REPAIRING ATTENTION COLLAPSE THROUGH THE LOSS WITH A SPECTRALLY GATED PENALTY

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention collapse can stall transformer training without a numerical failure: attention concentrates on a few tokens and the loss stops improving. We ask whether a term in the loss can bring a run back from collapse. This is not obvious, because a penalty on attention entropy acts through the softmax gradient, and that gradient becomes small as attention saturates. We propose SPECGATE, an attention-entropy penalty whose weight is set for each head by a gate that follows the spectral sharpness of the head's attention scores and passes no gradient of its own. In a 124M-parameter language model, SPECGATE brings runs from an established collapse to 0.07 nats above the healthy run of the same seed, against 1.05 nats without treatment, and raises the loss of healthy runs by 0.004 nats. The same penalty at constant weight, including a separately tuned one, ends 0.12 to 0.14 nats above SPECGATE. Starting SPECGATE at the first spectral alarm rather than after collapse has set in lowers the final loss by 0.22 nats, and stopping it 250 updates after the start of repair raises the final loss by 0.10 nats relative to continuous treatment. At equal numbers of updates, a gate on the entropy deficit and a fixed schedule are equivalent to the spectral gates within 0.02 nats; across the tested interventions, when and how strongly the penalty acts moved the final loss more than the signal that sets its weight. After collapse, QK-Clip, which rescales the weights, ends 0.03 nats above SPECGATE at its published threshold; with a threshold tuned on this task for the final loss, it ends 0.02 nats below SPECGATE and runs faster.

1 INTRODUCTION

A transformer run can fail without producing a NaN. Attention logits grow, attention distributions approach one-hot vectors, and the training loss settles on a plateau well above that of a healthy run (Dehghani et al., 2023; Zhai et al., 2023). The same failure appears in small models trained at high learning rates (Wortsman et al., 2023). Once it has happened, the practical question is whether training can be brought back from it or whether the run has to be rolled back or restarted.

Remedies act at different places. Some change the architecture, for example by normalizing queries and keys or by reparametrizing linear layers; some rescale weights after the optimizer step; some add a term to the loss (Table 1). They also differ in when they act: several act at every step, others only while a statistic of attention exceeds a threshold. Interventions against attention collapse thus differ in where they act and in when.

Table 1: **Interventions against attention collapse by where and when they act.**

Acts on	At every step	Only on demand
Architecture	QK-norm (Dehghani et al., 2023); σReparam (Zhai et al., 2023)	QK-Guard (Xie et al., 2026)
Weights or updates	logit-change control (Anson & Aitchison, 2025)	QK-Clip (Kimi Team, 2025)
Loss	EAR (Attanasio et al., 2022)	thresholded entropy (Jha & Reagen, 2025); SPECGATE (ours)

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

We study remedies in the loss, which change neither the architecture nor the optimizer step. Attention-entropy penalties have been used to reduce reliance on identity terms in fine-tuned classifiers (Attanasio et al., 2022) and to mitigate entropic overload in language models with fewer nonlinearities, where collapse in deeper layers is prevented with weight and spectral normalization instead (Jha & Reagen, 2025). A penalty acts on the model only through gradients, and gradients through a softmax become small near the vertices of the simplex (Mei et al., 2020), so a saturated head leaves an entropy term little to act on. Whether a penalty can bring a run back once collapse is established is open.

We propose SPECGATE, an attention-entropy penalty whose weight is set by a gate. The gate follows the spectral sharpness of the attention scores, which also serves as an alarm, and passes no gradient of its own. To separate what a penalty computes from when and how strongly it acts, we hold the failed model fixed: from states saved once collapse was established, and from states of healthy runs, we continue training with treatments that differ in one respect, namely the kind of penalty, its weight, whether the gate controls it, or when it starts and stops. We also ask where such repair ends: at higher learning rates, against a recipe tuned in advance, and within a fixed remaining time.

We contribute (i) SPECGATE, a detached spectral gate on attention penalties; (ii) repair from checkpoints of an established collapse, in the main recipe and, with settings unchanged, at another context length, corpus and model size, compared with the same penalties at constant weight, including a separately selected weight, with a reduced learning rate and with query/key clipping at its published and at a tuned threshold; (iii) the gate on two penalties whose gradients behave differently at saturation, in collapsed and in healthy regimes, and started when the spectral alarm fires; (iv) all contrasts placed on one axis by whether they change what a penalty computes or when and how strongly it acts.

2 RELATED WORK

Controlling attention logits. σReparam reparametrizes linear layers with spectral normalization and a learned scale and relates low attention entropy to training instability (Zhai et al., 2023). Query/key normalization prevents uncontrolled logit growth in large vision transformers (Dehghani et al., 2023) and in small language models at high learning rates (Wortsman et al., 2023). Hong & Lee (2025) trace entropy collapse to the variance sensitivity of the softmax and keep entropy stable with query/key normalization or ReLU attention. QK-Clip rescales the query and key weights of a head after the optimizer step whenever its largest logit exceeds a threshold (Kimi Team, 2025). QK-Guard leaves training unchanged until attention logits begin to saturate and then switches on per-head query/key normalization (Xie et al., 2026).

Attention penalties. EAR penalizes tokens with low self-attention entropy (Attanasio et al., 2022). Jha & Reagen (2025) penalize the entropy of a head only when it deviates from a learnable threshold by more than a tolerance. These penalties serve other aims than the repair of a failed run. The gate studied here reads a spectral quantity separate from the penalized entropy (Appendix Q).

Detection and repair. Monitors derived from internal signals report training instability thousands of steps before the loss diverges (Huang et al., 2026), and spectral profiles at initialization predict which configurations will diverge (Kim et al., 2026); these works detect or prevent. Collapsed heads in pretrained BLOOM models can be restored by reinitializing their query, key and value weights (Schallon, 2026). Rank collapse, where token representations approach rank one as attention layers are composed (Dong et al., 2021), is a different phenomenon from the saturation of individual heads studied here.

3 COLLAPSE AND ITS SPECTRAL SIGNAL

Four kinds of quantity play different roles here. Attention statistics, the normalized entropy and the spectral sharpness, describe a head; a collapse label, combining attention with task loss, decides when a run counts as failed; an alarm decides when a monitor starts treatment; a gate sets how strongly an active penalty acts on each head. Success is measured by validation loss alone.

Attention depends on differences between logits, not on their shared offset, so we measure scores after centering each row. For one head, $S = QK^\top/\sqrt{d_k} \in \mathbb{R}^{T_q \times T_k}$ holds the scores before the softmax, and row i attends to m_i valid keys with probabilities $A_{i,:} = \mathrm{softmax}(S_{i,:})$. We write $\bar{H}_{\ell,h}$ for the Shannon entropy of a row divided by $\log m_i$, averaged over valid rows and sampled sequences of head h in layer ℓ. The spectral sharpness of a head is

$$s = \frac{\|L(S)\|_2}{\sqrt{T_q} + \sqrt{T_k}}, \qquad L(S) = \big(S - \mathrm{rowmean}(S)\big) \odot M, \tag{1}$$

where M masks invalid keys, the row mean uses valid keys only and $\|\cdot\|_2$ is the largest singular value. It measures the strongest amplification by the centered score matrix. The denominator removes the scale of a random matrix with independent entries; whether attention scores follow that scale is an empirical question (Appendix I). For each head, $s_{\ell,h}$ averages this quantity over the sequences sampled for that head, and $s_{\max}$ is the maximum of $s_{\ell,h}$ over layers and heads. We estimate s with five power iterations on the centered score matrix (Appendix D).

Proposition 1 (Entropy floor from spectral sharpness)**.** *For the score matrix of one sequence with sharpness s, every valid row satisfies $\bar{H}(A_{i,:}) \geq 1 - 2(\sqrt{T_q} + \sqrt{T_k})\, s/\log m_i$.*

Low entropy therefore requires large s. The bound links the two statistics but sets none of the empirical thresholds below and does not identify a failed run. The converse does not hold: a sharp head can still attend usefully, and the bound is vacuous once $s > \log m_i/\big(2(\sqrt{T_q} + \sqrt{T_k})\big)$. The proof is in Appendix C; Zhai et al. (2023) give a related lower bound on attention entropy that decreases exponentially in the spectral norm of the attention logits.

Low entropy alone does not identify a failed run. Our corpus holds 227 runs across model sizes, learning rates and schedules; 48 of them end with a head below normalized entropy 0.05 and still reach near-healthy loss (Appendix E). We therefore call a run collapsed only when attention and task performance fail together: its lowest head entropy is below 0.05 and its validation loss is at least 0.25 nats above that of healthy training. In the corpus both are read at the end of training, against the median of entropy-healthy runs with the same dataset and width. For the repair experiments they are read during training, against a healthy reference curve of the same recipe at the same update, and must hold at two consecutive validations; we call the state saved at the second of these validations an *established* collapse.

