UNDERSTANDING AND CONTROLLING ADAMW LOSS SPIKES

Anonymous authors
Paper under double-blind review

ABSTRACT

Loss spikes can disrupt AdamW training despite gradient clipping. Mismatched first- and second-moment time scales can let momentum outpace its normalization, amplifying the adaptive update ($\widehat{u}_t = \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon)$). We study how its magnitude and allocation affect immediate loss. Across three GPT-2 355M runs, at the unit threshold $|\widehat{u}_{t,i}| = |\widehat{m}_{t,i}/(\sqrt{\widehat{v}_{t,i}} + \epsilon)| > 1$, the median fraction of coordinates newly crossing the threshold is 0.571% at 19 captured loss-spike onset states versus 0.0033% at nine ordinary states (over 170×). Globally shrinking the adaptive update lowers immediate loss at all 19 states, while selectively limiting new tail entrants improves further at equal norm. Tensor-relative history helps allocate attenuation, and replay favors sustained over onset-only control. These findings motivate Dual-Scale Update Control, a lightweight yet effective combination of coordinate clipping and a tensor-relative growth constraint, with negligible optimizer-state overhead and no extra forward or backward passes. Across three GPT-2 774M runs per configuration, it reduces post-warmup spike episodes from 17 to zero and improves mean validation loss from 2.916 to 2.856. Half-rate and global norm-matched controls retain spikes despite similar validation quality. Dual control also achieves zero post-warmup spikes in the tested larger NanoGPT and Qwen3 models, supporting transfer across architectures.

1 INTRODUCTION

Loss spikes can threaten stable large language model (LLM) training: they may degrade final convergence and, in severe cases, lead to non-finite loss or divergence. Reports across model settings document such instability (Chowdhery et al., 2023; Zeng et al., 2023). With the mechanism unclear, recovery has sometimes involved rolling back to an earlier checkpoint and skipping nearby batches (Chowdhery et al., 2023). This repeats computation and leaves samples unused. Understanding which optimizer updates cause damage offers a route to preventing such disruptions.

Gradient clipping limits the input to AdamW. The optimizer then averages gradients and their squares on different time scales and normalizes one moment by the other (Kingma and Ba, 2015; Loshchilov and Hutter, 2019). The resulting parameter displacement can differ substantially from the clipped gradient. Update clipping already addresses this distinction (Shazeer and Stern, 2018), and nearby or matched moment coefficients are established alternatives (Ma et al., 2022; Orvieto and Gower, 2025). The sharp coordinate bound of Peng et al. (2025) quantifies the available amplitude: its asymptotic value is 7.27 at $(\beta_1 = 0.9, \beta_2 = 0.999)$ and 1.0 for matched moments ($\beta_1 = \beta_2$). This bound leaves two empirical questions open: which updates training realizes, and how those updates affect loss.

We investigate these questions through observations and interventions from identical training states. At the unit threshold $|\widehat{u}_{t,i}| > 1$ for AdamW's bias-corrected adaptive update ($\widehat{u}_t = \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon)$), the median fraction of newly entering coordinates is 0.571% at 19 captured loss-spike onset states versus 0.0033% at nine ordinary states across three GPT-2 355M runs (over 170×). Uniformly scaling every coordinate of the native adaptive update to one quarter lowers immediate loss at all 19 spike-onset states. At 16 of them, this reduced update also beats omitting the adaptive update entirely. At equal adaptive norm, limiting new tail entrants at threshold 0.45 improves on uniform shrinkage at all 19. Thus both the size of the step and its allocation across parameters affect damage. Blockwise and fixed-budget interventions then show that tensor-relative update history helps choose where to

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

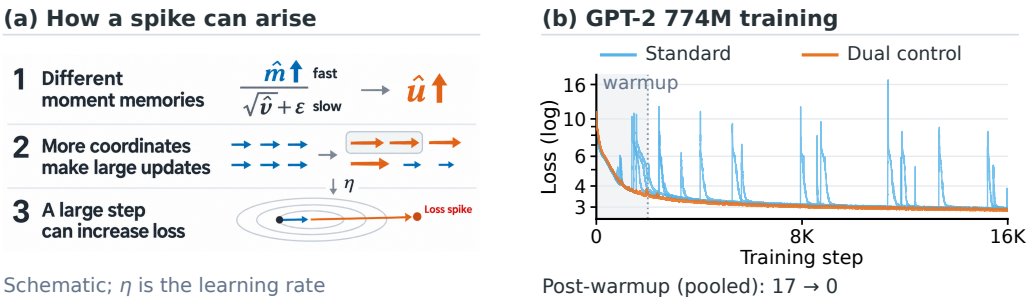


Figure 1: **Mechanism and mitigation.** (a) Three conceptual steps. (1) With different moment memories, momentum can grow faster than its normalization scale. (2) More coordinates enter the large-update tail, with tensor-local growth. (3) The learning rate η turns the update into parameter displacement. The blue small step enters a lower-loss region; enlarging it can reach higher loss (orange). Direction and allocation also affect damage. The contours are illustrative, and moment mismatch alone does not imply a spike. (b) Raw 774M training loss (log axis), with pooled three-seed post-warmup counts. Table 2 reports damage and validation separately.

mitigate. Finally, sustained clipping outperforms an identical first-step intervention, showing why a control must also address later updates.

These findings motivate Dual-Scale Update Control. A coordinate cap limits large outputs; a second constraint limits a tensor's mean absolute update relative to its recent history. The rule acts on AdamW's output while retaining the native moment recurrences and base learning-rate schedule. The local experiments motivate these design principles; complete training runs evaluate the resulting algorithm. At 774M, three runs per configuration show fewer strictly detected episodes and better paired validation than conventional AdamW. Half-rate and online global norm-matched controls retain episodes despite similar validation, supporting the value of selective attenuation. Component ablations and transfer tests establish the scope of these gains. Figure 1 illustrates the mechanism from moment response to loss damage alongside the full-training result.

Contributions. We (i) distinguish update capacity from measured tail growth; (ii) use shared-state interventions to establish the roles of displacement, allocation, and continued control in loss damage; and (iii) evaluate the resulting coordinate-and-tensor rule in complete training and larger-model stress tests.

2 Update Capacity and Realized Displacement

We distinguish the amplitude AdamW permits from the realized displacement whose loss effect can be tested.

2.1 What the two moment kernels permit

For a coordinate, let g_t be the gradient supplied after native clipping. AdamW starts with zero buffers and computes

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

With $\widehat{m}_t = m_t/(1-\beta_1^t)$ and $\widehat{v}_t = v_t/(1-\beta_2^t)$,

$$\theta_{t+1} = (1-\eta_t\lambda)\theta_t - \eta_t \widehat{u}_t, \qquad \widehat{u}_t = \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t}+\epsilon}. \tag{2}$$

The normalized update, adaptive displacement, and decoupled weight decay are distinct. Define the unit-mass temporal kernels

$$a_{t,k} = \frac{(1-\beta_1)\beta_1^k}{1-\beta_1^t}, \qquad b_{t,k} = \frac{(1-\beta_2)\beta_2^k}{1-\beta_2^t}, \qquad 0 \le k < t. \tag{3}$$

Then $\widehat{m}_t = \sum_k a_{t,k} g_{t-k}$ and $\widehat{v}_t = \sum_k b_{t,k} g_{t-k}^2$: the same history is averaged differently.

Table 1: **Complete 355M B0–B5 sweep at peak LR 10^{-3}.** All/post-warmup counts pool seeds 1234, 2027, and 3407 under the broader burden detector; a post-warmup episode begins after the 2K-step warmup. The Area per spikes and Spike loss peak columns report the sample mean ± sample SD over all detected episodes, including warmup. These event-level summaries are descriptive; seeds are the independent training replicates. N/A denotes one event and a dash an empty event set. Validation is mean ± sample SD at 16K ($n = 2$ except B3, where $n = 3$).

Run and (β_1, β_2)	C_∞	All spikes	Post-warmup spikes	Area per spikes	Spike loss peak	16K validation
B0 (.900, .900)	1.000	1	0	$0.948 \pm$ N/A	$4.402 \pm$ N/A	2.968 ± 0.002
B1 (.900, .950)	1.165	4	1	2.458 ± 1.579	4.390 ± 0.538	2.961 ± 0.003
B2 (.900, .990)	2.345	19	13	76.567 ± 145.556	5.436 ± 1.746	2.964 ± 0.007
B3 (.900, .999)	7.270	18	13	310.969 ± 204.079	8.231 ± 2.105	2.996 ± 0.015
B4 (.990, .999)	2.299	3	3	7.543 ± 2.578	3.343 ± 0.195	2.966 ± 0.009
B5 (.999, .999)	1.000	0	0	–	–	3.224 ± 0.001

Coordinate capacity. For geometric kernels, the following is the finite-time bound of Peng et al. (2025, Theorem 1), expressed using normalized weights. For any real gradient history with $\widehat{v}_t > 0$,

$$\frac{|\widehat{m}_t|}{\sqrt{\widehat{v}_t}} \le C_t, \qquad C_t^2 = \sum_k \frac{a_{t,k}^2}{b_{t,k}} = 1 + \chi^2(a_t \| b_t). \tag{4}$$

Weighted Cauchy–Schwarz proves the inequality; equality holds for $g_{t-k} \propto a_{t,k}/b_{t,k}$ (Appendix B). Thus $C_t \ge 1$, with equality for matched kernels. If $\beta_1^2 < \beta_2$,

$$C_\infty^2 = \frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2 - \beta_1^2)}. \tag{5}$$

At fixed $\beta_1 = 0.9$, lowering β_2 from 0.999 to 0.95 reduces C_∞ from 7.270 to 1.165; matching gives 1. This quantifies an amplitude effect of the practical beta choice while leaving first-moment averaging unchanged. It is not a prediction of spike probability or validation quality. Positive ϵ reduces the output; zero history gives zero. Appendix B.1.1 makes the algebraic equivalence to the earlier bound explicit.

Table 1 relates this envelope to observed excursions at fixed model size and peak learning rate. It uses the broader 3% relative-rise burden detector to retain mild excursions; the local interventions and main mitigation comparison use the strict 10% protocol defined in Section 3.1. Severity in this table pools all episodes, including warmup.

At $\beta_1 = 0.9$, the .99 and .999 second-moment settings each have 13 post-warmup episodes, but the latter has $4.1\times$ the mean all-episode area. Across the full grid, similar capacity can accompany very different burdens, and matched $(.999, .999)$ has no detected episode but worse validation than matched $(.9, .9)$. Capacity therefore describes available amplitude. Realized damage also depends on the trajectory, which changes when beta changes. We next introduce a reference for the observed tail and then test displacement directly.

2.2 A STATIONARY REFERENCE FOR THE COORDINATE TAIL

For independent stationary gradients with mean μ and variance σ^2, the moment calculation of Balles and Hennig (2018) gives

$$\mathbb{E}\widehat{m}_t^2 = \mu^2 + \sigma^2 \rho_t(\beta_1), \qquad \mathbb{E}\widehat{v}_t = \mu^2 + \sigma^2, \qquad \rho_t(\beta) = \frac{(1-\beta)(1+\beta^t)}{(1+\beta)(1-\beta^t)}. \tag{6}$$

At zero mean, the late-time ratio-of-moments reference is $\sqrt{(1-\beta_1)/(1+\beta_1)} = 0.2294$ for $\beta_1 = 0.9$. This ordinary reference lies well below the capacity 7.27. Under the additional Gaussian assumption, the proxy $U_t^{\mathrm{ref}} = \widehat{m}_t/\sigma$ has $\mathbb{E}|U_\infty^{\mathrm{ref}}| = 0.1830$ and $\mathrm{sd}(|U_\infty^{\mathrm{ref}}|) = 0.1383$. Mean plus one or two standard deviations gives 0.321 or 0.460, motivating the nearby control strengths 0.3 and 0.45. The proxy approximates the actual update when the denominator concentrates and $\epsilon \ll \sigma$. These scales calibrate magnitudes; they do not identify damaging coordinates or predict spike frequency. Appendix B.2 gives the assumptions and finite-time analysis.

2.3 From amplitude to loss damage

For $\widehat{u}_t \neq 0$, let $A_t = \|\widehat{u}_t\|_\infty / C_t \in (0, 1]$ and $\widetilde{u}_t = \widehat{u}_t / \|\widehat{u}_t\|_\infty$. With $\Omega_t = \mathrm{diag}(\omega_{t,i})$, $0 \leq \omega_{t,i} \leq 1$, the coordinate displacement is

$$\widehat{u}_{t,i} = C_t A_t \widetilde{u}_{t,i}, \qquad w_{t,i} = \omega_{t,i} \widehat{u}_{t,i}, \qquad \Delta^{\mathrm{ad}}_{t,i} = -\eta_t w_{t,i}. \tag{7}$$

Here C_t is available amplitude, A_t its realized fraction, and $\widetilde{u}_t$ the update's shape. The native displacement scale $\eta_t C_t A_t$ alone determines neither tail breadth nor loss damage.

Hold weight decay fixed at $\bar{\theta}_t = (1 - \eta_t \lambda)\theta_t$. For the same objective L, let $H_t = \nabla^2 L(\bar{\theta}_t)$. Expanding the local loss change coordinate by coordinate gives

$$\begin{aligned} \Delta L_t(\Omega_t) &= L(\bar{\theta}_t - \eta_t w_t) - L(\bar{\theta}_t) \\ &\approx \sum_i \left[-\eta_t \partial_i L(\bar{\theta}_t) w_{t,i} + \tfrac{\eta_t^2}{2} H_{t,ii} w_{t,i}^2 \right] \\ &\quad + \eta_t^2 \sum_{i<j} H_{t,ij} w_{t,i} w_{t,j}. \end{aligned} \tag{8}$$

For a gradient-aligned coordinate with positive diagonal curvature, linear descent competes with a quadratic cost; the final sum retains cross-coordinate interactions. Global scaling sets all $\omega_{t,i} = \gamma$: linear terms scale as γ, curvature terms as γ^2. Selective attenuation varies these factors by coordinate, so equal norms need not give equal loss changes. Section 3 measures tail recruitment, then tests global scaling and matched-norm allocation by direct loss evaluation (Appendix B.4).

3 From Update Growth to Immediate Damage

3.1 Measurements and shared-state interventions

GPT-2 uses a common OpenWebText fallback token stream, sequence length 1024, global batch 480, gradient clipping at 1, and FP16 with dynamic loss scaling and FP32 moments. Default AdamW settings are $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.1, and peak rate 10^{-3}. The 16K-step budget includes 2K warmup steps and the prefix of a 50K cosine schedule. This is a high-β_2 stress setting, not a claim about the original GPT-2 pretraining recipe; lower and matched betas are explicit baselines (Appendix C).

Immediate loss is evaluated on the same batch after one candidate displacement and before a further update. *Episode damage* measures grouped training-loss excursions; *training quality* uses validation and completion. The strict detector requires a 10% relative rise, a 0.15 absolute rise, and a robust score of eight. Post-warmup episodes begin after step 2000 (Appendix C). Three runs supply 19 captured loss-spike onset states (8/6/5 by seed), nine ordinary controls, and three recovery controls. Replay restores parameters, moments, scaler, data, and RNG state; immediate comparisons share weight decay. These interventions test the next displacement at a captured state, without assuming that it caused the original detected rise. Later branches receive their own gradients. Seeds are the independent training replicates; the 19 states are nested measurements.

