compensatory gate adaptation	Scaling elicits compensatory gate adaptation.

Projected QK normalization provides a known positive control: ten paired training comparisons improve last-window loss, with median improvement about 0.018975. Prior work introduced normalization variants and their stability benefits (Henry et al., 2020; Wortsman et al., 2024; Zhai et al., 2023).

Longer navigation studies compare fixed, global-signal and internal-signal policies. In one equal-cost comparison, they encounter 6/9/4 eventful segments covering 6/3/3 root mechanisms. The segment counts share trajectories, and context-history controls explain much of the apparent early lead. Five 4,096-update continuations in a later stage contain no qualifying event; four in another stage contain two eventful trajectories tied to the same expert.

K Overhead, reproducibility scope, and limitations

K.1 End-to-end monitor cost

The 160-update profiles discard the first 32 updates and summarize relative updates 33–96 and 97–160. Fifteen distinct configurations fill 16 logical slots because one selected MLA configuration aliases its state bundle. Successive windows are not independent reruns. Dense profiles use four GPUs with microbatch 32 and accumulation 1, unlike the main recovery experiments. MoE uses eight GPUs, microbatch 16, accumulation 1 and expert parallelism 8. All use global batch 128, sequence 4,096, BF16/O2, stage-1 sharding and tensor/pipeline/context parallel degrees 1.

MLA's global and selected/all-on profiles use different trainer revisions with the same runtime libraries; its ratio includes that revision difference. GQA and MoE use a common trainer revision within each setting. Costs are reported separately for the three profiling configurations.

Dense selected enables the prospective guard and sampled optimizer state, without generic monitor families. Dense all-on adds attention, activation, gradient/update, matrix-health and numerical-health collectors. MoE all-on additionally enables MoE health, but its requested Adam guard/state are unavailable under Muon. The MoE selected request enables no generic family and neither requested collector is available; its timing is retained only as an unsupported-request observation. Generic families, when enabled, use interval 16; scalar training logging uses interval 1.

Table 32: Measured step time in the profiling configurations; percentages are relative to the same row's global baseline. All-on means the available model-specific bundle. †MoE selected is an unsupported request, not functioning internal monitoring. MLA's cross-revision ratio is descriptive.

Setting	Global s/step	Selected s/step	Change	All-on s/step	Change
MLA	1.005469	1.122180	11.61%	1.808711	79.89%
GQA	0.933070	1.070461	14.72%	2.000812	114.43%
MoE	1.183367	1.304461†	10.23%†	1.432508	21.05%

These profiles motivate separate cadences for expensive spectral/function probes and lightweight operational counters. Earlier Matrix Health self-time is 0.12–0.19%, while Table 32 measures end-to-end step time.

K.2 Protected-total maintenance

For each dense setting, four serial 64-update runs use OFF/ON/ON/OFF order, with 47 stable measured intervals per run and two repeats per condition on four B300 GPUs. OFF retains attention, activation, gradient/update and numerical health in both settings, plus matrix health in GQA, at interval 16. Batch decompositions match the main recovery settings: GQA 16/2 and MLA 8/4, compared with 32/1 in the earlier profiles. Table 33 reports repeat ranges around this detector-enabled baseline; Table 32 uses its separate global-monitor baseline.

GQA has 178 and MLA 266 logical summary scalars; four-rank replicated payload totals are 2,848 and 4,256 bytes. The exploratory gradient-prior extension requires substantially more retained information: 1,754,468 bytes per rank for GQA and 2,559,844 for MLA before other buffers.

Table 33: Incremental summary maintenance. Payload is one FP32 scalar per logical tensor per replica, excluding temporary arrays, metadata, intact native state and communication.

Setting	OFF mean [range], s/step	ON mean [range], s/step	Overhead	Payload
GQA	2.441681 [2.400787,2.482574]	2.538447 [2.528404,2.548489]	3.96%	712 B
MLA	2.807904 [2.764383,2.851426]	2.932234 [2.802809,3.061660]	4.43%	1,064 B

The support detector remains enabled in both maintenance conditions. Separately recorded in-place repairs have maximum-rank alarm-to-action-record intervals of 64.749 ms (GQA) and 99.260 ms (MLA), including validation, synchronization and event logging. Alarm-to-next-completed-update intervals are 17.784 s and 23.213 s, including diagnostics and the next forward/backward/update. Each is one observed event, not an isolated reconstruction kernel cost or a latency distribution. Full-state recovery request-to-first-replayed-update latencies are 117.262/119.424 seconds for internal/global MLA and 113.417/123.802 for GQA, one observation each. Different boundaries and execution paths preclude inferring a speedup from these measurements.

K.3 Conditional operating regime

The demonstrated use case is a live process whose second-moment storage path has introduced a lossy cast while weights, momentum, clocks and the independent total remain intact. The process can stop that cast and reject the pending update. The experiment deliberately creates this condition; its frequency in routine training is unknown. A crash that loses the in-memory total, shared corruption of gradient inputs, or continued faulty storage violates this contract. Compact recovery is consequently a candidate complement to checkpoints, not their replacement.

A cost comparison must first require acceptable outcomes for both strategies. Let T be useful training updates, N the number of faults satisfying the contract, δ_s, δ_b the additional steady-state seconds per update for summary protection and a comparator, and R_s, R_b their per-event recovery costs, including any replay. Let K_s, K_b cover periodic persistence cost not included in those terms. For equal acceptable outcomes, summary protection saves time when

$$T(\delta_s - \delta_b) < N(R_b - R_s) + K_b - K_s. \tag{38}$$

If both systems retain the same ordinary checkpoints, their common persistence cost cancels. A reset comparator may have negligible recovery latency but different transient and quality consequences; a full rollback preserves more information but incurs reload and replay.

Measurements cover retained payload, incremental maintenance and observed action intervals; end-to-end protection cost remains to be evaluated in a minimal deployment configuration.

K.4 Limitations

We evaluate PRISM through matched studies on research-scale models and controlled faults. Larger-scale deployments and additional fault regimes are directions for future validation. The internal corpus and tokenizer constrain exact external replication; Appendix C details evaluation and statistical assumptions.

Protected-summary recovery addresses second-moment corruption before a destructive parameter update, assuming intact momentum, weights, clocks, gradient inputs and an independently maintained total. It offers compact approximate recovery alongside checkpointing, with outcomes depending on the training context and horizon. Prospective tests on new contexts would assess how a fixed PRISM qualification procedure supports intervention selection.