The label uses no spectral quantity, so it can test one. A monitor reads $s_{\max}$ every 50 updates on a small batch, at below 1% of training time, and raises an alarm when two consecutive readings exceed 300. This threshold decides when treatment starts; how strongly the treatment then acts is set separately. In this corpus, 108 runs are collapsed under the label and 119 are not; the alarm fires in 104 collapsed runs and in 25 non-collapsed ones. An entropy criterion with a similar detection rate (106 of 108) fires in 69 non-collapsed runs (Appendix E).

4 A SPECTRALLY GATED ATTENTION PENALTY

An entropy penalty pushes a head toward broader attention. Sharp attention can also be useful, so the question is when and how strongly to apply the penalty. A penalty in the loss therefore has two parts: a term $P_{\ell,h}$ that sets what is optimized, and a weight $\alpha_{\ell,h}(t)$ that sets when and how strongly it acts on head h of layer ℓ at update t, so that $R = \mathrm{mean}_{\ell,h}[\alpha_{\ell,h}(t)\, P_{\ell,h}]$. SPECGATE takes the entropy deficit as term and a weight that rises with the spectral sharpness of the head:

$$R_{\mathrm{Gate}} = \lambda \operatorname*{mean}_{\ell,h}\big[g(s_{\ell,h})\,(1 - \bar{H}_{\ell,h})\big], \qquad g(s) = \mathrm{sg}\Big[\sigma\Big(\frac{s - s_0}{w}\Big)\Big], \tag{2}$$

where σ is the logistic function, sg stops differentiation, and, for the numerical entropy calculation, each probability is replaced by its maximum with 10^{-8} before evaluating both factors in $p \log p$, without renormalizing (Appendix U). The gate is detached: it is recomputed from the current model at every update, and the gradient of R_{Gate} is $\lambda\, \mathrm{mean}_{\ell,h}[g(s_{\ell,h})\, \nabla(1 - \bar{H}_{\ell,h})]$. The gate thus rescales the entropy gradient and adds none of its own.

Its center $s_0 = 50$ and width $w = 10$ were set once from healthy runs by a calibration rule and left unchanged (Appendix K); at $s = 30$ the gate keeps about 0.12 of the weight. We propose $\lambda = 3$ as the default of the method. It was selected on two development seeds of the repair task, where the

grid is flat between 3 and 10, and is kept unchanged in all 124M repair and panel-set comparisons; other methods run at their published defaults and, where stated, also at settings selected on the same development seeds. A separate 30M study of protection from initialization, run later with its own seeds, selected 3 at the two lower of its four learning rates and 10 at the two higher ones (Appendix V). It supports a default in this range, not the value 3 at every learning rate.

A saturated head leaves an entropy term little to act on.

Proposition 2 (Vanishing entropy gradient)**.** *For $p = \mathrm{softmax}(z)$, $\partial H/\partial z_j = -p_j(\log p_j + H)$. As p approaches a one-hot vector with tail mass ε at fixed dimension, $\|\nabla_z H\|_2 = O(\varepsilon \log(1/\varepsilon))$.*

The gradient shrinks with saturation but remains nonzero at any finite saturation. Softmax gradients are known to become small near the vertices of the simplex, which slows gradient methods there (Mei et al., 2020); Proposition 2 states this for the entropy of one attention row. The statement concerns logit gradients; parameter updates also pass through the logit Jacobian, gradient clipping and the optimizer (Appendix C), so equal weights need not give equal updates. The gate does not remove this attenuation; it sets the weight by which the small gradient is multiplied.

We compare alternative terms and weighting rules. As a second term we use uniform cross-entropy, $\mathrm{KL}(U_m \,\|\, p)/\log m$ with U_m uniform over the m valid keys, in place of $1 - \bar{H}$ and with the same gate, at $\lambda = 1$ chosen on the same development seeds. Its logit gradient $(p_j - 1/m)/\log m$ does not vanish at saturation. A penalty with hinge adds to SPECGATE a term on the spectral sharpness above s_0, which reaches a saturated head without passing through its entropy (Appendix B). Both penalties are also run with the gate replaced by one, keeping the probability floor, per-head averaging, spectral computation and random-number stream unchanged. The constant-weight penalties run at the coefficient of their gated counterpart and, for the entropy penalty, also at a coefficient selected separately on the same development seeds; query/key clipping runs at its published threshold and at one tuned on the same development seeds.

5 RESULTS

We train a 124M-parameter decoder on WikiText-103 for 5,000 updates with a cosine schedule (Appendix F), at learning rate 3×10^{-4} or 1.2×10^{-3}, and report unregularized validation loss at update 5,000 in nats. Every comparison is paired: all arms of a seed start from the same saved state, with identical weights, optimizer moments, schedule position and data stream. We use two sets of saved states. The *confirmation set* holds eight seeds trained at 1.2×10^{-3} until two consecutive validations met the collapse label of Section 3. The *panel set* holds sixteen new seeds per regime, assigned together with the protocol before any run: collapsed states at 1.2×10^{-3}, saved at the first checkpoint that meets the label, and healthy states at 3×10^{-4}, saved at update 1,750 without any filter. Except for the three further configurations at the end of Section 5.1, all results use this recipe.

Each repair is read against two references. The untreated continuation of the same state shows whether a treatment improved on continuing the failed run; the healthy reference, trained from initialization at the lower learning rate, shows how close the run came to healthy training. The task target is the final loss of healthy training at 3×10^{-4} plus 0.10 nats (4.741), reached when two consecutive validations meet it; target attainment is assessed per run, so means and target counts can rank arms differently. Except in one comparison, arms train for the same number of updates; that comparison asks which method gives the lowest loss within a fixed remaining time.

We report paired mean differences with intervals and call a difference practically small only when an equivalence test against a ± 0.02-nat margin passes. Primary contrasts of the gate ablation, the mean gate, uniform cross-entropy and query/key clipping on the panel set use 97.5% paired intervals, all others 95%. Arms marked † were added after the first results; their protocols were fixed before they ran (Appendix A lists all studies).

5.1 REPAIR FROM AN ESTABLISHED COLLAPSE

We first ask whether a penalty can bring a run back from an established collapse. On the confirmation set, collapse was established after 1,500 or 1,750 updates. Each treatment branches from that state, and the untreated arm continues the same run.

SPECGATE ends at 4.716 against 5.695 for the untreated continuation, a paired difference of -0.978 $[-1.072, -0.884]$ that is negative in all eight seeds; six of eight runs reach the target (Table 2). The gap to healthy runs of the same seed is 0.072. Halving the learning rate at the same state ends at 5.332, and no run reaches the target. Gated uniform cross-entropy ends at 4.718.

The penalty needs weight to do so. On two separate development seeds, SPECGATE ends at 5.39 with coefficient 0.3, at 4.83 with 1, at 4.72 with 3 and at 4.73 with 10. A logit gradient small but nonzero at saturation (Proposition 2) thus reaches collapsed heads once the weight is large enough.

Other configurations. With $\lambda = 3$ and the gate unchanged, we branched established collapses of three further configurations†, four prefixes each: context length 1024, OpenWebText, and a 350M-parameter model, the last labeled with a healthy rule revised after the original rule failed (Appendix H.4). SPECGATE ends 0.18, 0.35 and 1.80 nats below the untreated continuation and 0.12, 0.10 and 0.12 nats below the entropy penalty at constant weight, each 95% paired interval, taken separately, beyond a ± 0.02-nat margin; its distance to the healthy reference is 0.29, 0.11 and 0.11 nats, and 2 of 12 runs reach the target of their configuration.

Time course and stopping. On six new collapsed states†, the spectral sharpness under SPECGATE fell below the gate center within 300 updates of the branch in every run, while the loss advantage over the untreated continuation built up until update 5,000 (Figure 1). Switching SPECGATE off 250 updates after the branch ended 0.102 nats above continuous treatment $[0.061, 0.143]$ and 0.826 below the untreated continuation; in all six runs $s_{\max}$ was above the gate center again 50 updates after the switch and stayed there.

Table 2: **Repair on the confirmation set.** Eight seeds; every arm branches from the same established collapse per seed and trains to update 5,000. Δ: arm minus untreated continuation, 95% paired interval. Target: healthy final loss plus 0.10 nats at two consecutive validations. Gap: arm minus healthy run of the same seed. The frozen protocol named the penalty with hinge (Appendix B) as primary arm; all arms are in Appendix F.

Arm	Final loss	Δ vs. untreated	Target	Gap
SPECGATE ($\lambda = 3$)	4.716	-0.978 $[-1.072, -0.884]$	6/8	0.072
Penalty with hinge ($\lambda = 3$)	4.712	-0.983 $[-1.078, -0.888]$	6/8	0.067
Gated uniform-CE ($\lambda = 1$)	4.718	-0.976 $[-1.068, -0.884]$	6/8	0.074
Learning rate $\times 0.5$	5.332	-0.363 $[-0.442, -0.284]$	0/8	0.687
Untreated	5.695		0/8	1.050

5.2 GATED AND UNGATED PENALTIES

The gate could contribute to this repair, or the penalty alone could carry it. We compare each penalty with the same penalty whose gate is replaced by one, at equal coefficient, in both regimes of the panel set. Both versions keep the probability floor, per-head averaging, spectral computation and random-number stream; only the applied weight differs.