3.2 Loss-spike onsets recruit a broad, first-moment-led tail

New entrants into the coordinate tail are

$$\mathcal{R}_t(\tau) = \{i : |\widehat{u}_{t-1,i}| \leq \tau,\ |\widehat{u}_{t,i}| > \tau\}. \tag{9}$$

At $\tau = 0.45$, their median fraction is 4.021% at spike onsets versus 0.516% at ordinary states (7.8×); at $\tau = 1$, it is 0.571% versus 0.0033% (Figure 2). These are new entries, not total tail mass. For nonzero magnitudes,

$$\Delta \log |\widehat{u}_{t,i}| = \Delta \log |\widehat{m}_{t,i}| - \Delta \log(\sqrt{\widehat{v}_{t,i}} + \epsilon). \tag{10}$$

The median within-state contributions at spike onsets are $+1.593$ and -0.028 at $\tau = 0.45$. Growth is first-moment-led; the denominator partly offsets it. This identifies a spike-onset signature, not the source of the initiating gradient pattern. Crossing 0.45 also does not itself require mismatch-specific capacity above one.

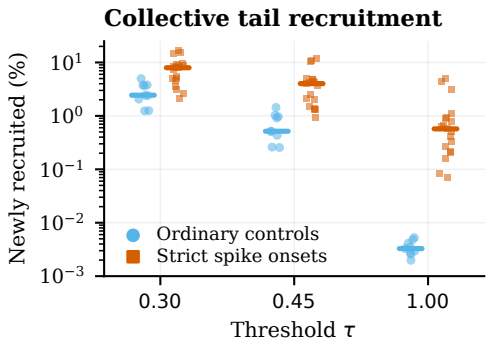


Figure 2: **Collective tail recruitment.** Points are 19 strictly detected loss-spike onset states or nine ordinary states; bars mark pooled event medians on a log scale. Events are nested in three seeds.

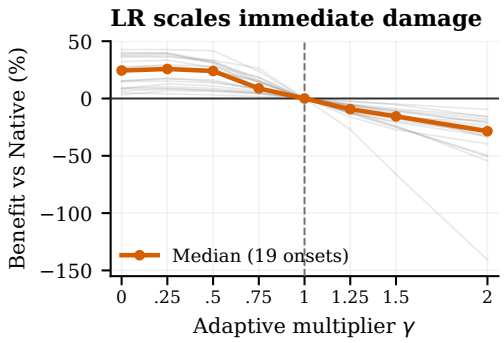


Figure 3: **Immediate loss versus adaptive displacement.** Gray curves show 19 spike-onset states; orange is their pooled median. Positive means lower post-update loss than Native ($\gamma = 1$).

3.3 Uniformly shrinking the adaptive update reduces immediate damage

At each captured spike-onset state, multiply every coordinate of the native adaptive displacement by the same scalar $\gamma \in \{0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 2\}$. This sets $\omega_{t,i} = \gamma$ for every coordinate in Equation (7), extended to $\gamma > 1$ for the amplification diagnostic. The moments, direction, evaluation batch, and decoupled weight decay are fixed. Immediate loss increases monotonically as this global multiplier rises from 0.25 to 2 at all 19 spike-onset states; reducing the native $\gamma = 1$ update to one quarter therefore helps at every state. The quarter-sized adaptive update also beats omitting it ($\gamma = 0$) at 16/19 states (Figure 3). Global shrinkage is thus useful as an immediate intervention, while retaining a small adaptive step often improves further over complete suppression. This finite-difference result supports the distinction in Equation (8) without requiring its quadratic approximation over the entire interval. It also leaves a second question: at the same displacement norm, does the choice of coordinates matter?

3.4 Coordinate allocation matters at equal norm

A *no-recruit* transform preserves the current update on coordinates already above τ before onset and caps the others at τ. Compare it with a radial control:

$$u^{\mathrm{NR}} = T_\tau(\widehat{u}_t), \qquad u^{\mathrm{rad}} = \frac{\|T_\tau(\widehat{u}_t)\|_2}{\|\widehat{u}_t\|_2}\widehat{u}_t. \tag{11}$$

Both match adaptive norm and weight-decay displacement at the shared state. No-recruit varies $\omega_{t,i}$ by coordinate; radial control uses one factor. At $\tau = 0.45$, no-recruit has lower immediate loss at 19/19 spike-onset states. The pooled median advantage is 6.597% of Native's post-update loss; the mean $\pm$ sample SD of the three *within-seed medians* (8.681, 2.333, 7.455) is 6.16 ± 3.37 percentage points. At $\tau = 1$, the advantage is positive at 17/19, with pooled median 3.174%. The audited relative norm mismatch stays below 3.28×10^{-8} (Figure 4).

Equal norm is not equal damage. The comparison identifies an allocation and direction effect, not a unique causal role for new tail entrants. It can also reflect better preservation of other useful coordinates. The match is global, so it can reflect redistribution both between and within tensors. Only the first shared-state step has equal norms across branches; later replay effects include changed trajectories.

3.5 Blockwise damage and a tensor-history reference

After workflow development on three discovery events, six additional events from the same three seeds test where attenuation helps. These events were also used in earlier diagnostics. With native adaptive *displacement* $d = \eta_t \widehat{u}_t$, drop one block at a time and compare with whole-model shrinkage to the same norm: $d^{-b} = d - P_b d$ and $d^{\mathrm{rad},b} = \|d^{-b}\|_2 d / \|d\|_2$. The diagonal factors are zero on

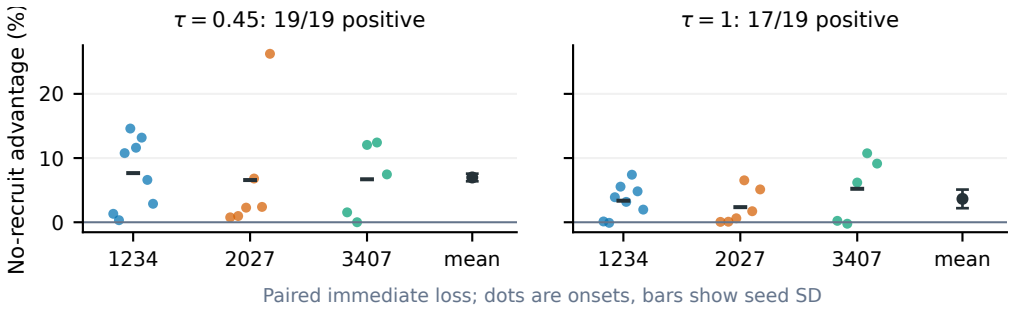


Figure 4: **Coordinate allocation at equal norm.** Each dot is a paired spike-onset state, grouped by seed; positive values favor no-recruit. Black ticks mark seed means; the rightmost point shows their mean and sample SD. Zero denotes equal immediate loss.

the dropped block and one elsewhere. Figure 5 reports

$$H_{e,b} = L^{\mathrm{rad}}_{e,b} - L^{\mathrm{drop}}_{e,b}, \quad B_{e,b} = L^{\mathrm{Native}}_{e} - L^{\mathrm{drop}}_{e,b}. \tag{12}$$

Block 0 has $H_{e,b} > 0$ and $B_{e,b} > 0$ at all six events. The maps establish a spatial effect beyond global norm. Because different deletions remove different energy, their effects are conditional on the remaining updates and do not define an intrinsic block ranking. To move from identifying a location to choosing an adaptive reference, we next compare allocations under a common energy budget.

A second comparison removes the same 10% of adaptive squared norm, allocating attenuation uniformly, by current tensor magnitude, by relative historical growth, or by shuffled historical scores. Here "history" is the median of each tensor's raw mean-absolute update over its preceding 128 successful updates, not an EMA controller. Historical allocation beats all three equal-budget controls at 6/6 validation events (Figure 5(b)); differences between allocations are near zero at three ordinary states. Thus recent history adds useful allocation information beyond current magnitude in these selected states. This single-step result motivates a tensor-relative reference; the online filter and its threshold are evaluated as part of the final algorithm (Appendix E.4).

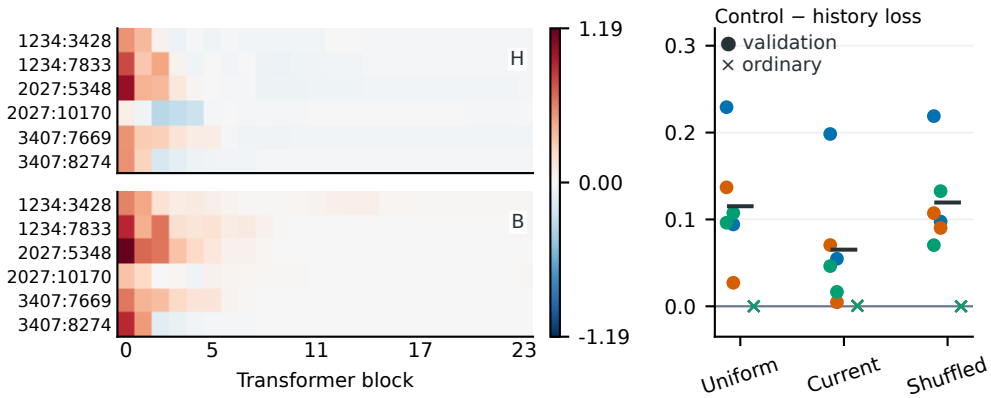


Figure 5: **Local damage and the allocation of attenuation.** Positive H favors dropping a block over equal-norm global shrinkage; positive B also improves on Native. The allocation panel retains validation and ordinary states separately; Native is not an equal-budget control.

4 Why Control Must Continue Beyond One Update

An immediate benefit need not persist along the subsequent trajectory. Clipping only the spike-onset update at 0.45 and continuing the same clipping have identical immediate losses, yet sustained clipping lowers the 32-step probe peak at 19/19 spike-onset states. Median peak ratios are 2.583 for Native, 2.488 for onset-only clipping, and 1.063 for sustained clipping. The advantage remains

10/10 where none of the three branches skips an update. The identical initial intervention therefore leaves substantial scope for later updates to amplify or suppress the excursion.

In a separate three-event memory diagnostic, both branches first take the same native step. One then rescales stored moments as $v' = a^2 v$ and $m' = mD'/D$, where D, D' are the exact bias-corrected denominators including ϵ. Parameters and the current normalized output are unchanged. Different later responses show that the stored state matters beyond the initial displacement (Appendix F). The beta time scales are unchanged, and the next clipped gradient is shared before later trajectories diverge. Together, these results motivate a rule that operates throughout training and adapts to recent update scale.

5 From the Evidence to Dual-Scale Update Control

5.1 Design principles and alternative controls

Coordinate interventions motivate limiting large outputs; fixed-budget allocation motivates a tensor-relative reference; sustained replay motivates control at every update. The online cap applies without the pre-onset set used by no-recruit. The tensor rule replaces the diagnostic's fixed budget and median with a threshold and an EMA. Full-training ablations test this implementation of the three design principles.

Matching moments and lowering the learning rate remain useful alternatives. Matching minimizes capacity at one; an explicit cap can impose a smaller limit independently of the moment time scales. Selective attenuation also allows some outputs to retain their native size. Section 6 evaluates training outcomes alongside beta and learning-rate alternatives.

5.2 An absolute cap and a relative-growth constraint

Set $\widetilde{u}_{t,i} = \mathrm{clip}(\widehat{u}_{t,i}, -c, c)$, using $c = 0.3$ by default and $c = 0.45$ as a less aggressive check. Unlike no-recruit, this rule needs neither spike-onset detection nor a fixed coordinate set. Section 2.2 explains its scale, not its optimality. For tensor ℓ, measure the *raw* magnitude $r_t^{(\ell)} = |\ell|^{-1} \sum_{i \in \ell} |\widehat{u}_{t,i}|$. Given a historical reference $b_{t-1}^{(\ell)}$, retain the largest factor $s \in [0, 1]$ satisfying $s(r_t^{(\ell)} + \varepsilon_r) \le \kappa b_{t-1}^{(\ell)}$. Thus

$$s_t^{(\ell)} = \min\left\{1, \frac{\kappa b_{t-1}^{(\ell)}}{r_t^{(\ell)} + \varepsilon_r}\right\}, \qquad u_{t,i}^* = s_t^{(\ell)} \,\mathrm{clip}(\widehat{u}_{t,i}, -c, c). \tag{13}$$

This solves a stated amplitude constraint, not loss minimization. We use a lagged EMA with $\alpha = 127/128$:

$$b_t^{(\ell)} = \begin{cases} \alpha b_{t-1}^{(\ell)} + (1-\alpha) r_t^{(\ell)}, & s_t^{(\ell)} = 1, \\ b_{t-1}^{(\ell)}, & s_t^{(\ell)} < 1. \end{cases} \tag{14}$$

The first observation initializes the EMA; tensor scaling starts after 128 successful observations, while coordinate clipping remains active. We use $\kappa = 2$ and FP32 machine epsilon for ε_r. Lagging and freezing avoid incorporating a burst into its own reference. These are online design choices, not uniquely inferred by the median-based diagnostic. Algorithm 1 gives the implementation; Appendix G details initialization and skip handling.

5.3 What the two constraints control

The method has exactly the form in Equation (7), with

$$\omega_{t,i} = s_t^{(\ell(i))} \min\left\{1, \frac{c}{|\widehat{u}_{t,i}|}\right\}, \qquad u_{t,i}^* = \omega_{t,i} \widehat{u}_{t,i}. \tag{15}$$

Here $\ell(i)$ is the containing tensor, and the clipping factor is one at zero. A below-cap coordinate in a non-triggered tensor keeps its native update at that state; other adaptive displacements shrink. The base schedule is unchanged, but effective coordinate step sizes and subsequent gradients can change. Clip-0.3 affects about 15.2% of coordinates in audited late-run snapshots, so control is not restricted to rare events.

The output obeys

$$|u^*_{t,i}| \le c, \qquad \mathrm{mean}_{i\in\ell}\,|u^*_{t,i}| \le s_t^{(\ell)} r_t^{(\ell)} \le \kappa b_{t-1}^{(\ell)}, \tag{16}$$

with the second inequality chain applying after 128 observations. Sparse large coordinates and broad sub-cap growth can activate different constraints. Tensor scaling preserves the clipped tensor's direction, not necessarily the whole-model direction. These are adaptive-output bounds, not loss guarantees. The rule adds one scalar reference and count per tensor, plus reductions and elementwise operations; it requires no additional forward or backward pass and does not reset moments.

6 Full-Training Evaluation

6.1 Moment mismatch and learning rate across trajectories

We evaluate episode burden, validation quality, and completion separately. Full-run beta and learning-rate changes also change future gradients, so these comparisons characterize training outcomes rather than isolated causal effects of the capacity bound. The 355M learning-rate grid extends the fixed-rate beta sweep in Table 1.

Under the strict detector, across four rates per moment setting and three seeds, conventional moments have mean post-warmup cumulative areas 91.25, 1017.25, 1349.64, and 2522.36 at $\{0.25, 0.5, 1, 2\} \times 10^{-3}$. Counts are not monotone. Lower rates can improve validation too: 0.5×10^{-3} gives 2.9688 ± 0.0093, versus 2.9956 ± 0.0146 at 10^{-3}. Matched moments have no post-warmup episodes at 0.5 and 1×10^{-3}, but five and eleven at 2 and 4×10^{-3}. Appendix D reports all rows and endpoint provenance. At 774M, $(0.9, 0.95)$ also improves burden and validation (Table 2). Fully matched $(0.9, 0.9)$ reaches zero post-warmup episodes and validation 2.8640 ± 0.0060 in three complete runs (Appendix K.5), establishing a strong alternative to output control under the same training contract.

6.2 Dual control: complementary effects and transfer

Table 2 compares three complete 774M runs per method. Clip-0.3 leaves four post-warmup episodes and EMA-2 thirteen; their combination leaves zero under the frozen detector. Both components already reduce the event count and mean per-event area. Relative to conventional AdamW, the combination improves validation in each paired seed, with the mean changing from 2.9160 ± 0.0305 to 2.8558 ± 0.0043. Its clearest additional benefit over Clip-0.3 alone is the reduction in remaining strict episodes, from four to zero; their small validation gap is not an established population-level advantage. The looser combined cap gives similar validation with one remaining post-warmup episode. A broader legacy detector retains two post-warmup episodes for the default combination, so zero strict detections should be read at the stated threshold (Appendix K).

Global attenuation controls. We add two controls trained from initialization for the same three seeds: half the learning-rate schedule, and online global scaling that matches the norm of a virtual Dual adaptive displacement at the control's own current state. The latter preserves AdamW's adaptive direction while matching the attenuation in norm; it excludes decoupled weight decay. Table 2 shows 15 and 27 post-warmup episodes, respectively, versus zero for Dual. Global scaling substantially reduces damage: mean post-warmup cumulative area falls from 1693.61 for Standard to 120.87, despite its larger episode count. Half-rate validation (2.8563 ± 0.0147) is nearly identical to Dual (2.8558 ± 0.0043), while global matching gives 2.8640 ± 0.0096. Thus the additional observed benefit is episode suppression at comparable endpoint quality. Neither global control reproduces it, supporting the value of selective attenuation under this training contract. Appendix K.4 specifies the control and per-seed results.

Transfer and completion. We test the same rule across models and precisions (Appendix Table 7). On Qwen3-0.6B in BF16, dual control reduces total/post-warmup episodes from 12/10 to 0/0 over three complete runs and improves every paired validation endpoint. Architecture, tokenizer, and precision change together.

All three guarded NanoGPT-1.5B runs complete 16K steps with no post-warmup episode and validation 2.7747 ± 0.0046; two Standard runs diverge at 10,914 and 9,985; the third is administratively stopped at 12,252. Standard outcomes are observed prefixes without a common 16K endpoint.

Table 2: **774M controls and ablations.** Three complete 16K runs per row; strict recovery episodes. Area per spikes and Spike loss peak pool all episodes (mean ± sample SD); validation pools seed endpoints likewise. C_∞ bounds the unguarded moment kernel.

Configuration	C_∞	All spikes	Post-warmup spikes	Area per spikes	Spike loss peak	16K final validation
Moment kernels without a guard						
Conventional (.9, .999)	7.270	23	17	315.167 ± 250.457	8.509 ± 2.998	2.9160 ± 0.0305
Reduced mismatch (.9, .95)	1.165	10	6	8.800 ± 7.713	4.629 ± 0.971	2.8704 ± 0.0156
Global controls with conventional moments						
Half learning rate	7.270	20	15	231.710 ± 245.813	7.109 ± 2.664	2.8563 ± 0.0147
Global norm match	7.270	33	27	21.031 ± 29.441	4.822 ± 1.213	2.8640 ± 0.0096
Update guards with conventional moments						
Clip-0.3	7.270	10	4	21.070 ± 13.332	4.562 ± 0.684	2.8586 ± 0.0055
Clip-0.45	7.270	19	12	41.063 ± 34.167	4.768 ± 1.107	2.8661 ± 0.0057
EMA-2	7.270	18	13	14.197 ± 10.187	4.565 ± 0.812	2.8603 ± 0.0065
Clip-0.3+EMA-2	7.270	2	0	2.239 ± 1.070	5.363 ± 0.354	$\mathbf{2.8558 \pm 0.0043}$
Clip-0.45+EMA-2	7.270	5	1	10.541 ± 4.674	5.181 ± 0.699	2.8570 ± 0.0099

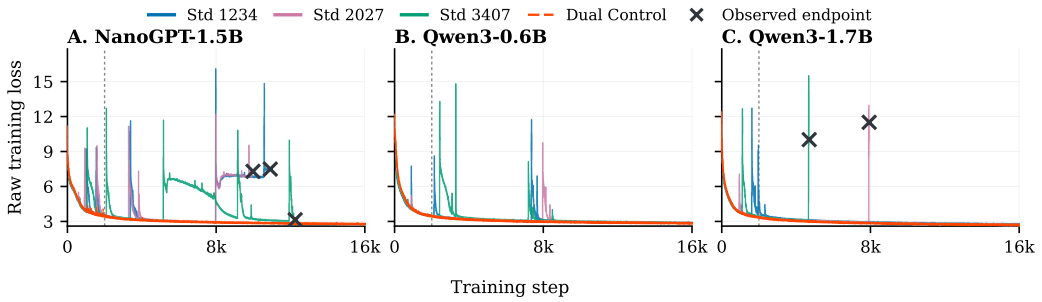


Figure 6: **Transfer trajectories.** Raw loss for NanoGPT-1.5B, Qwen3-0.6B, and Qwen3-1.7B. Standard colors (blue/magenta/green) and Dual line styles (solid/dashed/dotted) follow seeds 1234/2027/3407. The dotted line ends warmup at 2K; crosses mark early endpoints.

On Qwen3-1.7B in FP16, all three Dual runs reach 16K with no detected episodes and validation 2.7973 ± 0.0016; Standard seeds 2027 and 3407 diverge at 7,933 and 4,706. Its seven episodes are observed-prefix lower bounds. Only seed 1234 allows a paired 16K validation comparison (2.8132 versus 2.7963). Figure 6 shows the individual trajectories.

Appendix H gives transfer provenance and per-seed results.

Scope. We deliberately use a peak learning rate of 10^{-3} and the first 16K steps of a 50K-step cosine schedule. The high rate and extended decay are intended to elicit more spikes during observation, providing more events to study their mechanisms and a demanding setting in which to evaluate mitigation. Selected 355M states and three-seed 774M comparisons support this tested stress setting, without identifying a unique trigger or covering every tuned learning-rate schedule.

7 Conclusion

AdamW's moment kernels bound update amplitude. Shared-state interventions link displacement and allocation to immediate loss; replay motivates sustained coordinate and tensor control. In full training, Dual suppresses strict episodes where half-rate and global norm-matched controls do not, with competitive validation. Ablations and transfer tests support its benefit under the tested contracts.

AI USE STATEMENT

In preparing this work, we used generative AI tools to polish the writing, identify and retrieve relevant literature, summarize and analyze prior work, assist in refining proofs, and create the conceptual schematic in Figure 1(a). We reviewed all AI-assisted material and conducted a manual literature survey to check AI-suggested research ideas for potential overlap with prior work. The authors take full responsibility for the final manuscript, including any text, claims, or artifacts developed with AI assistance.

REPRODUCIBILITY STATEMENT

The appendices specify model architectures, optimizer and control settings, schedules, data processing, detector thresholds, replay timing, and endpoint rules (Appendix C). The experiment repository retains per-run configurations, seed IDs, raw training and validation metrics, replay responses, and source manifests. GPT-2 uses a tokenized OpenWebText fallback rather than a canonical reconstruction; the Qwen token stream is decoded from that fallback and retokenized. Shared-state comparisons restore parameters, optimizer moments, loss-scaler and RNG states, and data position; restoration audits check batch hashes. Immediate loss is evaluated before another branch update. Training seeds are the independent replicates, while sampled spike-onset states and probes are nested observations. Nonfinite or stopped runs contribute only observed-prefix outcomes, and final-validation means require logged 16K endpoints.

REFERENCES

Jinze Bai et al. Qwen technical report. arXiv:2309.16609, 2023.

Zhiwei Bai, Zhangchen Zhou, Jiajie Zhao, Xiaolong Li, Zhiyu Li, Feiyu Xiong, Hongkang Yang, Yaoyu Zhang, and Zhi-Qin John Xu. Adaptive preconditioners trigger loss spikes in Adam. arXiv:2506.04805v2, 2026.

Lukas Balles and Philipp Hennig. Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of PMLR, pages 404–413, 2018.

Xiao Bi et al. DeepSeek LLM: Scaling open-source language models with longtermism. arXiv:2401.02954, 2024.

Tom Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Matias D. Cattaneo and Boris Shigida. The effect of mini-batch noise on the implicit bias of Adam. arXiv:2602.01642, 2026.

Aakanksha Chowdhery et al. PaLM: Scaling language modeling with Pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

Aaron Grattafiori et al. The Llama 3 herd of models. arXiv:2407.21783, 2024.

Tianjin Huang, Ziquan Zhu, Gaojie Jin, Lu Liu, Zhangyang Wang, and Shiwei Liu. SPAM: Spike-aware Adam with momentum reset for stable LLM training. arXiv:2501.06842, 2025.

Tianjin Huang, Zhangyang Wang, Haotian Hu, Zhenyu Zhang, Gaojie Jin, Xiang Li, Li Shen, Jiaxing Shang, Tianlong Chen, Ke Li, Lu Liu, Qingsong Wen, and Shiwei Liu. GradientStabilizer: Fix the norm, not the gradient. In *International Conference on Machine Learning*, 2026.

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

Teven Le Scao et al. What language model to train if you have one million GPU hours? arXiv:2210.15424, 2022.

Aixin Liu et al. DeepSeek-V2: A strong, economical, and efficient mixture-of-experts language model. arXiv:2405.04434, 2024a.

Aixin Liu et al. DeepSeek-V3 technical report. arXiv:2412.19437, 2024b.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pre-training approach. arXiv:1907.11692, 2019.

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

Chao Ma, Lei Wu, and Weinan E. A qualitative study of the dynamic behavior for adaptive gradient algorithms. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, volume 145 of PMLR, pp. 671–692, 2022.

Igor Molybog, Peter Albert, Moya Chen, Zachary DeVito, David Esiobu, Naman Goyal, Punit Singh Koura, Sharan Narang, Andrew Poulton, Ruan Silva, Binh Tang, Diana Liskovich, Puxin Xu, Yuchen Zhang, Melanie Kambadur, Stephen Roller, and Susan Zhang. A theory on Adam instability in large-scale machine learning. arXiv:2304.09871, 2023.

Antonio Orvieto and Robert M. Gower. In search of Adam's secret sauce. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28(3) of PMLR, pp. 1310–1318, 2013.

Hanyang Peng, Shuang Qin, Yue Yu, Fangqing Jiang, Hui Wang, and Wen Gao. SoftSignSGD(S3): An enhanced optimizer for practical DNN training and loss spikes minimization beyond Adam. arXiv:2507.06464, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI technical report, 2019.

Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

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 PMLR, pp. 4596–4604, 2018.

Gaoxiang Tang, Huanran Chen, and Ziming Liu. Beyond quadratic loss: The stability phase diagram of Adam. arXiv:2609.18314, 2026.

Hugo Touvron et al. LLaMA: Open and efficient foundation language models. arXiv:2302.13971, 2023a.

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023b.

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. arXiv:2309.14322, 2023.

Aiyuan Yang et al. Baichuan 2: Open large-scale language models. arXiv:2309.10305, 2023.

An Yang et al. Qwen2 technical report. arXiv:2407.10671, 2024a.

An Yang et al. Qwen2.5 technical report. arXiv:2412.15115, 2024b.

An Yang et al. Qwen3 technical report. arXiv:2505.09388, 2025.

Aohan Zeng et al. GLM-130B: An open bilingual pre-trained model. In *International Conference on Learning Representations*, 2023.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham Kakade. Deconstructing what makes a good optimizer for autoregressive language models. In *International Conference on Learning Representations*, 2025.

A Related Work and Scope

Operational and architectural responses. PaLM restored checkpoints and skipped nearby batches to recover from loss spikes (Chowdhery et al., 2023). Other interventions target particular model components: GLM-130B shrinks embedding gradients (Zeng et al., 2023), and Baichuan 2 normalizes the output embedding (Yang et al., 2023). Architectural normalization also has trade-offs: in a separate model study, Le Scao et al. (2022) found that an extra normalization after the input embedding reduced zero-shot performance. Molybog et al. (2023) link instability to time-correlated gradient estimates in earlier layers and discuss smaller ϵ and batch size as possible remedies, with numerical and throughput costs. Our intervention instead acts on the realized AdamW update while retaining its moment state.

Beta choice and stability. RoBERTa reports improved large-batch stability with $\beta_2 = 0.98$ (Liu et al., 2019). Ma et al. (2022) identify oscillation, spike, and divergence regimes, favoring nearby betas in full-batch and large-batch experiments. Zhao et al. (2025) compare tied betas with Signum; Orvieto and Gower (2025) establish strong tuned LM performance and an online mean/variance interpretation. Thus matching is not an unexplained empirical observation. Peng et al. (2025) already provide the finite-time amplitude bound and matching experiment. Our emphasis is its kernel expression together with shared-state size/allocation interventions, spatial controls, and episode dynamics.

Complementary mechanisms and controls. Preconditioner curvature and update alignment offer other accounts of instability (Molybog et al., 2023; Wortsman et al., 2023; Bai et al., 2026). Implicit-bias predictions can reverse with batch size (Cattaneo and Shigida, 2026); geometry also affects beta stability regions (Tang et al., 2026). Neither lower nor equal betas universally optimize all outcomes. Adafactor uses update clipping (Shazeer and Stern, 2018); SPAM uses gradient-spike control and moment resets (Huang et al., 2025). GradientStabilizer instead stabilizes gradient norm using its running statistics (Huang et al., 2026). Our output control retains moments; Appendix I includes the tested SPAM comparison and its failed path.

We use three independent seeds in a stressful setting. Local effects concern selected states and do not identify a unique natural trigger or online filter. Dual control is evaluated with $(0.9, 0.999)$; the $(0.9, 0.95)$ baseline does not establish its additional benefit on top of that recipe. We claim neither universally optimal thresholds nor superiority over all tuned lower-beta or lower-rate baselines.

B Theory: Capacity, Typical Scale, and Loss Damage

B.1 Proof and finite-time form of the coordinate bound

For normalized positive kernels a_k, b_k and any real history, weighted Cauchy–Schwarz gives

$$\left(\sum_k a_k g_k\right)^2 = \left(\sum_k \frac{a_k}{\sqrt{b_k}}\sqrt{b_k} g_k\right)^2 \le \left(\sum_k \frac{a_k^2}{b_k}\right)\sum_k b_k g_k^2. \tag{17}$$

Equality holds when $\sqrt{b_k} g_k$ is proportional to $a_k/\sqrt{b_k}$, equivalently $g_k \propto a_k/b_k$. Furthermore,

$$\chi^2(a\|b) = \sum_k \frac{(a_k - b_k)^2}{b_k} = \sum_k \frac{a_k^2}{b_k} - 2\sum_k a_k + \sum_k b_k = C_t^2 - 1. \tag{18}$$

Thus $C_t \ge 1$, with equality precisely when the normalized kernels coincide. For $t > 1$, Adam's geometric kernels coincide only when $\beta_1 = \beta_2$; at $t = 1$ both contain one unit weight. The bound applies coordinatewise. Its extremal histories need not be simultaneously realizable along a coupled neural-network trajectory.