For the entropy penalty, gating lowers the final loss by 0.144 nats after collapse $[-0.149, -0.138]$ and by 0.119 in healthy continuations $[-0.121, -0.116]$, in all 32 pairs (Figure 8 in Appendix L). Thirteen of sixteen gated runs reach the target after collapse, and no ungated run does. In healthy continuations the ungated penalty ends 0.122 nats above untreated training and the gated penalty 0.004 $[0.003, 0.005]$. Selecting the ungated coefficient separately on the development seeds does not close the gap: on the confirmation set, where it is chosen at 1, it still ends 0.124 nats above the gated penalty $[0.095, 0.153]$.

Uniform cross-entropy† gives the same gate a penalty whose logit gradient does not vanish at saturation. At coefficient 1, gating lowers its final loss by 0.096 nats after collapse $[-0.100, -0.091]$ and by 0.059 in healthy continuations $[-0.061, -0.057]$; 14 of 16 gated runs reach the target after collapse, and no ungated run does. With the gate, the two penalties end at 4.707 and 4.708 after collapse and at 4.640 and 4.651 in healthy continuations. In a secondary comparison fixed in the protocol, at their separately selected coefficients, they are equivalent within the ± 0.02-nat margin in both regimes: uniform cross-entropy minus entropy is 0.0005 nats after collapse $[-0.0024, 0.0034]$

and 0.011 in healthy continuations $[0.010, 0.013]$ (95% intervals). For both penalties and in both regimes, the gate lowered the final loss at equal coefficient.

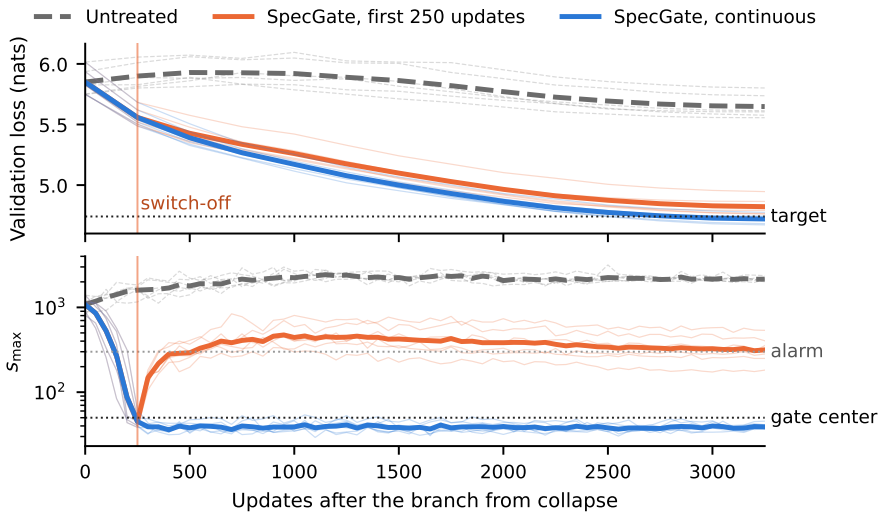

Figure 1: **Time course of repair and of the spectral sharpness.** Six new collapsed states† (Appendix T), branched at update 1,500 or 1,750 and trained to update 5,000; the axis ends at the shorter horizon. Top: validation loss. Bottom: largest spectral sharpness over layers and heads, log scale. Thin lines: seeds; thick lines: mean (geometric for $s_{\max}$). The vertical line marks where the short arm sets its weight to zero; before it, the two treated arms are identical. The figure does not show whether a repair completed before the switch would need a continued weight.

Table 3: **All arms on the panel set.** Final loss at update 5,000 and target attainment on sixteen states per regime; within a regime all arms share the state. †: added after the first results, protocol fixed before the runs. ‡: not the published default; tuned on development seeds of this task for the final loss. Paired intervals are in the appendices listed in Table 4.

	After collapse		Healthy	
Arm	Loss	Target	Loss	Target
Untreated	5.614	0/16	4.636	16/16
SpecGate (entropy, gated, 3)	4.707	13/16	4.640	16/16
Entropy, ungated, 3	4.851	0/16	4.758	0/16
Uniform-CE, gated, 1†	4.708	14/16	4.651	16/16
Uniform-CE, ungated, 1†	4.803	0/16	4.710	16/16
Uniform-CE, ungated, 0.3†	4.739	9/16	4.668	16/16
Mean gate (entropy, 3)†	4.714	12/16	4.643	16/16
Entropy-deficit gate (entropy, 3)†	4.711	12/16	4.642	16/16
Replayed schedule (entropy, 3)†	4.712	12/16	4.653	16/16
QK-Clip, $\tau = 100$†	4.737	10/16	4.636	16/16
QK-Clip, $\tau = 30$†‡	4.690	16/16	4.638	16/16
Entropy-gated regularizer, 10^{-5}†	5.517	0/16	4.635	16/16
Entropy-gated regularizer, 10^{-4}†‡	4.769	1/16	4.644	16/16

5.3 Weight, head allocation and signal of the gate

The gate could act by lowering the weight overall, by sharing it among heads, or through the signal it reads. On fixed validation probes, the mean gate is 0.933 at the collapsed states and 0.105 at the healthy ones. On the final training batches it is near 0.03 in both regimes (Appendix L).

A smaller constant weight does not reproduce this. Ungated uniform cross-entropy at coefficient 0.3† ends 0.031 nats above the gated penalty after collapse $[0.028, 0.034]$ and 0.017 above it in healthy continuations $[0.016, 0.019]$; 9 of 16 runs reach the target after collapse against 14.

A mean gate† gives every head the average of the per-head gates, so on the same state and batch it spends the same total gate weight; gradients and later gate values can still differ (Appendix N). It ends 0.007 nats above the per-head gate after collapse $[0.002, 0.011]$ and 0.003 above it in healthy continuations $[0.002, 0.004]$; 12 of 16 runs reach the target after collapse.

A penalty gated by the entropy it penalizes†, the regularizer of Jha & Reagen (2025) at a coefficient selected on the same development seeds at the upper edge of its grid, ends 0.061 nats above SPECGATE after collapse $[0.055, 0.067]$ (95% paired interval) and is equivalent to it in healthy continuations; 1 of 16 runs reaches the target after collapse (Appendix Q).

Two further controls† keep the construction of SPECGATE and change only what sets the gate (Appendix R). A gate on the entropy deficit of each head ends 0.004 nats above SPECGATE after collapse $[0.002, 0.006]$ and 0.002 above it in healthy continuations $[0.002, 0.003]$. A fixed schedule that replays the mean gate of two development runs over updates since the fork, the same for every head and run, ends 0.002 below the mean gate after collapse $[-0.006, 0.002]$ and 0.011 above it in healthy continuations $[0.010, 0.012]$. All four differences are equivalent within ± 0.02 nats after Holm correction (95% paired intervals). In healthy continuations the replayed schedule ends 0.018 nats above the untreated run, against 0.004 for SPECGATE. Within the construction of SPECGATE, replacing the gate by a constant weight changed the final loss more than changing how the weight is shared across heads or which signal sets it.

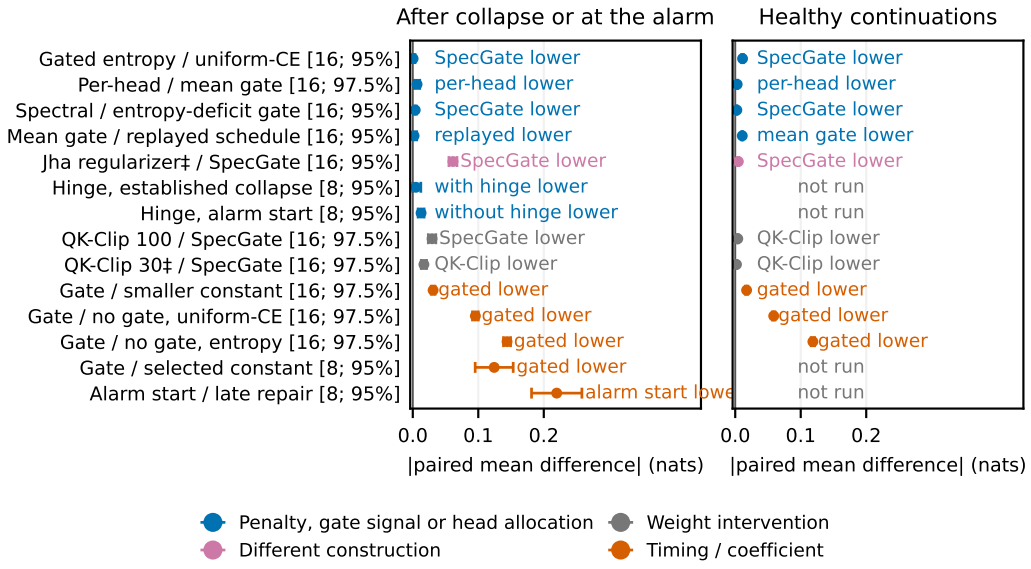

Figure 2: **Contrasts grouped by what they change.** Points show the magnitude of the paired mean difference in final loss; bars map each signed interval onto its magnitude, and an interval that contains zero starts at zero. Labels give the number of paired states and the interval level. Each row names the arm with the lower mean. The comparisons come from separately designed protocols and were grouped after their results were known; the figure is not a pooled test. ‡: setting tuned on development seeds of this task for the final loss, not the published default (Appendices P and Q).