Set $q = \beta_1^2/\beta_2$. Summing the geometric series gives

$$C_t^2 = \frac{(1-\beta_1)^2}{1-\beta_2}\frac{1-\beta_2^t}{(1-\beta_1^t)^2}\begin{cases}(1-q^t)/(1-q), & q \ne 1,\\ t, & q = 1.\end{cases} \tag{19}$$

If $q < 1$, the limit is

$$C_\infty^2 = \frac{\beta_2(1-\beta_1)^2}{(1-\beta_2)(\beta_2 - \beta_1^2)}. \tag{20}$$

When $q \geq 1$, there is no finite asymptotic envelope of this form, although every finite-time bound remains valid. For positive ϵ, $|\widehat{u}_t| \leq |\widehat{m}_t|/\sqrt{\widehat{v}_t} \leq C_t$. If $\widehat{v}_t = 0$, every gradient in the positively weighted history is zero, so the numerator is also zero; the implemented positive epsilon yields zero adaptive output.

B.1.1 Relation to the finite-time bound of Peng et al.

Peng et al. (2025, Theorem 1) already give a finite-time bound, not merely an asymptotic one. For $q = \beta_1^2/\beta_2 < 1$, their coefficient is

$$B_t = \frac{(1-\beta_1)\sqrt{1-\beta_2^t}\sqrt{1-q^t}}{(1-\beta_1^t)\sqrt{1-\beta_2}\sqrt{1-q}}. \tag{21}$$

Squaring this expression gives exactly C_t^2 above. Their equality history has a common sign and $g_{t-k} \propto (\beta_1/\beta_2)^k$, which is also $g_{t-k} \propto a_{t,k}/b_{t,k}$ for a fixed t. Their reported limiting value 7.27 at $(0.9, 0.999)$ is therefore the same quantity. The geometric amplitude bound and the observation that matching minimizes it are prior results. Here the expression through arbitrary normalized kernels and $1 + \chi^2(a_t\|b_t)$ makes the temporal mismatch explicit; the finite sum also remains meaningful when $q \geq 1$. We use this formulation to organize the displacement interventions, not to claim a new geometric Adam bound.

B.2 Reference scales for normalized updates

B.2.1 General stationary case

The squared mass of a normalized geometric kernel is

$$\rho_t(\beta) = \sum_{k=0}^{t-1} \left(\frac{(1-\beta)\beta^k}{1-\beta^t} \right)^2 = \frac{(1-\beta)(1+\beta^t)}{(1+\beta)(1-\beta^t)}. \tag{22}$$

For one coordinate, suppose the gradients are independent and stationary over the moment-estimation horizon, with

$$\mathbb{E}g_s = \mu, \qquad \mathrm{Var}(g_s) = \sigma^2, \qquad \mathrm{Var}(g_s^2) < \infty. \tag{23}$$

With the normalized geometric weights from Section 2,

$$\widehat{m}_t = \sum_{k=0}^{t-1} a_{t,k} g_{t-k}, \qquad \widehat{v}_t = \sum_{k=0}^{t-1} b_{t,k} g_{t-k}^2. \tag{24}$$

Independence gives

$$\mathbb{E}\widehat{m}_t = \mu, \qquad \mathrm{Var}(\widehat{m}_t) = \sigma^2 \rho_t(\beta_1), \tag{25}$$

$$\mathbb{E}\widehat{m}_t^2 = \mu^2 + \sigma^2 \rho_t(\beta_1), \qquad \mathbb{E}\widehat{v}_t = \mu^2 + \sigma^2. \tag{26}$$

These identities use only the first two moments and temporal independence; they do not require a Gaussian distribution. Write

$$q \equiv \mu^2 + \sigma^2, \qquad z \equiv \frac{|\mu|}{\sigma}, \qquad \kappa_v \equiv \frac{\mathrm{Var}(g_s^2)}{q^2}. \tag{27}$$

Because $\widehat{v}_t$ is a weighted average of independent g_s^2 terms,

$$\frac{\mathrm{Var}(\widehat{v}_t)}{q^2} = \kappa_v \rho_t(\beta_2). \tag{28}$$

The corresponding second-moment reference scale is

$$S_t = \sqrt{\frac{\mathbb{E}\widehat{m}_t^2}{\mathbb{E}\widehat{v}_t}} = \sqrt{\frac{z^2 + \rho_t(\beta_1)}{1+z^2}}. \tag{29}$$

An exact inequality requires an explicit meaning of "concentrated." Under the idealized assumption $\widehat{v}_t = \mathbb{E}\widehat{v}_t = q$ almost surely, Cauchy–Schwarz gives

$$\mathbb{E}|\widehat{u}_t| = \frac{\mathbb{E}|\widehat{m}_t|}{\sqrt{q}+\epsilon} \le \frac{\sqrt{\mathbb{E}\widehat{m}_t^2}}{\sqrt{q}+\epsilon} \le S_t. \tag{30}$$

If instead the denominator obeys the deterministic relative bound $\widehat{v}_t \ge (1-\delta)q$ almost surely for some $\delta \in (0,1)$, then the exact statement weakens to

$$\mathbb{E}|\widehat{u}_t| \le \frac{S_t}{\sqrt{1-\delta}}. \tag{31}$$

Finally, if concentration holds only with high probability, split on the same event. Cauchy–Schwarz controls its contribution, the pointwise capacity bound controls the lower-tail event, and Chebyshev's inequality gives

$$\begin{aligned} \mathbb{E}|\widehat{u}_t| &\le \frac{S_t}{\sqrt{1-\delta}} + C_t \Pr[\widehat{v}_t < (1-\delta)q] \\ &\le \frac{S_t}{\sqrt{1-\delta}} + C_t \min\left\{1, \frac{\kappa_v \rho_t(\beta_2)}{\delta^2}\right\}. \end{aligned} \tag{32}$$

The fixed-denominator identity above is exact under its stated idealization. With a random denominator, Equation (32) gives a correction rather than an unconditional bound by S_t itself. This bound need not be numerically sharp. Positive ϵ can only reduce the left-hand sides.

Noise-dominated corollary. The nonzero mean changes the reference scale according to the exact identity

$$S_t^2 = \rho_t(\beta_1) + \frac{[1-\rho_t(\beta_1)]z^2}{1+z^2}. \tag{33}$$

Consequently, a nonzero signal can only increase S_t above $\sqrt{\rho_t(\beta_1)}$. If the gradient is noise dominated in the stronger, quantitative sense $z^2 \ll \rho_t(\beta_1)$, then

$$S_t = \sqrt{\rho_t(\beta_1)}\left[1 + \frac{1-\rho_t(\beta_1)}{2\rho_t(\beta_1)}z^2 + O\left(\frac{z^4}{\rho_t(\beta_1)^2}\right)\right] \approx \sqrt{\rho_t(\beta_1)}. \tag{34}$$

Thus $|\mu| \ll \sigma$ alone is not the precise condition for recovering the zero-mean scale; one needs $\mu^2/\sigma^2 \ll \rho_t(\beta_1)$. At late time and $\beta_1 = 0.9$, this requires $|\mu|/\sigma \ll \sqrt{1/19} = 0.229$.

B.2.2 Gaussian special case

Only now impose the stronger Gaussian model

$$g_s = \mu + \xi_s, \qquad \xi_s \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{35}$$

This specialization is used only to obtain a distributional approximation and closed-form clipping calibration; the general bound above does not depend on Gaussianity. The moment identities agree with the decomposition of Balles and Hennig Balles and Hennig (2018). In particular, $\rho_t(\beta_1)$ multiplies the gradient-noise variance in $\mathbb{E}\widehat{m}_t^2$, not the squared signal μ^2.

Because $\widehat{m}_t$ is a linear combination of Gaussian variables,

$$\widehat{m}_t \sim \mathcal{N}\big(\mu, \rho_t(\beta_1)\sigma^2\big). \tag{36}$$

For the denominator, Gaussian moments give $\mathrm{Var}(g_s^2) = 2\sigma^4 + 4\mu^2\sigma^2$. Therefore

$$\mathrm{Var}(\widehat{v}_t) = \rho_t(\beta_2)(2\sigma^4 + 4\mu^2\sigma^2). \tag{37}$$

Writing $q = \mu^2 + \sigma^2$ and $z = |\mu|/\sigma$, its squared coefficient of variation is

$$\frac{\mathrm{Var}(\widehat{v}_t)}{q^2} = \rho_t(\beta_2)\frac{2+4z^2}{(1+z^2)^2} \le 2\rho_t(\beta_2). \tag{38}$$

For $\beta_2 = 0.999$ at late time, $\rho_t(\beta_2) \simeq 0.000500$, so the coefficient of variation of $\widehat{v}_t$ is at most 3.16% under this model. This is a reference-model calculation, not an empirical claim about every onset. Define the proxy

$$U_t^{\mathrm{ref}} = \frac{\widehat{m}_t}{\sqrt{q}} \sim \mathcal{N}\left(\frac{\mu}{\sqrt{\mu^2+\sigma^2}}, \frac{\rho_t(\beta_1)\sigma^2}{\mu^2+\sigma^2}\right). \tag{39}$$

The distribution is exact for the proxy under the Gaussian assumption. The actual update is approximately this proxy when the denominator concentrates and $\epsilon \ll \sqrt{q}$; its numerator and denominator need not be independent for the proxy to be defined.

At $\mu = 0$, $U_t^{\mathrm{ref}} \sim \mathcal{N}(0, \rho_t(\beta_1))$ exactly. For $\mu \neq 0$, the same zero-mean approximation requires $\mu^2/\sigma^2 \ll \rho_t(\beta_1)$, not merely $|\mu| \ll \sigma$. Writing $s_u^2 = \rho_t(\beta_1)$, the zero-mean proxy has

$$\mathbb{E}|U_t^{\mathrm{ref}}| = s_u\sqrt{2/\pi}, \qquad \mathrm{sd}(|U_t^{\mathrm{ref}}|) = s_u\sqrt{1-2/\pi}. \tag{40}$$

At late time and $\beta_1 = 0.9$, these are 0.18305 and 0.13829. The mean plus one or two standard deviations is 0.32134 or 0.45964. For the nearby thresholds used in the experiments,

$$\Pr(|U_\infty^{\mathrm{ref}}| > 0.3) = 19.10\%, \qquad \Pr(|U_\infty^{\mathrm{ref}}| > 0.45) = 4.98\%. \tag{41}$$

These are reference tail probabilities, not measured clipping rates, new-entrant probabilities, or probabilities of loss damage. The late-time reference also does not describe the first Adam update: $|\widehat{u}_1| = |g_1|/(|g_1| + \epsilon)$ can be near one.

B.2.3 Empirical consistency check

For comparison, we aggregate exact raw, pre-clipping update statistics at steps 12K, 13K, 14K, 15K, and 16K in each of the three protected 774M runs. Because $|\widehat{u}|^2 = \widehat{u}^2$, their pooled coordinate–snapshot summary gives

$$\widehat{\mathbb{E}}|\widehat{u}| = 0.16536, \qquad \sqrt{\widehat{\mathbb{E}}\widehat{u}^2} = 0.21110, \qquad \widehat{\mathrm{Var}}(|\widehat{u}|) = 0.21110^2 - 0.16536^2 = 0.01722. \tag{42}$$

The resulting standard deviation is 0.13122. These quantities use the exact raw L_1, raw squared L_2, and total coordinate count; they do not mix in the separately logged per-tensor sampled RMS diagnostic. Therefore $c = 0.3$ is $1.81\widehat{\mathbb{E}}|\widehat{u}|$ and lies 1.03 empirical standard deviations above the mean. The observed 15.2% coordinate activation rate is consistent with treating the cap as deliberate upper-tail regularization rather than as a detector reserved for extremely rare coordinates. A zero-mean Gaussian with standard deviation 0.21110 would predict $\mathbb{E}|\widehat{u}| = 0.16843$, $\mathrm{sd}(|\widehat{u}|) = 0.12725$, and $\Pr(|\widehat{u}| > 0.3) = 15.53\%$. The agreement is close at this bulk-tail boundary, although it does not establish a Gaussian law for the much rarer extreme tail.

These measurements are from guarded trajectories. Measuring the raw update before the current clip does not undo the effect of earlier guarded updates on the trajectory. The summary is a consistency check, not independent pre-selection calibration or proof of a Gaussian law for all coordinates.

B.2.4 Temporal dependence and the scope of the reference

The independent-gradient calculation excludes temporal covariance. With stationary marginal variance but correlated gradients,

$$\mathrm{Var}(\widehat{m}_t) = \sigma^2 \sum_k a_{t,k}^2 + 2\sum_{j<k} a_{t,j}a_{t,k}\,\mathrm{Cov}(g_{t-j}, g_{t-k}). \tag{43}$$

Signal, temporal dependence, changing gradient scale, and denominator fluctuations can all alter the actual update distribution. In particular, changing β_2 changes the accuracy of the denominator approximation, even though the zero-mean ratio-of-moments reference depends only on β_1. None of these reference assumptions is needed for Proposition 1. Nor do they determine the tensor-history multiplier κ.

B.3 Global scaling in the element-wise loss expansion

At the common weight-decay point $\bar{\theta}_t$, uniform scaling sets $\omega_{t,i} = \gamma$ with $\widehat{u}_t$ fixed. Equation (8) then gives

$$\begin{aligned}\Delta L_t(\gamma I) \approx & -\eta_t \gamma \sum_i \partial_i L(\bar{\theta}_t)\widehat{u}_{t,i} \\ & + \tfrac{\eta_t^2\gamma^2}{2}\left[\sum_i H_{t,ii}\widehat{u}_{t,i}^2 + 2\sum_{i<j} H_{t,ij}\widehat{u}_{t,i}\widehat{u}_{t,j}\right].\end{aligned} \tag{44}$$

Here $H_t = \nabla^2 L(\bar{\theta}_t)$. With positive total alignment and bracketed curvature, linear descent dominates for small positive γ, while the quadratic term grows faster. Coordinate and interaction terms may have either sign. Section 3 evaluates finite loss changes directly, without assuming this approximation holds over the tested range. In $\widehat{u}_{t,i} = C_t A_t \widetilde{u}_{t,i}$, changing beta need not leave the realized fraction, update shape, or loss geometry fixed.

B.4 Exact element-wise loss identity and limits of control

For $w_{t,i} = \omega_{t,i}\widehat{u}_{t,i}$, suppose L is twice continuously differentiable along the segment from $\bar{\theta}_t$ to $\bar{\theta}_t - \eta_t w_t$. Taylor's formula gives the exact identity

$$\begin{aligned}\Delta L_t(\Omega_t) = & -\eta_t \sum_i \partial_i L(\bar{\theta}_t) w_{t,i} \\ & + \eta_t^2 \sum_{i,j} w_{t,i} w_{t,j} \int_0^1 (1-a)\, \partial_{ij} L(\bar{\theta}_t - a\eta_t w_t)\, da.\end{aligned} \tag{45}$$

Freezing the Hessian entries at $H_{t,ij} = \partial_{ij} L(\bar{\theta}_t)$, using $\int_0^1 (1-a)\, da = 1/2$, and combining symmetric off-diagonal terms yields Equation (8). All cross-coordinate interactions remain; element-wise control does not make the loss separable. Attenuation may remove useful alignment or alter an interaction, so it need not reduce loss. For coordinate-dependent learning rates, use displacements $-\eta_{t,i}\omega_{t,i}\widehat{u}_{t,i}$ and match their norm, not merely the normalized-update norm.

The common base point also matters: lower loss than Native establishes a relative benefit, not necessarily lower loss than the weight-decay-only state. Block effects are conditional on the other updates and need not add. The candidate-loss evaluations establish these effects directly, without estimating a Hessian or deriving the EMA scale factor from a local loss model.

C Experimental Protocols and Outcome Definitions

C.1 Training contracts

The GPT-2 models have 12/24/36/48 blocks and widths 768/1024/1280/1600 for the 124M/355M/774M/1.5B settings. They use 12/16/20/25 attention heads, respectively. The actual 355M parameter count is 354,871,296. All use sequence length 1024 and global batch 480 across eight GPUs; per-GPU microbatch times accumulation is 60×1, 30×2, 15×4, and 12×5, respectively. The nominal 16K-step budget processes 7.86432B tokens. The schedule has 2000 warmup steps followed by the prefix of a 50K-step cosine schedule with a learning-rate floor one tenth of the peak. The 16K endpoint is therefore an equal-budget comparison, not the end of the scheduled decay. GPT-2 uses the repository's tokenized OpenWebText fallback data; it is not presented as an independently reconstructed canonical OpenWebText benchmark. FP16 uses native dynamic loss scaling; optimizer states and relevant reductions retain higher precision.

Qwen3-0.6B runs use BF16, peak LR 10^{-3}, the same 16K-step budget, and the existing reconstructed data pipeline. Its data provenance is `openwebtext_fallback_gpt2_roundtrip_reconstructed`: the fallback GPT-2 token stream is decoded and retokenized for Qwen. Qwen3-1.7B uses FP16, the same peak LR, betas, warmup, 16K budget and 50K-step schedule, and the same reconstructed Qwen-tokenized data provenance. Each Standard–dual pair shares its random initialization, and strict final validation

uses the same fixed 8K-sequence set within this 1.7B campaign. Absolute validation loss is not compared between architectures/tokenizers. The Qwen experiments test Standard against dual control within each setting.

C.2 Frozen full-run detector and episode grouping

For raw global training loss L_t, the detector uses the median $\widetilde{L}_t$ and median absolute deviation of the preceding 100 finite observations. The robust score is

$$z_t = \frac{L_t - \widetilde{L}_t}{1.4826\,\mathrm{MAD}_t + 10^{-12}}. \tag{46}$$

A candidate simultaneously satisfies $z_t \geq 8$, $L_t - \widetilde{L}_t \geq 0.15$, and $L_t \geq 1.10\widetilde{L}_t$. Emitted candidates are separated by more than 20 steps, with the history continuing to update at every finite observation. Consecutive emitted candidates at most 200 steps apart form one recovery episode. The episode ends 200 steps after its last candidate, or at the run end. The window can therefore extend more than 200 steps from its onset. For episode e with onset baseline B_e, define

$$R_e = \frac{\max_{t\in e} L_t}{B_e}, \qquad A_e = \sum_{t\in e} \max(L_t - B_e, 0). \tag{47}$$

Post-warmup episodes have onset after step 2000, and post-warmup cumulative damage is the sum of their A_e. Mean per-event severity and cumulative per-run burden answer different questions. Ratios are formed per event before averaging. A method with zero detected post-warmup episodes has zero post-warmup episode area by this definition; it can still have smaller subthreshold loss fluctuations. A nonfinite loss is recorded as a numerical failure, not silently replaced by a finite value for area or validation calculations.

C.3 Shared-state replay and probe timing

The 19-onset cohort contains eight, six, and five onsets for seeds 1234, 2027, and 3407. It also includes nine ordinary controls and three recovery controls. The candidate threshold matches the full-run detector, while nearby replay candidates are merged within a 200-step window. The 19 saved onset states form a selected replay cohort distinct from the grouped full-run episodes. Restoration audits compare parameters, optimizer state, random-number state, data position, and batch hashes. The initial no-recruit and radial adaptive norms agree to a relative difference below 3.28×10^{-8}; decoupled weight decay is shared.

The *immediate* matched-minibatch loss is evaluated after the candidate trigger displacement and before a further branch update. The first stored fixed-probe point is evaluated after one additional branch update. Consequently, equality of immediate onset-only and sustained-clipping loss does not require equality at the first stored probe point. The 32-point propagation comparison uses those same probe times in every branch. Longer branch studies use separately specified 64-, 128-, or 256-attempt windows; these are not pooled as interchangeable endpoints. Native FP16 skips remain part of each branch's actual trajectory.

C.4 Seed replication and endpoint provenance

The independent training replicates are the three seeds. Onsets, probe batches, randomized directions, and repeated branch orders are nested measurements, not extra independent training runs. We report paired outcomes and descriptive summaries, not event-level population-significance claims. Strict final validation requires a logged evaluation at step 16,000. Legacy loss trajectories that reach 16K without that evaluation contribute to event accounting but not to the strict endpoint mean. Endpoint recovery or matched reruns do not replace the original trajectory's spike statistics.

D Complete Beta, Learning-Rate, and Scale Results

D.1 The six-setting moment sweep

Table 1 gives the complete numerical comparison. Figure 7 retains all 18 raw, unsmoothed training-loss trajectories. Holding $\beta_1 = 0.9$, larger β_2 increases the available coordinate capacity and is

associated with more severe realized excursions. Across other changes in beta, capacity remains insufficient to determine event frequency or final quality.

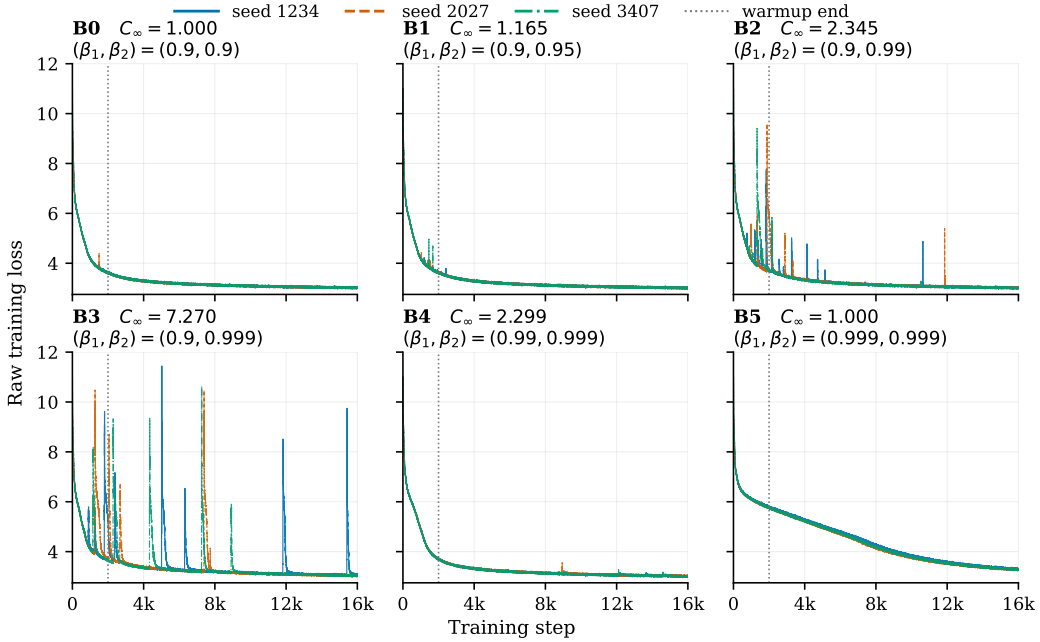


Figure 7: **Raw loss for the complete 355M B0–B5 sweep.** Each panel contains seeds 1234, 2027, and 3407 on identical axes, without smoothing. The dotted vertical line marks the end of the 2K-step warmup. Table 1 reports the pooled counts, all-episode severity, and strict 16K validation endpoints.

Strict 16K validation for $(.9, .9)$, $(.9, .95)$, $(.9, .99)$, $(.99, .999)$, and $(.999, .999)$ is respectively 2.968 ± 0.002, 2.961 ± 0.003, 2.964 ± 0.007, 2.966 ± 0.009, and 3.224 ± 0.001, each from seeds 2027 and 3407 ($n = 2$). Conventional $(.9, .999)$ uses three endpoints, 2.996 ± 0.015, with seed 1234 supplied by the matched passive-telemetry rerun. Episode counts still use the three original trajectories; endpoint recovery does not replace their event records.

D.2 Learning rate under matched and conventional moments

The two grids contain 24 complete loss trajectories in total, including shared baseline settings reused from the beta sweep. Figure 8 shows all 24 original, unsmoothed training-loss trajectories. The matched grid uses peak LR $\{0.5, 1, 2, 4\} \times 10^{-3}$, and the conventional grid uses $\{0.25, 0.5, 1, 2\} \times 10^{-3}$. We do not imply that the unequal endpoint settings are crossed. Table 3 reports both counts and all-episode severity. Counts need not be monotone in learning rate; matched moments at the highest rate have more but smaller excursions than at 2×10^{-3}. Final validation worsens in both high-rate settings despite different event burdens.

For matched moments at 10^{-3}, the strict validation mean uses seeds 2027 and 3407. The canonical comparison does not fold the separately recovered seed-1234 endpoint into the original trajectory. For conventional moments at 10^{-3}, seed 1234's strict endpoint comes from a matched passive-telemetry Standard rerun; the original sweep still supplies its spike statistics. Thus counts and endpoint means have explicit provenance, and the latter are not a per-trajectory mediation analysis.

D.3 Spike burden rises with scale under the fixed stress contract

Under the fixed contract, the number of detected Standard-run spikes grows from 10 at 124M to 18 at 355M and 23 at 774M across three complete seeds (Table 4). Exact tail audits show the same ordering: the probability that a monitored state has $\max_i |\widehat{u}_i| > 4$ is 0.0244, 0.0410, and 0.0584. This tail telemetry comes from one detailed seed per scale and is supporting evidence, whereas burden rates aggregate three seeds. At 1.5B, two runs terminate with non-finite loss and the

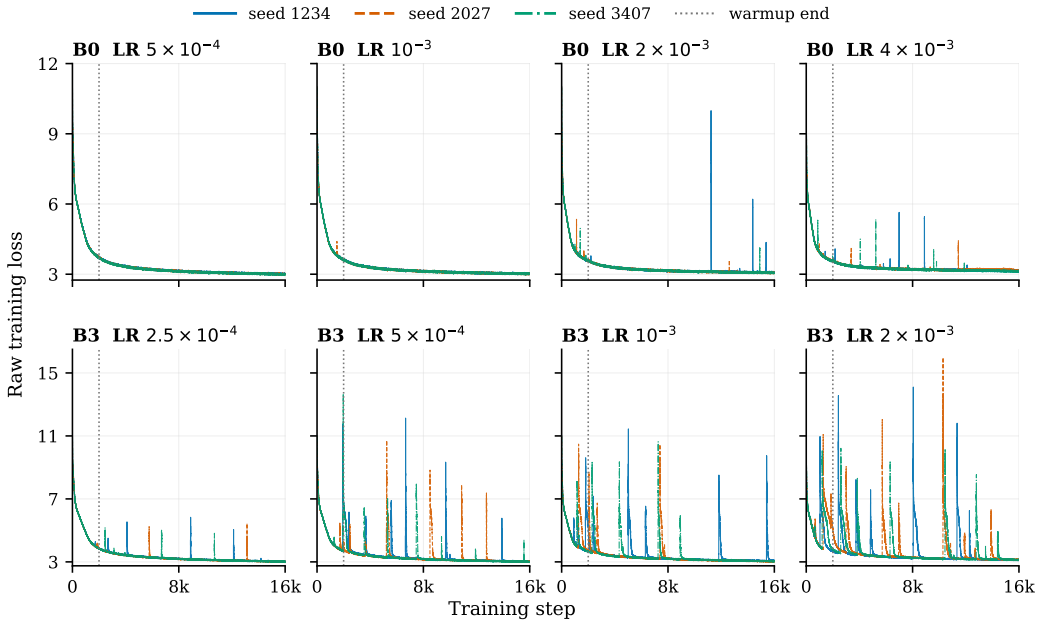


Figure 8: **Raw loss across the complete 355M learning-rate grids.** The upper row uses matched moments ($B0$), and the lower row uses conventional moments ($B3$). Each panel shows the original pre-update training loss for seeds 1234, 2027, and 3407 through 16K steps, without smoothing. The dotted line marks the end of the 2K-step warmup. All panels within a row share a vertical scale; the $B3$ row extends higher to include its largest spikes. The 10^{-3} trajectories also appear in Figure 7.

Table 3: **Complete 355M learning-rate grid at $\beta_1 = 0.9$.** B0 uses $\beta_2 = 0.9$ and B3 uses $\beta_2 = 0.999$. All/post-warmup counts use the strict 10% detector and pool three complete seeds. As in Table 1, the Area per spikes and Spike loss peak columns report the sample mean ± sample SD over all detected episodes, including warmup; seeds are the independent training replicates. A dash denotes no event, and N/A denotes one event. Validation is the strict 16K mean ± sample SD ($n = 2$ for B0 at peak LR 10^{-3}; $n = 3$ otherwise).

Run and peak LR	C_∞	All spikes	Post-warmup spikes	Area per spikes	Spike loss peak	16K validation
B0, 0.5×10^{-3}	1.000	0	0	–	–	2.953 ± 0.005
B0, 1.0×10^{-3}	1.000	1	0	$0.948 \pm$ N/A	$4.402 \pm$ N/A	2.968 ± 0.002
B0, 2.0×10^{-3}	1.000	7	5	22.375 ± 40.501	5.510 ± 2.155	3.017 ± 0.010
B0, 4.0×10^{-3}	1.000	12	11	5.934 ± 4.239	4.474 ± 0.782	3.093 ± 0.019
B3, 0.25×10^{-3}	7.270	9	9	30.417 ± 13.095	5.180 ± 0.396	2.970 ± 0.004
B3, 0.5×10^{-3}	7.270	24	19	182.466 ± 206.973	7.014 ± 2.810	2.969 ± 0.009
B3, 1.0×10^{-3}	7.270	18	13	310.869 ± 204.102	8.231 ± 2.105	2.996 ± 0.015
B3, 2.0×10^{-3}	7.270	25	21	385.620 ± 269.355	8.789 ± 3.274	3.080 ± 0.011

third was stopped after those divergences established that the fixed contract was outside the reliable training regime.

Figure 9 reports a separate analysis using the strict 10% detector and observed finite post-warmup exposure. Its pooled episode rates are 1.90, 3.10, 4.05, and 4.79 per 10K steps for 124M, 355M, 774M, and 1.5B; corresponding excess-loss-area rates are 820, 964, 1,210, and 2,100. The strict 1.5B count is 13 over 27,149 observed post-warmup steps, whereas Table 4 retains 20 episodes under the broader legacy detector. These counts belong to different protocols and should not be combined.

The completed rows use three seeds per scale; the exact-tail observer uses one detailed seed per scale. The ordering concerns episode burden under a fixed training contract, while mean area per

Table 4: Standard Adam scale study under the same global batch, learning rate, and 16K-step budget. Counts use the detector in Table 1 and pool three seeds. Unlike that table, post-warmup severity is the descriptive sample mean ± sample standard deviation across individual post-warmup spikes; the three seeds are the independent training replicates. Strict 16K validation uses all three seeds for 124M–774M; the 355M seed-1234 endpoint is supplied by the matched passive-telemetry Standard rerun described in Table 1. The 774M row uses exactly the same three Standard trajectories as Table 2. The 1.5B row is right-censored: seeds 1234 and 2027 diverge at steps 10,914 and 9,985, while seed 3407 was stopped at step 12,252 after divergence had been established. Its spike counts are therefore observed-prefix lower bounds, its severity statistics pool the 20 observed post-warmup spikes, and neither final validation nor the exact-tail probability is reported. The final column otherwise gives the monitored-state probability that the maximum normalized update exceeds four.