5.4 TREATMENT STARTED AT THE ALARM

The alarm of Section 3 can also start the treatment. We reran the confirmation seeds from initialization at 1.2×10^{-3} with the alarm armed, so that SPECGATE switches on at the update after the alarm. The alarm fired after 800 to 900 updates in every seed, and until then each run is identical to its untreated one.

Started at the alarm†, SPECGATE is active for about 4,150 updates and ends at 4.497. Started from the established collapse of the same seed, as in Table 2, it is active for 3,250 or 3,500 updates and ends at 4.716. The paired difference is -0.220 $[-0.258, -0.181]$; all eight alarm-started runs reach the target, against six of eight.

From an established collapse on the confirmation set, the penalty with hinge ends 0.005 nats below SPECGATE (descriptive 95% interval $[-0.012, 0.003]$). Started at the same alarm on the same seeds†, it ends at 4.509, and SPECGATE without it ends 0.013 nats lower $[-0.017, -0.008]$.

5.5 QUERY/KEY CLIPPING

Query/key clipping acts on the weights instead of the loss: after the optimizer step it rescales the query and key projections of a head whenever its largest logit exceeds a threshold τ (Kimi Team, 2025). We run it at its published threshold $\tau = 100$ from the same states as all other arms. After collapse it ends 0.029 nats above SPECGATE $[0.024, 0.035]$ and reaches the target in 10 of 16 runs against 13. In healthy continuations the two are equivalent within the prespecified ± 0.02-nat margin (difference -0.004 $[-0.005, -0.003]$; Table 3). We additionally compare against a threshold tuned for this task‡, selected on the development seeds by the final loss compared here. At the selected $\tau = 30$, QK-Clip ends 0.017 nats below SPECGATE after collapse $[-0.021, -0.013]$, reaches the target in 16 of 16 runs and is equivalent to it in healthy continuations (Appendix P). In the comparison at a fixed remaining time, on six new collapsed states†, each arm received the same wall-clock budget (Appendix T). In this time QK-Clip at the tuned threshold‡ completes 2,925 updates on average and SPECGATE 2,472; QK-Clip ends 0.068 nats below SPECGATE $[-0.084, -0.052]$, beyond the margin, and reaches the target in 5 of 6 runs against none.

6 DISCUSSION

Timing and form. Figure 2 places the contrasts of the confirmation and panel sets on one axis, sorted by which part of the penalty they change. After an established collapse, changing within the construction of SPECGATE what the term computes, which signal sets the weight or how the weight is shared across heads moved the final loss by 0.0005 to 0.007 nats; started at the alarm, adding the hinge moved it by 0.013. In the repair and alarm-start comparisons, changing when and how strongly the penalty acts moved it by 0.031 to 0.220 nats: a smaller or a separately selected constant weight, no gate, a start at the alarm and, on six new states, a stop after 250 updates (0.102). In healthy continuations the two ranges are 0.002 to 0.011 and 0.017 to 0.119. The complete entropy-gated regularizer of Jha & Reagen (2025), which changes signal, direction, thresholds and functional form together, ended 0.061 nats above SPECGATE; this comparison does not separate these changes. QK-Clip lies at 0.029 at its published threshold and at 0.017 at a threshold tuned on this task‡. These are observed mean differences between variants of one recipe, grouped after their results were known, not a single test. This pattern is consistent with results outside attention. Regularization of image classifiers matters mainly during an early phase and can be removed afterwards at little cost (Golatkar et al., 2019); dropout helps when restricted to early or to late training (Liu et al., 2023); and entropy terms in reinforcement learning are sensitive to a fixed coefficient but work with one that follows the current entropy (Haarnoja et al., 2018; Zhang et al., 2026). Where an intervention acts can also matter: with trigger and firing step fixed, a query/key action rescues a low-precision collapse that a rescaling inside the softmax does not (Xie et al., 2026); all interventions compared here act on the query/key channel. A second boundary concerns the learning rate.

Loss-based repair and its limits. A loss term acts on the model only through gradients. In our runs it repaired an established collapse once its weight was large enough; with the same weight and gate it lowered the final loss after collapse at another context length, corpus and model size, although 10 of 12 of those runs stayed short of the healthy target. Gated, it left healthy training nearly unchanged, and it needed no direct gradient on the spectrum. Its reach ends where updates become large. In a separate 30M protection study with settings selected per learning rate, at rate 0.1 without warmup QK-Clip ended at 5.46, the penalties at 7.33 to 7.49 and untreated training at 7.31; the hinge and gated uniform cross-entropy kept all runs out of the collapse label there, SPECGATE none. At high but not extreme rates the gated penalty with hinge ended below QK-Clip in three of ten conditions, and four remained unresolved (Appendix J). A penalty also does not replace a stable

recipe chosen in advance: when learning rate, warmup and penalty weight were tuned jointly at equal training time, untreated training ended lowest (4.218 against 4.368 with SPECGATE). Under a fixed remaining budget after collapse, a gated penalty applied in short bursts after each alarm ended below rollback and restart (5.013 against 5.115 and 5.207 on four development seeds, none at the target). Both studies rank settings on development seeds (Appendix H). QK-Clip acts on the weights and SPECGATE on the loss, so the two compose; whether their effects add is untested.

Recovery from collapse. From an established collapse, a gated penalty closed the gap to healthy runs of the same seed to 0.072 nats. On the confirmation set the loss advantage built up at nearly the same pace with the penalty with hinge and uniform cross-entropy, whose sharpness fell faster (Appendix S). On six new states, stopping after 250 updates raised the final loss by 0.102 nats relative to continuous treatment and let the sharpness rise again; at the switch it was still at or above the center in two of six runs, so the experiment does not show whether a completed repair needs a continued weight. Started at the alarm, 800 to 900 updates after initialization, the same penalty ended 0.220 nats lower still, and all eight runs reached the target. Deep networks can carry lasting deficits when an early problem is not corrected in time (Achille et al., 2019), and collapsed heads in pretrained models can be restored by reinitialization (Schallon, 2026). Attention collapse in our recipe lies between these cases: most of the loss gap is undone, while attention stays sharper than in healthy training; after repair the lowest head entropy ends at 0.017 to 0.080 with SPECGATE, against 0.195 to 0.258 in the healthy runs (Appendix F). Earlier intervention left a smaller final gap, but treatment also lasted longer and the healthy reference trained at the lower learning rate, so the effect of time spent in collapse is not isolated.

Detection and dosing. The same spectral sharpness bounds the entropy from below, raises the alarm and sets the dose. For the dose after an established collapse, the spectral sharpness was not needed: a gate on the entropy deficit and a replayed schedule without feedback ended within 0.005 nats of the spectral gates, against 0.144 nats between gated and ungated penalties. Feedback showed as a small difference in healthy continuations, 0.018 against 0.004 nats above untreated training. Computing the spectral sharpness costs time: at equal remaining wall-clock time on six new states†, the replayed schedule, which needs no power iterations, completed 2,676 updates against 2,472 and ended 0.022 nats below SPECGATE $[-0.027, -0.017]$; a difference beyond the margin is not established (Appendix T). Where the spectral sharpness differed from entropy in our data was detection: as an alarm it fired in 25 of 119 non-collapsed runs, against 69 for an entropy criterion with a similar detection rate. We did not start the same treatment from an entropy alarm, so the 0.22 nats gained by starting at the spectral alarm do not show that this alarm is the better trigger.

Limitations. The main results come from one 124M recipe on one corpus. Arms marked † were added after the first results, on states that had been inspected, except the transfer repairs and the comparisons on six new states, which were added after all other 124M results were known. We did not compare alarm-started against always-on treatment. In a benchmark, the active penalty added 35.9% time per update (60.9% in the reference implementation) and monitoring every 50 updates 0.34% (Appendix W). Beyond this recipe, the collapse label produced qualifying states at context 1024, on OpenWebText and, under a healthy rule revised after the original rule failed, at 350M parameters (Appendix H.3); repair from these prefixes with unchanged settings rests on four prefixes per configuration (Section 5.1).

Conclusion. In these comparisons of attention penalties, changing when or how strongly the penalty acts moved the final loss more than changing, within the construction of SPECGATE, what it penalizes or which signal sets its weight. SPECGATE supplies that timing from the quantity that also detects collapse: it brought runs back from an established collapse, 0.98 nats below the untreated continuation, left healthy training nearly unchanged, ended below the penalties at constant weight and, after collapse, below QK-Clip at its published threshold, did more when started at the alarm, and lowered the final loss in three further configurations with unchanged settings. QK-Clip at a threshold tuned on this task for the final loss ended 0.017 nats below it at equal updates and 0.068 at equal time. Whether this ordering holds at larger scale and in other architectures is open.