Model	All spikes	Post-warmup spikes	Area per spikes	Spike loss peak	16K final validation / status	$\Pr(\max_i \lvert\widehat{u}_{t,i}\rvert > 4)$
124M	10	8	430.548 ± 249.128	8.330 ± 1.557	3.189 ± 0.007	0.0244
355M	18	13	311.536 ± 188.532	8.362 ± 2.123	2.996 ± 0.015	0.0410
774M	23	17	298.845 ± 217.366	8.632 ± 3.165	2.916 ± 0.030	0.0584
1.5B	≥ 25	≥ 20	287.671 ± 332.029	9.415 ± 3.096	2 diverged; 1 stopped	–

episode is not monotone. At 1.5B, two seeds diverge before 11K and the third stops administratively at 12,252. Their observed-prefix rates are descriptive; the comparison neither isolates a causal effect of parameter count nor supplies a complete 16K scaling point.

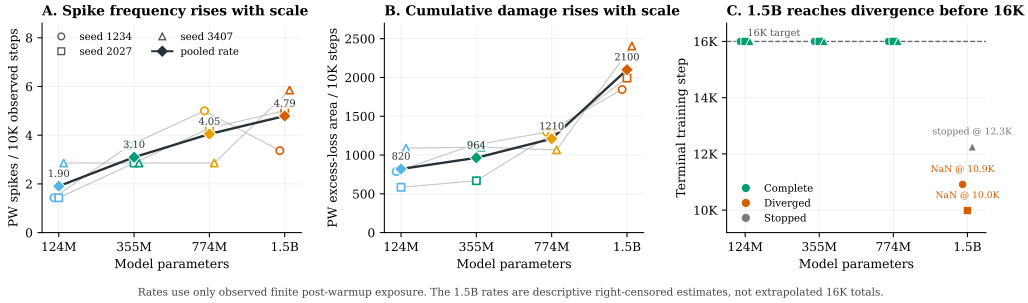


Figure 9: Spike burden under the fixed training contract and strict 10% detector. Hollow markers are individual seeds and filled diamonds are exposure-pooled rates; Table 4 uses the broader legacy detector. (A) Post-warmup spike frequency rises monotonically with model size. (B) Cumulative excess-loss area per 10K observed post-warmup steps shows the same ordering, even though mean area per individual spike is not monotone. (C) All 124M–774M runs complete 16K steps, whereas two 1.5B seeds terminate with non-finite loss and the remaining seed is administratively stopped at step 12,252. The 1.5B rates use only observed finite prefixes and are descriptive right-censored estimates, not extrapolated 16K totals.

E Immediate Interventions and Additional Controls

E.1 Recruitment, Direction, and Dose Response

Recruitment records new entries into a tail, not its net growth. At threshold 0.45, the median newly entered fraction is 4.02075%, while the median net increase is 1.54002%; 18/19 events have positive net growth. At threshold one, 17/19 have positive net growth. Numerator log contributions are positive and denominator contributions negative for recruited coordinates at all 19 onsets. This does not establish that denominator history is irrelevant: it only identifies the immediate decomposition of the selected coordinates' growth.

No-recruit's median paired immediate advantage over equal-norm radial scaling is 6.597 pp at threshold 0.45, with within-seed medians 8.681, 2.333, and 7.455 pp. At threshold one, it is 3.174 pp and positive at 17/19 onsets. Over 32 stored probe points, the threshold-0.45 peak advantage is positive at 17/19, with median 10.142 pp. For sustained no-recruit replay, the initially large set is fixed at

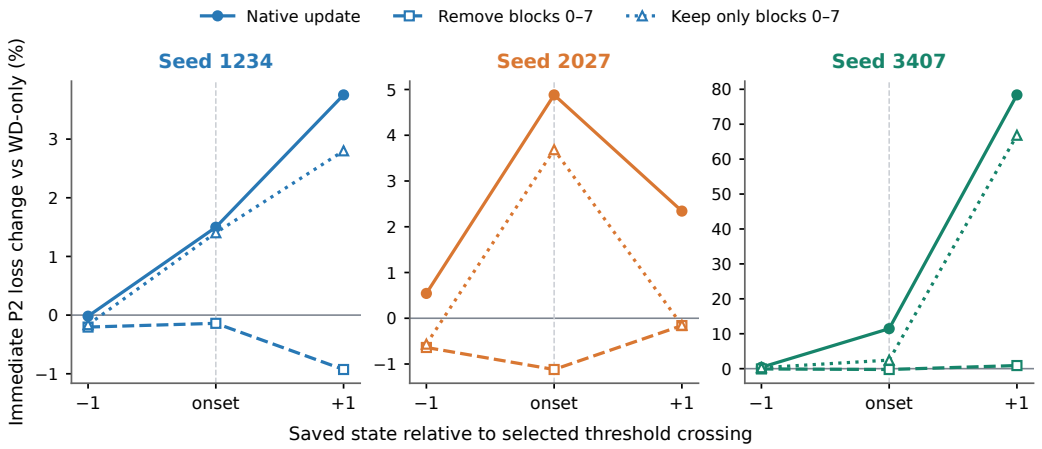


Figure 10: **Front-block updates contribute to immediate loss at selected 355M states.** Blue, orange, and green denote seeds 1234, 2027, and 3407; solid circles use the native update, dashed squares remove adaptive updates in blocks 0–7, and dotted triangles retain only those blocks. Each panel shows three nearby saved states from one reused event, aligned to a retrospectively selected fixed-probe threshold crossing. Values are the percentage change in fixed-probe P2 loss relative to applying weight decay without the adaptive update (gray horizontal line). Each panel has its own vertical scale. The nine points represent three events, not nine independent spikes. These are same-state, immediate interventions, not sustained training controls; matched-norm random controls are discussed in the text.

onset. Each radial branch matches the no-recruit transform evaluated at its own current state. The multistep comparison is a policy comparison after trajectories diverge; it does not maintain equal norms across independent branches. The earlier direction–radius crossed experiments use a separate cohort and remain supporting geometry evidence, not additional independent onsets.

The dose-response sweep holds the update direction fixed and modifies the adaptive displacement, not weight decay. All 19 positive-range curves are monotone over 0.25–2 times the native step. Zero adaptive displacement is not part of the universal monotonic statement; 16 quarter steps improve over zero. Later nine-event static experiments also find useful small positive displacements, supporting the same distinction between direction and finite step length.

E.2 Coordinate and module specificity

In an earlier 11-event cohort, deleting the largest 256 coordinates does not consistently stop propagation; the median peak improvement is -0.003545 in the stored loss-difference metric. It is therefore not supported to attribute most spike damage to a handful of maximal coordinates. The cohort and protocol differ from the 19-onset study and are not pooled.

A module screen evaluates three threshold-selected states in each of three reused events (Figure 10). Removing updates in the first eight Transformer blocks lowers immediate fixed-probe loss relative to the native update at all nine states; retaining only those blocks can itself produce substantial damage. The intervention does not always erase the loss increase: at seed 3407's last state, the removal branch remains 0.89% above the weight-decay-only reference. These nine states are not nine independent events. Random controls match coordinate count and remaining adaptive norm, but do not exactly match removed energy. Subsequent front-module, radial, and direction-replacement branches have event-dependent outcomes. The finding identifies a useful intervention location in the selected models and states, not a universal anatomical origin of AdamW spikes.

E.3 A single directional pulse versus continued control

A formal nine-event campaign includes three previously used discovery events and six later events, with native, radial, selected-direction, and random-direction controls. For the six later events, the selected pulse reduces cumulative damage relative to native and radial controls at 6/6, with median

relative changes -5.52% and -8.16%, respectively. It beats the random-direction control at only 3/6, and improves the peak relative to native at 4/6. In all three discovery events, cumulative damage is worse than native. These outcomes support testing the full subsequent trajectory, rather than promoting an immediate directional benefit into a generally effective mitigation rule.

For the 19-onset clipping comparison, median fixed-probe peak ratios are 2.583 for Standard, 2.488 for onset-only clipping, and 1.063 for sustained clipping. Onset-only clipping improves peak loss at 10/19 and cumulative excess at 9/19. Sustained clipping improves peak loss over onset-only at 19/19. Restricting to the ten events with no native skip in any of the three compared branches retains a 10/10 advantage. The effect is therefore not entirely attributable to FP16 skip differences, although full-window results include the actual numerical paths.

E.4 Additional blockwise and history-allocation validation

Table 5: **History allocation at six validation events.** Loss advantages are in nats/token. Each seed contributes the mean of its two event-level paired differences, with P1/P2 averaged within each event; uncertainty is the sample SD across three seed means.

Equal-budget control	Mean advantage $\pm$ seed SD	Positive events
Uniform	0.115 ± 0.042	6/6
Current magnitude	0.065 ± 0.053	6/6
Shuffled history	0.119 ± 0.034	6/6

The additional campaign used three discovery events to establish the workflow, then six validation events from the same three training seeds, with new fixed probes. The six events are distinct episodes, not six independent training runs. Block 0 has positive benefits relative to both Native and its equal-norm radial control at all six validation events. For the fixed 10% adaptive-energy reduction, the supplied event-level mean loss advantages of history-based allocation over uniform, current-magnitude, and shuffled-history allocation are respectively 0.115, 0.065, and 0.119 nats/token. All three comparisons favor history at 6/6 events. Advantages over these equal-budget controls are near zero in three ordinary states; this is not a claim that the forced 10% reduction is harmless relative to Native at every ordinary state.

The source response table contains both fixed probes (P1/P2) for each validation event and all four allocations. We average paired probe differences within each event, then the two events within each seed. Table 5 reports the means and sample SDs across the three seed means; each seed mean is positive. The six events were previously used in other diagnostics; their probes and repeated allocations are nested observations.

The frozen manifest defines a common 10% reduction in adaptive squared norm and a history reference from the preceding 128 successful raw updates. For tensor energy $E_\ell = \|d^{(\ell)}\|_2^2$ and nonnegative score q_ℓ, the retained factor is

$$f_\ell = \sqrt{1 - \min\{\nu q_\ell, 1\}}, \qquad \sum_\ell E_\ell \min\{\nu q_\ell, 1\} = 0.1 \sum_\ell E_\ell. \tag{48}$$

A scalar bisection chooses ν from this energy constraint, without using candidate loss. Scores are one for uniform allocation, current raw mean absolute update for current-magnitude allocation, and its ratio to the historical median plus FP32 epsilon for historical allocation. Shuffled-history controls permute the last scores among tensors with the same parameter suffix and exact shape. The earlier three-marker experiment below uses a different weighting rule and is reported separately. The six-event diagnostic provides evidence for allocating attenuation using relative history; it does not measure an online EMA controller or establish its preferred threshold.

Exploratory single-tensor response. A separate fixed-dose diagnostic kept one selected high-growth tensor at fractions $0, 0.25, 0.5, 0.75, 1$ of its native update. Keeping 25% reduced immediate loss by a reported mean 0.215 nats/token, whereas deletion reduced it by 0.271. Under the prespecified near-optimal criterion, all 12 validation probe comparisons selected zero retained update. Partial attenuation therefore helped, but the tested grid did not identify a near-optimal nonzero fraction. This diagnostic is not used to derive soft scaling, $\kappa = 2$, or the EMA implementation. Its selected-tensor, single-step objective also differs from the full-training controller.

A descriptive threshold check. For the ratio of current raw mean-absolute update to the median over the previous 128 successful updates, the supplied $G > 2$ summaries are 0.23% of tensors and 0.00067% of adaptive-update energy in the three ordinary states, versus 39.3% and 54.2% in the six validation events. These finite state comparisons do not establish an optimal threshold or a population false-positive rate. They are not treated as a derivation of EMA-2.

E.5 Earlier three-marker history-allocation diagnostic

A static experiment removes 10% of adaptive squared norm. It compares uniform tensor scaling, allocation based on current tensor magnitude, and allocation weighted by the positive part of the log ratio between current mean absolute update and its preceding 128-successful-update median. At the three discovery markers, historical allocation lowers immediate third-probe loss more than either control; the smaller paired advantage is 0.66%, 0.69%, and 5.31%, respectively. This supports historical scale as a useful reference at those states. It does not establish that a fixed-budget history policy works over complete training. The final algorithm uses a lagged EMA, threshold-based scaling, and coordinate clipping; it is not this static allocation rule.

E.6 Exploratory sustained policies, withdrawal, and ordinary states

Nine later control windows test fixed-budget tensor-history and coordinate policies for 64 attempts, then remove them. Most windows are observed for 128 attempts; the first discovery window is extended to 256. Their endpoints must therefore retain the different horizons. Coordinate selection can maintain stability during control in the three discovery events, with rebound after withdrawal in two. Random controls can also be effective. An exploratory history-plus-top-coordinate policy harms ordinary reference windows: third-probe peak ratios are 2.318, 1.246, and 2.303, compared with native 1.041, 1.001, and 1.002. In one later event its peak is 5.888 versus native 2.900. Table 11 reports all third-probe arm outcomes for these windows.

Ordinary references are selected by available metadata, not exact parameter-state matching; one seed's initial first-probe loss differs by approximately 14.7%. The later events are temporally distinct but have already appeared in other diagnostics, so they are not fresh blind tests. These restrictions prevent a claim of universal specificity. They also prevent conflating this fixed-budget exploratory combination with the final EMA-plus-cap algorithm evaluated in complete training.

An earlier feedback campaign completed 22 review windows and seven of 13 feedback windows before being stopped. The eighth was interrupted and is excluded; six missing or partial windows, including the intended ordinary feedback controls, are not treated as completed evidence. Open-loop gradient-tape and endogenous-gradient branches can have different outcomes, but this incomplete campaign does not establish a population-level feedback decomposition.

F Optimizer Memory and Natural Event Dynamics

F.1 Exact compensated-state construction

After a common native update, let n be the tensor's successful-update age. For a multiplier $a > 0$, define

$$v' = a^2 v, \qquad D = \sqrt{v/(1-\beta_2^n)} + \epsilon, \qquad D' = \sqrt{v'/(1-\beta_2^n)} + \epsilon, \qquad m' = m\frac{D'}{D}. \tag{49}$$

All operations are coordinatewise; beta, epsilon, and update age are unchanged. Then

$$\frac{m'/(1-\beta_1^n)}{D'} = \frac{m/(1-\beta_1^n)}{D}. \tag{50}$$

With nonzero epsilon this is more precise than simply using $m' = am$. The implemented output-preservation gate uses absolute tolerance 10^{-7} and relative tolerance 2×10^{-6}. The next actual clipped gradient is checked by a full-coordinate hash. Subsequent trajectories receive their own gradients; only the first post-intervention input is guaranteed common. The original scale study includes $a = 0.8, 1, 1.25$. The $a = 1.25$ branch motivates the response-dissection comparison shown in the main text.

F.2 Response hybrids retain branch-specific dynamics

Within each branch, native and compensated moment states receive the same current clipped gradient. Hybrid policies combine one candidate's module norm with the other's direction, or apply the compensated response only to the first eight blocks or its complement. These are within-state candidate matches; they do not enforce equal norms along independent future trajectories. Table 6 reports all event arms.

Table 6: 128-attempt fixed-probe peak ratios after the common native first step. Three reused discovery windows; ratios use each window's shared initial probe loss. Native skips are retained.