STATEMENT ON LLM USAGE

LLM assistants from Anthropic and OpenAI were used under author direction for editing and revising the text for clarity and consistency, for literature search, including the identification of related work in adjacent fields, and for checking derivations and proofs. They also assisted with hypothesis development, experiment design, implementation, numerical and statistical checks, and consistency checks between text and result files. All reported numbers were produced by the authors' analysis code from versioned result files. The authors reviewed all AI-assisted work and take full responsibility for the scientific claims, the text and the released artifacts.

REPRODUCIBILITY STATEMENT

The recipe, the collapse label, the calibration of the gate and every selection rule are given in Appendices F, E, K and V; Appendix A lists every study with its seeds and status, and Appendix G describes the run records. The protocols of the added studies were fixed before their first run; their commits or freeze records are given in the respective appendices. An anonymized package with the training code, the frozen protocols, the run logs and the scripts that regenerate every reported number, table and figure from these logs will be made available to the reviewers through the OpenReview forum.

REFERENCES

Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations (ICLR)*, 2019.

Ben Anson and Laurence Aitchison. Controlling changes to attention logits. *arXiv preprint arXiv:2511.21377*, 2025. URL `https://arxiv.org/abs/2511.21377`.

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. Entropy-based attention regularization frees unintended bias mitigation from lists. *Findings of the Association for Computational Linguistics (ACL)*, 2022. arXiv:2203.09192.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 2023. z-loss for output-logit stability.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning (ICML)*, 2023. introduces QK-normalization.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *ICML*, 2021.

Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 10678–10688, 2019.

Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, and Sergey Levine. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Jonghyun Hong and Sungyoon Lee. Variance sensitivity induces attention entropy collapse and instability in transformers. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Ruixuan Huang, Hantao Huang, Yifan Huang, Ansheng You, Zhenxing Zhang, and Shuai Wang. Mechanism-driven monitors for preemptive detection of LLM training instability. *arXiv preprint arXiv:2606.28116*, 2026.

Nandan Kumar Jha and Brandon Reagen. Entropy-guided attention for private LLMs. *arXiv preprint arXiv:2501.03489*, 2025. AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI).

Bum Jun Kim, Shohei Taniguchi, Makoto Kawano, Yusuke Iwasawa, and Yutaka Matsuo. Residual Koopman spectral profiling for predicting and preventing transformer training instability. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 3036–3060. PMLR, 2026. URL https://proceedings.mlr.press/v337/kim26d.html.

Kimi Team. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Zhuang Liu, Zhiqiu Xu, Joseph Jin, Zhiqiang Shen, and Trevor Darrell. Dropout reduces underfitting. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 22233–22248, 2023.

Jincheng Mei, Chenjun Xiao, Bo Dai, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Palmer Schallon. Surgical repair of collapsed attention heads in ALiBi transformers. *arXiv preprint arXiv:2603.09616*, 2026. URL https://arxiv.org/abs/2603.09616v1.

Mitchell Wortsman, Peter J Liu, Lechao Xiao, Katie Everett, Alex Alemi, Ben Adlam, John D Co-Reyes, et al. Small-scale proxies for large-scale transformer training instabilities. *arXiv preprint arXiv:2309.14322*, 2023.

Shuxiao Xie, Shuyang Xie, Yuan Cao, Dezhi Ran, Wei Yang, and Tao Xie. One QK channel, many sources: Guarding low-precision attention collapse. *arXiv preprint arXiv:2608.02091*, 2026. URL https://arxiv.org/abs/2608.02091v1.

Nikolay Yudin, Alexander Gaponov, Sergei Kudriashov, and Maxim Rakhuba. Pay attention to attention distribution: A new local lipschitz bound for transformers. *arXiv preprint arXiv:2507.07814*, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.06296.

Xiaoyun Zhang, Xiaojian Yuan, Di Huang, Wang You, Chen Hu, Jingqing Ruan, Kejiang Chen, and Xing Hu. Revisiting entropy regularization: Adaptive coefficient unlocks its potential for LLM reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 18005–18020, 2026. doi: 10.18653/v1/2026.findings-acl.895.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. ST-MoE: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. router z-loss.

A Overview of studies

Table 4 lists every study behind a number in the main text, with its states or seeds, its status and its primary endpoint. *Development*: settings are selected and evaluated on the same seeds. *Frozen*: arms, endpoints and analysis were fixed before the runs; for studies marked † this happened after the first results were known; ‡ marks a setting tuned on development seeds of this task for the final loss; for the panel-set and confirmation-set studies, other arms on the same states had been inspected. All repair studies branch every arm from the same saved state and report unregularized validation loss at update 5,000, except the comparison at equal remaining time, which reports it at the end of the time budget.

Table 4: **Studies, seeds and status.** Section and appendix give where each study is reported.

Study	Where	States or seeds	Status	Primary endpoint
Collapse corpus and alarm	§3, App. E	227 existing runs, 108 collapsed	analysis of existing runs	detections and false alarms against the label
Coefficient grid	§5.1, App. F	development seeds 3, 4	development	final loss
Repair, confirmation set	§5.1, Tab. 2	seeds 10–17	frozen	penalty with hinge minus untreated (prespecified primary); other arms, including SPECGATE, reported alongside
Gate ablation, confirmation seeds	abstract, §5.2, Fig. 2, App. L	seeds 10–17; constant weight also selected on seeds 3, 4	exploratory	gated minus ungated
Gate ablation, panel set	§5.2, App. L	16 collapsed and 16 healthy states, seeds 83300–83315	frozen	gated minus ungated
Uniform cross-entropy (D03)	§5.2, §5.3, App. O	panel set	frozen†	gated minus ungated; dose
Mean gate	§5.3, App. N	panel set	frozen†	per-head minus mean gate
Entropy-gated regularizer (D04)	§5.3, App. Q	panel set	frozen†	regularizer minus SPECGATE
Gate signal and schedule (D06)	§5.3, App. R	panel set	frozen†	entropy gate minus SPECGATE; schedule minus mean gate
Alarm start (D01)	§5.4, App. O	seeds 10–17	frozen†	alarm start minus late start
Query/key clipping (D02)	§5.5, App. P	panel set; threshold selected on seeds 3, 4	frozen†	QK-Clip minus SPECGATE
Equal remaining time (D08)	§5.5, §6, App. T	6 new collapsed states, seeds 20–25	frozen†	SPECGATE minus replayed schedule; SPECGATE minus QK-Clip ($\tau = 30^{\ddagger}$), at 302 s
Short treatment (D09)	§5.1, Fig. 1, App. T	same 6 states	frozen†	250 updates minus continuous treatment
Transfer feasibility	App. H.3	3 configurations, 4 healthy references and 4 prefixes each	admission rules; 350M rule revised	eligibility of prefixes
Transfer repair (D05)	§5.1, App. H.4	3 configurations, 4 prefixes each	frozen†	SPECGATE minus untreated
Protection from initialization, 30M	Discussion, App. J	2 development seeds per rate, 8 evaluation seeds	settings selected per rate; penalty comparisons added later, exploratory	final loss; collapse label
Joint tuning of the recipe	Discussion, App. H.2	development seeds 1004–1006	development	final loss at equal training time
Fixed remaining budget	Discussion, App. H	4 development seeds	development	final loss at budget end
Context-length checks	§3, App. I	2 to 3 seeds per statistic	predictions registered	control of $s_{\max}$
Cost	Discussion, App. W	benchmark, 3 blocks per method	measurement	time per update

B SPECTRAL HINGE

The penalty with hinge, SPECBREAK in the released code, adds to SPECGATE a term that acts directly on the spectral sharpness:

$$R_{\mathrm{Break}} = R_{\mathrm{Gate}} + \lambda_s \operatorname*{mean}_{\ell,h} \big[\mathrm{relu}(s_{\ell,h} - s_0)/s_0\big]. \tag{3}$$

Setting $\lambda_s = 0$ gives SPECGATE at the same entropy coefficient. In the released artifacts both coefficients are written as a multiplier of unit coefficients; the development selection in Appendix F scales the gate coefficient and the hinge coefficient together, and multiplier 3 corresponds to $\lambda = 3$ in Eq. (2).

Proposition 3 (Persistent spectral gradient)**.** *Let $\tilde{S} = L(S)$ have a simple largest singular value with vectors u, v, let $r_S = \mathrm{rank}(\tilde{S})$, and let $D_T = \sqrt{T_q}+\sqrt{T_k}$ be the normalizer of Eq. (1). For the per-sequence hinge $r(S) = \mathrm{relu}(s(S) - s_0)/s_0$ underlying Eq. (3), $\partial r/\partial S = L^*(uv^\top)/(s_0 D_T)$ on the active side, where L^* is the adjoint of masked row-centering, and*

$$\frac{1}{s_0 D_T \sqrt{r_S}} \;\le\; \|\partial r/\partial S\|_F \;\le\; \frac{1}{s_0 D_T}, \tag{4}$$

with the upper bound attained without masking. No factor involves the softmax output p, and at fixed matrix shape the lower bound is a positive constant.