Response policy	Seed 1234	Seed 2027	Seed 3407
Native	2.597	1.482	3.112
Compensated memory	1.000	.999	1.000
Compensated norm / native direction	1.000	1.000	2.609
Native norm / compensated direction	2.487	.999	3.098
Compensated front blocks	1.000	1.256	1.001
Compensated complement	2.135	.999	2.756

Neither size nor direction alone reproduces the compensated trajectory in every event. For seed 3407, both hybrid policies eventually develop large excursions, some delayed by 10–11 steps. In seed 2027, replacing only the front response delays rather than removes the rise. Later-phase reference windows have native and compensated peak ratios near one, but are not training-stage-matched controls. At the first common post-intervention input, full-model norm ratios are approximately 0.956–0.961 and direction changes 5.36–5.72 degrees for events; later references have similar initial changes. Those two numbers alone therefore do not identify event states.

The earliest native skips in any branch occur at absolute steps 2306 and 2151 for seeds 1234 and 3407; seed 2027 has none. Before the first skip in any branch, native/compensated normalized cumulative excess is 0.832/0 and 5.660/0.001 for the first and third seeds. Protection therefore appears before skip divergence, while the full 128-attempt comparison remains a policy total effect that includes later numerical behavior.

F.3 Natural trajectories and local source decomposition

The native first-moment output can be decomposed into a historical term and a new-gradient term using the actual common denominator:

$$\widehat{u}_t = h_t + n_t, \quad h_t = \frac{\beta_1 m_{t-1}}{(1-\beta_1^n)D_t}, \quad n_t = \frac{(1-\beta_1)g_t}{(1-\beta_1^n)D_t}. \tag{51}$$

Its squared norm is $\|h_t\|^2+\|n_t\|^2+2\langle h_t, n_t\rangle$; the cross term can be negative. Contributions therefore cannot be inferred by adding two positive norms. Natural traces show a short new-gradient pulse followed in some windows by persistent historical output. In seed 2027, one step after a pulse the new-gradient norm is about 0.957 of its early baseline while the historical norm remains 1.956 and the local total 1.808. This is an algebraic trajectory description, not causal mediation of a unique initiating source.

Simple local magnitude statistics are not sufficient. In the seed-2027 native and native-norm/compensated-direction branches, front-block maximum relative norms are 1.970 and 2.114 and relative maximal coordinates are 2.679 and 3.208, yet the latter branch stays stable. The parameter states have already diverged, so this is a boundary on scalar prediction, not another same-state direction experiment. Results depend somewhat on the early-window baseline and the threshold used to define the first rise.

Virtual input perturbations at fixed moments produce much larger responses for random directions than for gradient-aligned scaling. The gain uses the actual local relative input change; a global nominal perturbation is not a valid local denominator. Responses also occur in later references, and two tested amplitudes do not establish one linear regime. These are finite perturbation diagnostics, not measured Jacobian norms, validated warning signals, or closed-loop feedback gains.

G Algorithm Details and Evaluation Boundaries

Algorithm 1 Dual-Scale Update Control at one successful optimizer update

Require: Gradient g_t after native clipping; moments; tensor references b and counts n.
Update native moments and form the bias-corrected update $\widehat{u}_t$.
for each parameter tensor ℓ with a gradient **do**
  $r \leftarrow \mathrm{mean}_{i\in\ell} |\widehat{u}_{t,i}|$ ▷ raw, before clipping
  if $n^{(\ell)} = 0$ **then**
    $b^{(\ell)} \leftarrow r$
  end if
  $s \leftarrow 1$
  if $n^{(\ell)} \geq 128$ **then**
    $s \leftarrow \min\{1, \kappa b^{(\ell)}/(r + \varepsilon_r)\}$
  end if
  $u^{*(\ell)} \leftarrow s\ \mathrm{clip}(\widehat{u}_t^{(\ell)}, -c, c)$
  $\theta^{(\ell)} \leftarrow (1 - \eta_t\lambda)\theta^{(\ell)} - \eta_t u^{*(\ell)}$
  if $s = 1$ **then**
    $b^{(\ell)} \leftarrow \alpha b^{(\ell)} + (1 - \alpha)r$
  end if
  $n^{(\ell)} \leftarrow n^{(\ell)} + 1$
end for

A parameter tensor is the unit of historical scaling, not an entire Transformer block. Each tensor stores its own mean-absolute-update EMA and observation count. The first observation initializes the EMA. During the first 128 observations the tensor scale is one and the baseline updates normally. Thereafter scaling uses the previously stored baseline. A trigger is the implemented condition $s < 1$, including its numerical epsilon; the baseline is unchanged on that tensor's triggered steps. Counts advance at successful optimizer updates. Skipped FP16 steps do not advance the optimizer state. Checkpoint restoration includes the EMA and count, so resumed runs do not restart the history window.

The coordinate clip is computed from the bias-corrected adaptive update. The tensor reference is measured before this clip, and tensor scaling is applied to the clipped output. Weight decay is applied independently with the native group settings. The adaptive-output bounds in Equation (16) follow, once $n \geq 128$, from $s \leq 1$ and $\mathrm{mean}\,|\,\mathrm{clip}(\widehat{u}, -c, c)| \leq r$. If $s = \kappa b/(r + \varepsilon_r) < 1$, then $sr = \kappa b\, r/(r + \varepsilon_r) \leq \kappa b$; otherwise $r + \varepsilon_r \leq \kappa b$. The bound excludes the decoupled weight-decay displacement and does not bound loss change without additional geometric assumptions.

The guard's recurrence retains all gradients accepted by the native optimizer. It does not reset moments, discard a large-gradient minibatch, or require the loss detector used for retrospective evaluation. Component ablations use conventional moments. The additional $c = 0.45$ result is a sensitivity check, not a broad hyperparameter optimization. The method's scalar state is small relative to Adam's two per-coordinate buffers, but a clean wall-clock comparison is needed before asserting an efficiency or time-to-quality improvement.

The frozen reference also has a boundary: if a tensor's raw scale remains above κb indefinitely, its reference stays fixed and attenuation persists. This statistic alone cannot distinguish a persistent benign scale change from a damaging one. An exactly zero baseline can similarly suppress a later nonzero update. The stated recurrence has no positive baseline floor; adding one would change the evaluated rule.

H Transfer, Completion, and Per-Seed Results

The Qwen3-0.6B comparison consists of three paired 16K-step trajectories. Reported episode counts use the same frozen detector and episode grouping as the GPT-2 comparisons, rather than counts inferred from a plot. Table 8 gives the corresponding per-seed outcomes. For Qwen3-0.6B, the paired validation improvement is consistent across seeds; its size is not compared to GPT-2's because tokenizers and loss units differ.

Table 7: **Transfer and completion with conventional moments.** All rows use $(\beta_1, \beta_2) = (0.9, 0.999)$; Dual Control is Clip-0.3+EMA-2. Counts pool the indicated seeds under the frozen detector. The Area per spikes and Spike loss peak columns report the sample mean $\pm$ sample SD over all detected episodes. Final validation is mean $\pm$ sample SD across complete 16K-step seeds, or the individual value when only one seed completes. † marks statistics over right-censored observed prefixes, so the counts are lower bounds. Dashes denote unavailable outcomes, or undefined per-event statistics when no episode occurs; they never denote zero.

Model	Method	16K-complete runs	All / Post-warmup spikes	Area per spikes	Spike loss peak	16K final validation
NanoGPT 1.5B	Standard	0/3	$\geq 18/\geq 13^{\dagger}$	$424.306 \pm 288.077^{\dagger}$	$10.381 \pm 2.770^{\dagger}$	–
NanoGPT 1.5B	Dual Control	3/3	2/0	4.424 ± 1.482	4.179 ± 0.072	2.7747 ± 0.0046
Qwen3 0.6B	Standard	3/3	12/10	223.263 ± 287.970	8.331 ± 3.463	2.9462 ± 0.0119
Qwen3 0.6B	Dual Control	3/3	0/0	–	–	$\mathbf{2.9230 \pm 0.0017}$
Qwen3 1.7B	Standard	1/3	$\geq 7/\geq 3^{\dagger}$	$213.753 \pm 199.065^{\dagger}$	$10.666 \pm 3.843^{\dagger}$	$2.8132\ (n = 1)$
Qwen3 1.7B	Dual Control	3/3	0/0	–	–	$\mathbf{2.7973 \pm 0.0016}$

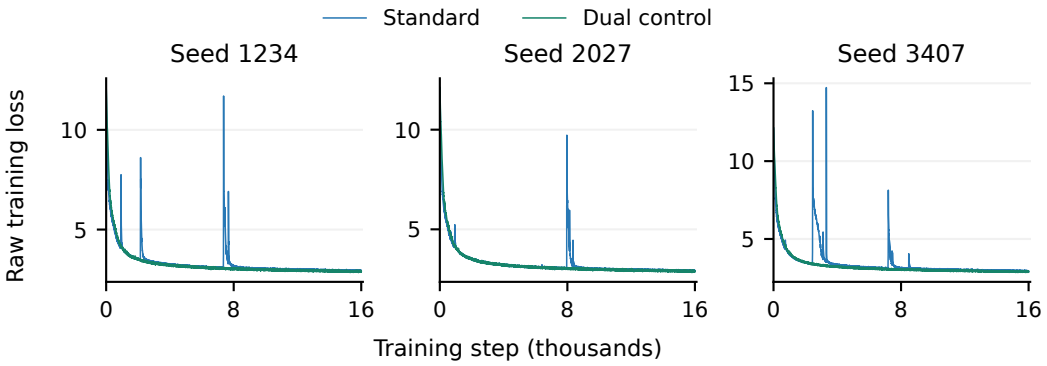


Figure 11: All three Qwen3-0.6B BF16 training seeds at peak LR 10^{-3}. Curves cover the full 16K steps and compare Standard with dual control. Each panel has its own vertical scale.

The Qwen3-1.7B Standard and dual-control runs use the same FP16 stress contract, conventional moments, and matched initialization within each seed. Table 9 distinguishes a terminated run from one that reached the 16K target. Standard seeds 2027 and 3407 produce non-finite loss; their event counts and post-warmup areas cover only finite observed prefixes. Neither has a strict 16K validation endpoint. Dual control reaches 16K on all three seeds, but a paired final-validation comparison exists only for seed 1234.

The guarded 1.5B paths all complete with zero post-warmup episodes; two warmup episodes remain in total. An auxiliary GradSkip threshold configured in parts of that study does not reject any accepted-path update. Ordinary dynamic-loss-scaler skips occur (7, 6, and 11 by seed) and are not that safeguard. The references diverge at completed steps 10,914 and 9,985, while seed 3407 stops administratively at 12,252. Their event counts are observed-prefix lower bounds, with no fabricated 16K loss or continuation after failure.

I Comparison with SPAM: Frequency and Damage

SPAM addresses gradient spikes using spike-aware clipping and momentum reset (Huang et al., 2025). Its benefit should be evaluated through both frequency and damage, as well as optimization

Table 8: Per-seed Qwen3-0.6B outcomes. Post-warmup area sums episode excess loss; all rows complete 16K steps.

Seed	Method	All / post-warmup	Post-warmup area	Final validation
1234	Standard	4 / 3	759.38	2.9449
1234	Dual control	0 / 0	0.00	2.9224
2027	Standard	3 / 2	541.34	2.9349
2027	Dual control	0 / 0	0.00	2.9217
3407	Standard	5 / 5	1308.22	2.9587
3407	Dual control	0 / 0	0.00	2.9250

Table 9: Per-seed Qwen3-1.7B FP16 outcomes. Terminal step is 16,000 for completed runs and the non-finite-loss step for divergent runs. Event counts and post-warmup area on divergent Standard paths are observed-prefix lower bounds; a dash denotes no strict 16K validation, not zero loss.

Seed	Method	Terminal step	All / Post-warmup spikes	Post-warmup area	16K final validation
1234	Standard	16,000	3 / 1	15.43	2.8132
1234	Dual Control	16,000	0 / 0	0	2.7963
2027	Standard	7,933	$\geq 2/\geq 1$	≥ 117.04	–
2027	Dual Control	16,000	0 / 0	0	2.7964
3407	Standard	4,706	$\geq 2/\geq 1$	≥ 125.96	–
3407	Dual Control	16,000	0 / 0	0	2.7992

quality and successful completion. Table 10 uses the same detector and episode accounting as the main experiments.

Our 774M implementation uses threshold 5000, reset interval 500, and 150-step reset-associated warmup, with dense moment storage and global bias-correction age. The comparison holds the base model, data, batch, FP16 contract, and external training schedule fixed, while SPAM adds its reset-associated learning-rate behavior. Alternative bias-age/reset and periodic-LR-only diagnostic branches are distinct implementations and are not pooled as independent seeds. The reported comparison is not a claim that all SPAM variants were optimally tuned.

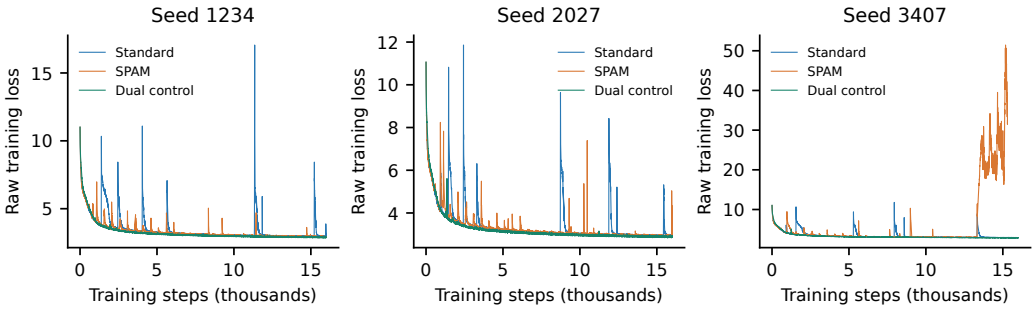


Figure 12: **SPAM can reduce damage even when detected episodes remain.** Complete raw loss curves for all three original seed paths. The third SPAM path becomes nonfinite at step 15,302; only finite observations are drawn, with its failure explicitly retained in Table 10. The resumed third-seed branch is not an extra independent replicate.

For the two completed SPAM runs, post-warmup counts increase from 7 to 13 and from 6 to 14, but cumulative damage decreases from 1812.60 to 451.52 and from 1775.05 to 424.69: reductions of 75.1% and 76.1%. Mean post-warmup event peak ratios also fall from 2.820 to 1.365 and from 2.474 to 1.441. This is clear attenuation of spike damage despite more detected episodes. Final validation improves in the first seed and worsens in the second; reduced damage alone does not ensure a quality improvement in every run.

Table 10: 774M SPAM comparison, including every original seed. Post-warmup area is cumulative per-run excess. The mean peak ratio averages event-wise peak/baseline ratios. Incomplete-prefix counts and areas are lower bounds; final validation is absent after failure.

Seed	Method	Post-warmup count	Post-warmup area	Mean peak ratio	Validation	Outcome
1234	Standard	7	1812.60	2.820	2.9510	16K
1234	SPAM	13	451.52	1.365	2.8858	16K
1234	Dual control	0	0.00	–	2.8595	16K
2027	Standard	6	1775.05	2.474	2.8956	16K
2027	SPAM	14	424.69	1.441	2.9273	16K
2027	Dual control	0	0.00	–	2.8511	16K
3407	Standard	4	1493.20	3.098	2.9013	16K
3407	SPAM	≥ 12	≥ 3912.83	2.178	–	Diverged
3407	Dual control	0	0.00	–	2.8568	16K

The third SPAM run diverges late, with at least 12 post-warmup episodes and 3912.83 observed cumulative post-warmup excess. Its incomplete prefix cannot be compared as if it were a completed 16K run or omitted from a three-seed assessment. A continuation from step 15,000 also diverges and is not counted as a new seed. We report both the substantial completed-run benefit and the failed path. These outcomes characterize the tested implementation and setting; they do not justify calling SPAM ineffective or claiming universal superiority of dual control.