The proof is in Appendix C. The proposition concerns logit gradients; parameter updates also pass through the logit Jacobian, gradient clipping and the optimizer.

C PROOFS

Proposition 4 restates Proposition 1 and adds the scalar bound it rests on.

Proposition 4 (Entropy floor from spectral sharpness, restated)**.** *Let $z \in \mathbb{R}^n$, $p = \mathrm{softmax}(z)$, $\Delta = \max_i z_i - \min_i z_i$. Then $H(p) \geq \log n - \Delta$; and for row i of a logit matrix S with row-centered $\tilde{S}$ and sharpness s as in Eq. (1), $H(p_i)/\log m_i \geq 1 - 2(\sqrt{T_q} + \sqrt{T_k})\, s/\log m_i$ for rows with $m_i \geq 2$ valid keys.*

Proof. $p_i = e^{z_i}/\sum_j e^{z_j} \leq e^{\Delta}/n$, so $-\log p_i \geq \log n - \Delta$ for all i, and $H = \sum_i p_i(-\log p_i) \geq \log n - \Delta$. For the second claim, $\|\tilde{S}_{i,:}\|_2 \leq \|\tilde{S}\|_2 = (\sqrt{T_q}+\sqrt{T_k})\, s$ and each entry obeys $|\tilde{S}_{ij}| \leq \|\tilde{S}_{i,:}\|_2$, so the row's range over valid keys is at most $2(\sqrt{T_q}+\sqrt{T_k})\, s$; the softmax ignores the row mean, so Δ_i on $\tilde{S}_{i,:}$ equals that on $S_{i,:}$. □

The constants are loose at operating sizes (implied trigger $s \approx 0.08$ at $T{=}256$ versus the working $s_0{=}50$); we use only the necessary-condition direction.

Proof of Proposition 2. Using $\partial p_i/\partial z_j = p_i(\delta_{ij} - p_j)$,

$$\frac{\partial H}{\partial z_j} = -\sum_i p_i(\delta_{ij} - p_j)(\log p_i + 1) = -p_j(\log p_j + 1) + p_j(1 - H) = -p_j(\log p_j + H).$$

For fixed row dimension, with $p_k = 1 - \varepsilon$ and $p_j = O(\varepsilon)$ for $j \neq k$: $H = O(\varepsilon \log \frac{1}{\varepsilon})$, the k-th component is $O(\varepsilon \log \frac{1}{\varepsilon})$, each other component is $O(\varepsilon \log \frac{1}{\varepsilon})$, and summing squares gives $\|\nabla_z H\|_2 = O(\varepsilon \log \frac{1}{\varepsilon})$. □

Proof of Proposition 3. The statement assumes a simple top singular value; the degenerate set has measure zero, and on it σ_1 admits a subgradient of the same form, so the hinge remains well-defined. For a simple largest singular value, $\nabla_{\tilde{S}}\sigma(\tilde{S}) = uv^\top$ with $\|uv^\top\|_F = 1$; the chain rule through the linear centering map L gives $\partial\sigma/\partial S = L^*(uv^\top)$; the statistic divides by $D_T = \sqrt{T_q}+\sqrt{T_k}$ and the $\mathrm{relu}\,/s_0$ scaling contributes $1/s_0$, so the active-side factor is $1/(s_0 D_T)$. *Unmasked:* $\tilde{S} = SC$ with $C = I - \frac{1}{T_k}\mathbf{1}\mathbf{1}^\top$ symmetric idempotent; $L^*(uv^\top) = u(Cv)^\top$, and the top right singular vector v of SC lies in $\mathrm{range}(C)$, so $Cv = v$ and the norm is exactly 1. *Causal:* $L^*(uv^\top)$ has rows $u_i(C_i M_i v)^\top$ with C_i centering over row i's valid prefix and M_i zeroing the rest; $\|L\|_2 \leq 1$ gives the upper bound. For the lower bound, L is an orthogonal projection, so $\sigma = \langle uv^\top, L(S)\rangle_F = \langle L^*(uv^\top), \tilde{S}\rangle_F \leq \|L^*(uv^\top)\|_F\|\tilde{S}\|_F \leq \|L^*(uv^\top)\|_F\sqrt{r_S}\,\sigma$; dividing by $\sigma > 0$ gives $\|L^*(uv^\top)\|_F \geq 1/\sqrt{r_S}$ and hence Eq. (4). Vanishing requires v constant on the valid prefix of every row with $u_i \neq 0$; causal prefixes are nested, so if the last query row is active this forces v constant on all of key space, whence $\tilde{S}v = 0$ (every centered row sums to zero over its valid keys), contradicting $\sigma(\tilde{S}) > s_0 > 0$. No factor in any case involves p. □

D ALGORITHM

Algorithm 1 SPECGATE, one training update

1: **given** λ, $s_0{=}50$, $w{=}10$, statistics sub-batch, $P{=}5$ power iterations
2: forward pass; compute task loss $\mathcal{L}_{\mathrm{task}}$
3: compute $s_{\ell,h}$ and $\bar{H}_{\ell,h}$ on the statistics sub-batch ▷ P power iterations; Eq. (1)
4: $g_{\ell,h} \leftarrow \mathrm{stopgrad}\,\sigma\big((s_{\ell,h} - s_0)/w\big)$
5: $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{task}} + \lambda\ \mathrm{mean}_{\ell,h}[g_{\ell,h}\,(1 - \bar{H}_{\ell,h})]$
6: backward; clip gradient norm at 1.0; optimizer step

The penalty with hinge adds $\lambda_s \operatorname{mean}_{\ell,h}[\operatorname{relu}(s_{\ell,h} - s_0)/s_0]$ to $\mathcal{L}$ (Appendix B); gated uniform cross-entropy replaces $1 - \bar{H}_{\ell,h}$ by $\mathrm{KL}(U_m \| p)/\log m$; the ungated controls replace $g_{\ell,h}$ by one; the mean gate gives every head $\operatorname{mean}_{\ell,h} g_{\ell,h}$.

Names. The released artifacts use earlier names. Table 5 maps them to the terms of the main text.

Table 5: **Names in the main text and in the artifacts.**

Main text	Artifacts and appendix tables
SPECGATE (gated entropy penalty)	`specgate`, gated Shannon, Shannon, G
penalty with hinge	SPECBREAK, `specbreak`
gated uniform cross-entropy	`klugate`, uniform-CE, uCE
ungated penalty	U
mean gate	M, E5
saturation-adaptive penalty (appendix only)	`adaptgate`, adaptive Shannon
confirmation set	E1, historical confirmation seeds 10–17
panel set	E2, seeds 83300–83315
added arms (†)	D01 alarm start, D02 QK-Clip, D03 uniform-CE
recovery study	`recovery_v27`

Aggregation and baseline details. For each layer/head, we estimate the norm separately for each sampled sequence, then average before applying the gate and hinge. Normalized entropy is averaged over those sequences and query rows with $m_i \geq 2$. The numerical treatment of the probability floor is described in Appendix U. The materialized path uses up to 8 probe sequences by default; the unfused probe size is configurable.

Norm estimation and scheduling. The materialized and unfused-probe paths differentiate through five power iterations. The matvec path detaches the estimated singular vectors and differentiates their final Rayleigh quotient. Its hinge acts every step; k controls only the additional entropy probe. The runs of the main text evaluate the penalty at every update, as in Algorithm 1. For causal scores, matrix-vector products factorize through prefix sums: $(\operatorname{tril}(S)v)_i = q_i \cdot \operatorname{cumsum}(v \odot k)_i / \sqrt{d_k}$, with a row-centering correction. These products are algebraically equivalent to the masked matrix operation; the finite-iteration norm and its gradient remain approximations.

E MONITORING PROTOCOL

Corpus (mechanical, no discretion): every completed language-model run whose method never suppresses or edits the statistic (untreated, EAR (Attanasio et al., 2022), JaSMin (Yudin et al., 2025) or z-loss (Chowdhery et al., 2023; Zoph et al., 2022); no delayed starts). **The outcome label is spectral-free**; so the alarm is never evaluated against a label built from its own statistic: collapsed $=$ final min-head entropy < 0.05 *and* final validation loss at least 0.25 nats above the median of entropy-healthy runs in the same (dataset, width) group; everything else, including the 48 entropy-flagged runs that finish near healthy loss, is non-collapsed and stays in the false-alarm denominator. This yields 108 collapsed and 119 non-collapsed runs. A group with no entropy-healthy in-corpus run has no reference, and its runs are labelled non-collapsed by default; in this corpus that is the (WikiText, width 1024) group, the five untreated 350M runs, all of which end entropy-flagged (two finish at 6.29). Three of them are among the $25/119$ spectral and five among the $69/119$ entropy false alarms, which is conservative for the spectral alarm; without them the counts are $22/114$ and $64/114$, the overlap upper bound of Figure 3 is $1{,}912$ in place of $5{,}696$, and the non-collapsed peak median is 150 in place of 151. The healthy reference is computed on these in-corpus runs and then applied unchanged to the held-out set, so one rule labels both, and Figure 3 uses the same labels. Grouping matches dataset and width but not training budget or batch size; thus the label does not isolate irreversible collapse from slower learning under a shorter budget. The reference also pools methods (at 30M, 34 untreated and 9 EAR runs with budgets from 800 to $50{,}000$ steps) and, at 124M, tokens per update; that group also holds nine GPT-2 fine-tuning runs judged against the from-scratch median. The bias runs both ways: harsh for short budgets and lenient for $5{,}000$-step runs, since an

untreated-only 5,000-step reference at 30M (4.666 in place of 4.754) would relabel four in-corpus runs as collapsed and no held-out run.