J All Third-Probe Outcomes for Exploratory Sustained Controls

Table 11 retains all nine exploratory windows and every policy for the third fixed probe. Policies remove a fixed update-energy budget using uniform, current-scale, or historical tensor allocation and selected or random coordinate allocation. The source labels are kept here for unambiguous correspondence: `history+top` combines the exploratory historical allocation and selected-coordinate rule; it is not Dual-Scale Update Control. "Withdrawal peak" is measured after control removal at attempt 64. The first discovery window has a 256-attempt extension; other rows use 128. Any unavailable historical allocation follows the recorded native fallback and is not credited as activated control.

Table 11: All exploratory sustained-control outcomes on probe P3. P_H is the full-window peak ratio, P_{off} the withdrawal peak, and skip is the native skipped-attempt count.

Window / seed	Type	Policy	H	P_H	P_{off}	Skip
0 / 1234	Discovery	Standard	256	2.600	1.839	2
0 / 1234	Discovery	history	256	2.671	1.709	2
0 / 1234	Discovery	absolute	256	2.564	1.685	2
0 / 1234	Discovery	uniform	256	2.758	1.790	3
0 / 1234	Discovery	top	256	2.448	2.448	1
0 / 1234	Discovery	radial	256	2.247	1.584	1
0 / 1234	Discovery	random0	256	2.712	2.712	2
0 / 1234	Discovery	random1	256	2.435	2.435	2
0 / 1234	Discovery	history+top	256	2.825	2.825	2
1 / 2027	Discovery	Standard	128	1.469	1.243	0
1 / 2027	Discovery	history	128	1.000	0.976	0
1 / 2027	Discovery	absolute	128	1.354	1.206	0
1 / 2027	Discovery	uniform	128	0.999	0.975	0
1 / 2027	Discovery	top	128	1.570	1.570	0
1 / 2027	Discovery	radial	128	1.473	1.473	0
1 / 2027	Discovery	random0	128	1.472	1.472	0
1 / 2027	Discovery	random1	128	1.572	1.572	0
1 / 2027	Discovery	history+top	128	1.454	1.454	0
2 / 3407	Discovery	Standard	128	3.122	1.908	2
2 / 3407	Discovery	history	128	3.209	3.209	1

Window / seed	Type	Policy	H	P_H	P_{off}	Skip
2 / 3407	Discovery	absolute	128	1.001	0.996	0
2 / 3407	Discovery	uniform	128	2.569	2.418	1
2 / 3407	Discovery	top	128	1.012	1.012	0
2 / 3407	Discovery	radial	128	1.001	0.999	0
2 / 3407	Discovery	random0	128	1.003	1.003	0
2 / 3407	Discovery	random1	128	2.643	2.643	2
2 / 3407	Discovery	history+top	128	2.755	2.755	2
3 / 1234	Reference	Standard	128	1.041	1.041	0
3 / 1234	Reference	history+top	128	2.318	2.318	1
4 / 2027	Reference	Standard	128	1.001	0.945	0
4 / 2027	Reference	history+top	128	1.246	1.246	0
5 / 3407	Reference	Standard	128	1.002	0.994	0
5 / 3407	Reference	history+top	128	2.303	2.303	2
6 / 1234	Later	Standard	128	2.216	1.681	0
6 / 1234	Later	history	128	1.001	0.998	0
6 / 1234	Later	absolute	128	2.454	1.728	0
6 / 1234	Later	uniform	128	1.950	1.562	0
6 / 1234	Later	top	128	1.001	0.998	0
6 / 1234	Later	radial	128	1.001	0.998	0
6 / 1234	Later	random0	128	1.003	1.003	0
6 / 1234	Later	random1	128	1.001	1.000	0
6 / 1234	Later	history+top	128	1.002	1.002	0
7 / 2027	Later	Standard	128	4.563	2.436	2
7 / 2027	Later	history	128	1.000	1.000	0
7 / 2027	Later	absolute	128	1.001	0.999	0
7 / 2027	Later	uniform	128	1.001	1.000	0
7 / 2027	Later	top	128	1.000	0.999	0
7 / 2027	Later	radial	128	1.000	0.999	0
7 / 2027	Later	random0	128	1.000	0.999	0
7 / 2027	Later	random1	128	1.000	0.999	0
7 / 2027	Later	history+top	128	1.000	0.999	0
8 / 3407	Later	Standard	128	2.900	2.204	1
8 / 3407	Later	history	128	1.001	1.001	0
8 / 3407	Later	absolute	128	3.007	2.355	1
8 / 3407	Later	uniform	128	3.408	2.255	2
8 / 3407	Later	top	128	1.000	0.999	0
8 / 3407	Later	radial	128	1.000	0.999	0
8 / 3407	Later	random0	128	0.999	0.999	0
8 / 3407	Later	random1	128	1.000	1.000	0
8 / 3407	Later	history+top	128	5.888	2.179	2

K Additional Paired Results and Reporting Conventions

K.1 Recruitment at three thresholds

Median new-entrant fractions at thresholds 0.3, 0.45, and 1 are 7.967%, 4.021%, and 0.571% at onsets, versus 2.439%, 0.516%, and 0.0033% at ordinary controls. Using unrounded fractions, the enrichment factors are 3.27, 7.80, and approximately 174.

At $\tau = 1$, the median numerator and denominator log contributions are $+2.026$ and -0.118, respectively. These describe selected coordinates; they do not remove the role of denominator history in setting their scale.

K.2 Same-state loss and later probe loss are different endpoints

For immediate candidate loss L_a^{imm}, the benefit relative to Standard is $100(L_{\text{Standard}}^{\text{imm}} - L_a^{\text{imm}})/L_{\text{Standard}}^{\text{imm}}$. A paired difference is calculated within each onset before taking its median. Thus the difference of two reported medians need not equal the median paired difference.

Table 12: Immediate no-recruit advantage over radial scaling at $\tau = 0.45$, in percentage points of Standard's matched-batch loss.

Seed	Onsets	Median paired advantage	Positive onsets
1234	8	8.681	8/8
2027	6	2.333	6/6
3407	5	7.455	5/5
Pooled	19	6.597	19/19

For fixed-probe values $P_{a,h}$ and shared pre-branch loss P_0, the 32-step peak benefit is $100(\max_h P_{\mathrm{Standard},h} - \max_h P_{a,h})/P_0$. The first stored probe follows an additional update, so it must not be renamed an immediate matched-minibatch outcome. The median paired no-recruit/radial peak advantage at $\tau = 0.45$ is 10.142 pp and is positive at 17/19 onsets. At $\tau = 1$, it is positive at 16/19. These comparisons include subsequent trajectory changes. The nine ordinary controls have median effects within 0.04 pp of Standard for the three clipping/EMA policies. Small probe-loss changes there do not mean that the policies never alter ordinary updates.

Table 13: Sustained replay benefits at 19 onsets. Values are median paired percentage-point reductions relative to Standard on fixed probes. Positive counts concern 32-step peak benefit.

Intervention	First stored probe	32-step peak	Positive onsets
Radial, $\tau = 0.45$	2.883	35.987	18/19
No-recruit, $\tau = 0.45$	4.815	47.164	18/19
Radial, $\tau = 1$	1.753	18.776	18/19
No-recruit, $\tau = 1$	3.007	52.280	17/19
Clip-0.45	7.809	110.031	19/19
EMA-2	7.564	147.810	19/19
Clip-0.45+EMA-2	9.528	155.395	19/19

K.3 Immediate displacement multipliers

Table 14: Median immediate loss benefit relative to the native multiplier one, in percentage points of its matched-minibatch loss.

Multiplier	0	.25	.5	.75	1	1.25	1.5	2
Benefit	24.43	25.73	24.02	8.83	0	−9.33	−15.59	−28.52

Every tested sub-unit multiplier lowers loss relative to one, and every super-unit multiplier raises it, at all 19 onsets. The universal monotone range is .25–2; it does not include zero.

K.4 Four-arm 774M comparison: global scaling and selective control

This comparison tests whether two forms of global attenuation reproduce Dual's full-training outcomes. Standard AdamW, half learning rate, global norm matching, and Dual each have three complete 16K runs with seeds 1234, 2027, and 3407. All start from initialization and use the 774M training contract in Appendix C: FP16, sequence length 1024, global batch 480, $(\beta_1, \beta_2) = (0.9, 0.999)$, weight decay 0.1, and input-gradient clipping at 1. Half rate scales both the peak and floor by $1/2$, to 5×10^{-4} and 5×10^{-5}; the other arms retain 10^{-3} and 10^{-4}. Warmup remains 2K steps and the cosine schedule remains 50K steps. The new controls reuse the existing Standard and Dual comparisons, rather than adding independent replications of those arms. Configuration audits verify the shared contract, and recorded per-rank data-probe hashes agree within each seed. Standard seed 3407 uses the original completed scale run; its telemetry and checkpoint settings differ, while its core training configuration and data-probe hashes match.

Online norm matching. At each accepted optimizer update, let Δ_t^{Adam} be the native adaptive displacement from the global control's own current moments and clipped gradient. Compute a virtual

Δ_t^{Dual} at that same state using coordinate cap $c = 0.3$ and the tensor EMA rule of Equations (13)–(14), including its 128-update initialization and frozen reference on trigger. Apply

$$\Delta_t^{\text{global}} = \gamma_t \Delta_t^{\text{Adam}}, \qquad \gamma_t = \frac{\|\Delta_t^{\text{Dual}}\|_2}{\|\Delta_t^{\text{Adam}}\|_2}, \tag{52}$$

with $\gamma_t = 0$ if the native displacement is zero. Decoupled weight decay is excluded from both norms and applied with the native setting. Adam's moment recurrences remain unchanged; the virtual tensor references evolve along this control's own trajectory. The target therefore matches a same-state virtual Dual step, not a step taken by the separately trained Dual run. Logged squared norms satisfy the matching identity to maximum relative error below 4×10^{-16} across all accepted updates; this checks the scalar norm calculation, not bitwise equality of floating-point parameter displacements. Native FP16 skips remain part of the trajectories, and no loss-triggered intervention is added.

Table 15: **Paired four-arm 774M outcomes.** Every row completes 16K steps. Counts use the strict recovery episodes defined in Appendix C; post-warmup cumulative area sums their excess loss. Each seed is one independent training replicate. Zero episodes imply zero episode area under this definition.

Configuration	Seed	All spikes	Post-warmup spikes	Post-warmup cumulative area	16K final validation
Standard AdamW	1234	9	7	1812.60	2.9510
	2027	8	6	1775.05	2.8956
	3407	6	4	1493.20	2.9013
Half learning rate	1234	8	6	1487.50	2.8579
	2027	3	2	738.59	2.8410
	3407	9	7	2094.14	2.8702
Global norm match	1234	11	9	106.66	2.8676
	2027	10	8	126.80	2.8531
	3407	12	10	129.15	2.8713
Dual (Clip-0.3+EMA-2)	1234	0	0	0.00	2.8595
	2027	1	0	0.00	2.8511
	3407	1	0	0.00	2.8568

What the comparison establishes. Each seed retains post-warmup episodes under both global controls, while Dual retains none. Global norm matching reduces mean post-warmup cumulative area from 1693.61 to 120.87 (about 93%), although its episode count is higher than Standard's. Half rate gives 1440.08 and final validation nearly equal to Dual's, with lower endpoints on seeds 1234 and 2027. Dual therefore has no established validation advantage over half rate. The observed distinction is the simultaneous absence of strict post-warmup episodes and competitive validation. These full runs support the practical value of selective attenuation beyond the two tested global controls. Because their trajectories diverge, they do not isolate a unique cause of each spike or prove that allocation alone explains the outcome. The comparison covers three seeds, one halved schedule, and one same-state norm target on the repository's OpenWebText fallback data; it is not an exhaustive tuning study.

K.5 Paired 774M validation endpoints

Table 16: Final 16K validation loss at 774M, paired by training seed.

Seed	Standard	Matched $(.9, .9)$	Dual, $c = 0.3$	Dual, $c = 0.45$
1234	2.9510	2.8646	2.8595	2.8673
2027	2.8956	2.8577	2.8511	2.8476
3407	2.9013	2.8698	2.8568	2.8560

The matched-moment endpoints come from the three completed 774M v25 runs with no coordinate or tensor guard. Their original configuration audits verify 774,090,240 parameters, FP16, global

batch 480, weight decay 0.1, and the common 16K budget; each endpoint is present in its original training log. The strict event analysis reports one warmup episode and zero post-warmup episodes across these runs. The two guarded settings improve on Standard in every paired seed. The small difference between their means is not a broad threshold-optimality result.

K.6 Sensitivity to a broader detector in an earlier analysis

Reapplying the detector to the same 774M trajectories with a 3% relative-rise threshold, instead of the main mitigation comparison's 10%, recovers the earlier broader-burden counts. All other thresholds and recovery grouping are unchanged; these are not additional training runs. Under that protocol, total/post-warmup counts are 23/17 for Standard, 14/8 for Clip-0.3, 24/17 for Clip-0.45, 27/19 for EMA-2, 5/2 for the dual rule with cap 0.3, and 7/3 for the dual rule with cap 0.45.

The stricter detector's zero post-warmup count is therefore not an absence of every smaller excursion. The broader analysis still reports fewer episodes for the combination, but two post-warmup episodes remain. Thus the zero-spike result refers specifically to the strict 10% protocol.

L Published Beta Settings

Table 17 lists explicit settings, not beta ablations. Several LM recipes use $(0.9, 0.95)$ rather than the original Adam default $(0.9, 0.999)$ (Kingma and Ba, 2015). This does not establish a universal transition beginning with GPT-2 or a causal explanation of model quality.

Table 17: **Explicit pretraining beta settings in primary reports.** All listed values concern training, not sampling parameters or third-party fine-tuning defaults. The report location establishes adoption of the pair; it does not isolate a stability or validation gain over $(0.9, 0.999)$.

Model / report	(β_1, β_2)	Location
GPT-3 (Brown et al., 2020)	$(0.9, 0.95)$	Appendix B
LLaMA (Touvron et al., 2023a)	$(0.9, 0.95)$	Section 2.3
Llama 2 (Touvron et al., 2023b)	$(0.9, 0.95)$	Section 2.2
Qwen (Bai et al., 2023)	$(0.9, 0.95)$	Section 2.4
DeepSeek LLM (Bi et al., 2024)	$(0.9, 0.95)$	Section 2.3
DeepSeek-V2 (Liu et al., 2024a)	$(0.9, 0.95)$	Section 3.1.2
DeepSeek-V3 (Liu et al., 2024b)	$(0.9, 0.95)$	Section 4.2

The checked training descriptions of GPT-2, Llama 3, and later Qwen reports did not establish an explicit pretraining beta pair (Radford et al., 2019; Grattafiori et al., 2024; Yang et al., 2024a;b; 2025). We do not infer it from family names, third-party reproductions, or fine-tuning configurations. This is a verification limit, not evidence that those models use different betas.

RoBERTa gives more direct practitioner evidence: Liu et al. (2019, Section 3.1) report improved large-batch stability at $\beta_2 = 0.98$, not 0.95. This does not attribute all benchmark gains to beta. Our own 0.95 comparison separately measures damage and validation (Table 2).