Alarm: $s_{\max} > 300$ on two consecutive 50-step readings. Detection among finally collapsed runs: an alarm at or before the first sustained validation degradation, or any alarm if no such event occurs. Degradation is the first of two consecutive validation readings at least 0.1 nats above the running minimum. Of 104 spectral detections, 48 have a degradation event and leads of 200–2,650 steps; 56 have none. A run can finish with poor loss without rising above its own earlier minimum, so the latter cases support collapse detection but not a measured early-warning lead. The monitor script, `scripts/monitor_both_alarms.py`, ships in the supplementary material and reproduces Table 6 from the released logs; the entropy-criterion comparator (min-head entropy < 0.05 twice) runs through the same script. Held-out protocol (preregistered before the runs): 14 fresh-seed runs spanning collapse-expected, healthy-expected, and boundary regimes; frozen rule applied unchanged; all outcomes reported. Of the seven held-out spectral detections, four have degradation events (leads 350–1400 steps) and three have none. Known limits: of the four in-corpus misses, two are injected mid-training LR spikes (`sp2_base_s3/s4`) whose validation loss jumps at step 2500 and whose alarm at 2600 comes 100 steps after that first elevated reading and 150 steps before the confirming reading at 2750; the other two (`a_base_lr1.2e-3_wu0_s4`, `cv_base_lr3e-3_st2000_s3`) never exceed $s_{\max} = 300$ (peaks 220 and 300) and never show a validation-loss increase; they are labelled collapsed only because their final loss exceeds the pooled-budget cutoff (one is a 2,000-step run, the other is within 0.005 nats and 0.002 entropy of the cutoffs); the hold-out is same-pipeline, not an external corpus, and at n=14 it is a pilot: 95% Clopper–Pearson intervals are $[0.59, 1]$ (sensitivity, $7/7$) and $[0, 0.41]$ (false-alarm rate, $0/7$). The in-corpus runs are correlated members of one experimental pipeline, not independent deployment draws; rates are within-pipeline operating characteristics. A deployment would add hysteresis or a longer consecutive-window requirement to trade false alarms against lead time; we report only the frozen rule.

Table 6: **The statistic as instrument.** Frozen alarm rule ($s_{\max} > 300$ on two consecutive readings), mechanical script `monitor_both_alarms.py` released; the outcome label is spectral-free (entropy flag + loss outcome; App. E), and entropy-flagged-but-near-healthy runs stay in every denominator. The advantage over an entropy-criterion alarm is specificity, not earliness (similar detection counts, 104 vs 106 of 108). Counts include alarms in finally collapsed runs with no degradation event; leads apply only when that event exists (App. E).

	spectral alarm	entropy-criterion alarm
in-corpus detection	104/108	106/108
in-corpus alarms on non-collapsed	25/119	69/119
held-out (14 preregistered runs)	7/7, 0/7 FA	7/7, 2/7 FA

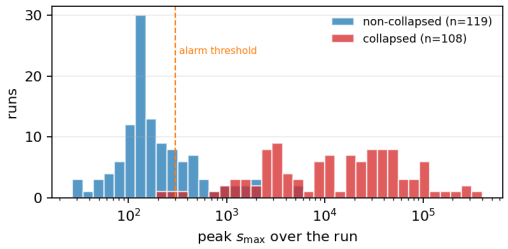

Figure 3: **Peak spectral level by outcome.** Largest $s_{\max}$ in each completed run without direct control of s, under the spectral-free outcome label of App. E, held-out corpus excluded (108 collapsed, 119 non-collapsed including the entropy-flagged runs). The dashed line is the monitor's threshold, which requires two consecutive readings above it; detection also depends on the loss trajectory, so the figure gives no alarm counts. Retrospective association; Table 6 evaluates warning performance.

What was frozen, and what we corrected. Frozen before the held-out runs: the alarm rule ($s_{\max} > 300$ on two consecutive readings), the entropy comparator, the outcome-label definition

together with its 0.05 and 0.25 constants, and the composition of the held-out set itself. Corrected afterwards: the population over which the healthy reference median is computed. It now uses in-corpus runs only and is applied unchanged to the held-out set, so held-out runs no longer contribute to the reference they are judged against. Under the two reference rules compared by the verification script, labels and scores are identical on the present corpus, in-corpus and held-out ($7/7$ and $0/7$ spectral, $7/7$ and $2/7$ entropy); that script ships with the others. The corpus grew during the campaigns and earlier manifests were not retained, so we do not claim the correction is inert in general.

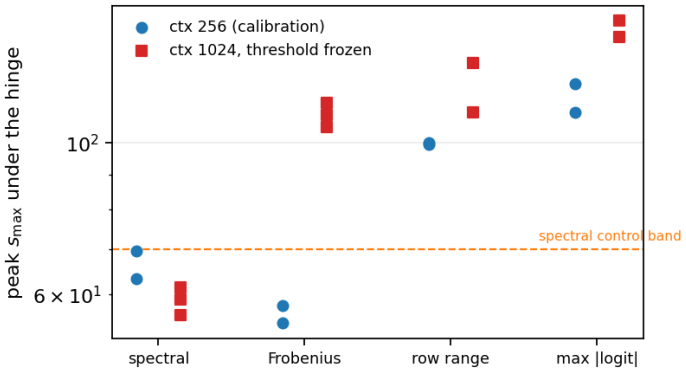

Figure 4: **Spectral control after a change of context length.** Peak $s_{\max}$ under matched hinges with thresholds calibrated at context 256 (circles) and *frozen* at context 1024 (squares). The spectral hinge keeps control (56–62); Frobenius, range, and max-logit hinges lose it (106–151), inside the bands registered before the runs. Cheaper statistics hold the spectrum at the calibration point and contain it less well away from it; their final losses stay close. Three seeds for the spectral and Frobenius hinges, two for the range and max-logit comparators.

F THE CONTROLLED RECOVERY STUDY

This appendix records the protocol, the runner and every per-seed value of the recovery study of Section 5.1; the alarm-started runs of Section 5.4 are described in Appendix O. Log names below are paths under `runs/recovery_v27/`.

Protocol and freeze. The protocol (`recovery_v27_protocol.md` under `experiments/recovery_v27/`) was written before any development run; its freeze record was completed on 2026-09-16 after the development results had been read and before the first confirmation prefix was launched, and the confirmation manifest (`manifest_conf.json`) records the commit of the frozen file. Nothing was deposited outside the repository; the record is local version control. Frozen quantities: the recipe (the 124M cell of App. G, 12 layers, 12 heads, width 768, context 256, WikiText-103, 16 sequences or 4,096 tokens per update, the statistics on a sub-batch of 4 sequences with five power iterations, AdamW $\beta = (0.9, 0.95)$, weight decay 0.1, gradient clipping 1.0, bf16 autocast, lr $1.2\text{e}{-3}$ with 500 warmup updates and cosine decay to zero over 5,000 updates, a monitor reading every 50 updates, validation every 250 updates on 40 batches, gate $s_0 = 50$ and $w = 10$ with both coefficients 1 at multiplier 1, alarm on two consecutive readings above 300, the regularizer at every update once active); the reference curve (the per-validation median of two runs of the same recipe at lr $3\text{e}{-4}$, seeds 3 and 4, `controls/ref_s3` and `ref_s4`, final 4.6407); the time-to-target threshold 4.7407 (reference final plus 0.10); the practical margin 0.02 nats; the confirmation seeds 10–17, fixed before any development result was read; the primary outcome (validation loss at update 5,000, SPECBREAK at the selected multiplier minus the untreated continuation, paired by seed, with a 95% paired t interval); the key comparator (gated uniform-CE); and the failure policy (a diverged arm ranks below any finite arm in the win/tie/loss ordering, two diverged arms give no ordering, an infrastructure failure may be retried once from the same state). No retry was needed: all 56 confirmation branches completed, 0 diverged, 0 errors (`outcome` events).

Runner. Every run uses one trainer, `src/train_recovery.py` with `recovery_losses.py`, `recovery_state.py` and `recovery_controller.py`; the historical trainers are not used for any arm. The sha256 of the four source files is written into the `config` event of every log and is identical across all 142 logs (the `source_revision` field records the enclosing repository commit, which advanced as manifests and scorers were committed while runs were in flight; the runner did not change). Three checks precede the study. Replay: an untreated run of seed 3 on the A40 node that trained the released `big_base_lr1.2e-3_s3` reproduces it (`check/replay_s3_mirella`: final validation loss 5.4640, final task loss 5.6518, alarm after update 950, the released values), and the two development prefixes end at the released same-node finals of seeds 3 and 4 (5.5160 and 5.7436). Restart identity: on the A100 node, a 300-update run with its state saved after update 150 (`check/gpu_restart_a`) and two branches from that state (`gpu_restart_b`, `gpu_restart_c`) log identical train, monitor and validation records from update 151 on, so a branch continues its parent bitwise. Branch verification: a branch loads the parent's weights, optimizer moments, schedule position and data-stream state, checks the checkpoint's sha256 against its sidecar and the hash of the next eight batches recorded there, and refuses any command-line override that is not declared for the arm (method, multiplier or learning-rate factor); the parent hash and the declared overrides are written into the branch's `config` event. Index convention: updates are numbered 1 to 5,000; the monitor reading at update u is taken on the forward pass of update u and processed after it; the validation at u evaluates the model after u updates; a branch from the state after update u makes $u + 1$ its first, treated, update; the legacy logs of the rest of the paper record step k for the model before update $k + 1$.

Eligibility. A prefix qualifies when two consecutive validations meet the label of Section 3, min-head entropy below 0.05 and validation loss at least 0.25 nats above the reference curve at that update, with the second validation at update $\geq$ 1,000 and the first at $\leq$ 2,500; the state after the second validation's update is saved and is the common parent of every arm. The reference is the per-update curve of the two lr 3e$-$4 runs, not the corpus group median of App. E. A prefix that never qualifies stays in the denominator; none occurred (2/2 development, 8/8 confirmation). Table 7 lists per seed the alarm, the first qualifying validation, the fork and the finals of the untreated prefix and of the same-seed healthy control (the reference recipe at lr 3e$-$4, monitor armed).

Table 7: **Alarm and eligibility per seed.** Alarm: the update of the second consecutive monitor reading above 300 in the untreated prefix. First qualifying validation: the first validation at which min-head entropy was below 0.05 and loss at least 0.25 nats above the reference curve; fork: the update of the second consecutive such validation, after which the state was saved. Finals are validation loss at update 5,000 of the untreated prefix and of the same-seed healthy control (lr 3e$-$4, monitor armed); – means the control raised no alarm. The development seeds were used for tuning only.

seed	stage	alarm	first qualifying validation	fork	untreated final	control final	control alarm
3	development	900	1,500	1,750	5.516	4.648	–
4	development	850	1,250	1,500	5.744	4.634	–
10	confirmation	850	1,250	1,500	5.819	4.637	–
11	confirmation	850	1,250	1,500	5.593	4.639	–
12	confirmation	800	1,250	1,500	5.716	4.650	–
13	confirmation	800	1,250	1,500	5.720	4.654	–
14	confirmation	850	1,250	1,500	5.775	4.637	–
15	confirmation	850	1,500	1,750	5.689	4.641	–
16	confirmation	900	1,500	1,750	5.499	4.655	–
17	confirmation	850	1,250	1,500	5.745	4.640	–

Development and selection. Two development seeds (3 and 4) supplied the tuning grid: coefficient multipliers 0.3, 1 and 3 for the four loss terms (the multiplier scales the gate coefficient and the hinge coefficient together; s_0, w, the alarm threshold and the cadence stay fixed) and factors 0.5, 0.2 and 0.1 for the learning-rate arm, which multiplies the remaining schedule. The selection rule, written before the grid ran: among a method's settings with finite loss on both seeds, the lowest mean final loss, ties toward multiplier 1, then the smaller multiplier. Three methods selected the boundary setting 3; the predeclared expansion by one step to 10 was worse for all three (Table 8), so 3 remained selected and is interior. The selected SPECBREAK multiplier is 3, a tuned variant; the

unit setting was confirmed alongside as its own arm. Table 8 is two seeds and not confirmatory. It is the source of the coefficient-sensitivity statement of Section 5.1: the SPECGATE term at multiplier 1 restores learning (4.83 against untreated 5.63) but ends 0.11 behind the selected SPECBREAK setting, and at $0.3\times$ it ends at 5.39, near the learning-rate arm, while SPECBREAK and uniform-CE stay within 0.03 across $0.3\times$ to $3\times$.

Table 8: **Development grid (two seeds, not confirmatory).** Mean final validation loss at update 5,000 over seeds 3 and 4 per method and coefficient multiplier, with the per-seed values beneath; for the learning-rate arm per factor applied to the remaining schedule. The $\times 10$ column is the predeclared one-step expansion, run for the three methods whose selected setting lay at the grid boundary. Selected by the frozen rule: SPECBREAK $\times 3$, uniform-CE $\times 1$, SPECGATE $\times 3$, saturation-adaptive Shannon $\times 3$, learning rate $\times 0.5$. Source: `score_dev.json` and the six expansion logs (`dev/s*_*_m10`).

method	$\times 0.3$	$\times 1$	$\times 3$	$\times 10$
SPECBREAK	4.7465	4.7231	4.7177	4.7305
seeds 3/4	4.773/4.720	4.751/4.695	4.747/4.689	4.757/4.704
gated uniform-CE	4.7219	4.7190	4.7255	
seeds 3/4	4.746/4.698	4.742/4.696	4.750/4.701	
gated Shannon	5.3875	4.8315	4.7199	4.7271
seeds 3/4	5.393/5.382	4.870/4.793	4.748/4.692	4.754/4.700
saturation-adaptive Shannon	5.4553	4.8188	4.7213	4.7387
seeds 3/4	5.347/5.563	4.839/4.799	4.748/4.694	4.763/4.714
learning rate	$\times 0.5$	$\times 0.2$	$\times 0.1$	
mean	5.3181	5.3622	5.4473	
seeds 3/4	5.269/5.368	5.326/5.399	5.402/5.492	
untreated	5.6298	seeds 3/4: 5.516/5.744		
healthy control (lr $3e{-4}$)	4.6407	seeds 3/4: 4.648/4.634		

Confirmation, per seed. Table 9 is the generated summary of Section 5.1 with its footer, Table 10 the per-seed finals, the first validation at or below the target and the normalized area under the post-fork validation-loss curve. In every case that reached the target the confirming validation was the next one, 250 updates later.

Table 9: **Confirmation summary with the generator's footer.** The table of Section 5.1 as written by `gen_recovery_table.py` from `manifest_conf.json` and `score.json`; columns as there.

arm	final loss	Δ [95% interval]	w/t/l	reached	gap
SPECBREAK $\times 3$	4.712 ± 0.030	$-0.983\ [-1.078, -0.888]$	8/0/0	6/8 (2,874)	0.067
SPECBREAK $\times 1$ (unit coefficients)	4.718 ± 0.031	$-0.976\ [-1.072, -0.881]$	8/0/0	5/8 (2,749)	0.074
gated uniform-CE $\times 1$	4.718 ± 0.030	$-0.976\ [-1.068, -0.884]$	8/0/0	6/8 (2,874)	0.074
gated Shannon $\times 3$	4.716 ± 0.034	$-0.978\ [-1.072, -0.884]$	8/0/0	6/8 (2,749)	0.072
saturation-adaptive Shannon $\times 3$	4.718 ± 0.033	$-0.977\ [-1.071, -0.882]$	8/0/0	5/8 (2,749)	0.073
learning rate $\times 0.5$	5.332 ± 0.077	$-0.363\ [-0.442, -0.284]$	8/0/0	0/8	0.687
untreated	5.695 ± 0.103			0/8	1.050

seeds 10/11/12/13/14/15/16/17; eligible prefixes 8/8 (planned 8)
fork s10 1,500, s11 1,500, s12 1,500, s13 1,500, s14 1,500, s15 1,750
s16 1,750, s17 1,500; horizon 5,000 updates; healthy final 4.6407
target 4.7407 (healthy final + 0.10)
Δ: arm minus untreated at the horizon, paired by seed, mean [95% paired t interval]
w/t/l: seeds won/tied/lost; gap: mean of arm minus same-seed control
final loss: mean $\pm$ sd over finite seeds
reached: seeds at the target / seeds evaluated (median time from activation)

Table 10: **Confirmation per seed.** Top: final validation loss at update $5{,}000$. Middle: the first validation at or below the target 4.741 (reference final plus 0.10); the target counts as reached when the next validation, 250 updates later, also qualifies, which held in every reached case; n.r. means not reached within the horizon. Bottom: normalized area under the post-fork validation-loss curve (trapezoid from the fork, with the prefix's validation at the fork prepended, to update $5{,}000$, divided by the span); lower is better. Source: `score.json`.

final validation loss at update $5{,}000$ arm	s10	s11	s12	s13	s14	s15	s16	s17
SPECBREAK $\times 3$	4.685	4.676	4.749	4.707	4.709	4.754	4.731	4.683
SPECBREAK $\times 1$ (unit coefficients)	4.692	4.677	4.755	4.718	4.712	4.761	4.743	4.691
gated uniform-CE $\times 1$	4.698	4.679	4.757	4.711	4.720	4.759	4.734	4.688
gated Shannon $\times 3$	4.706	4.673	4.760	4.706	4.705	4.765	4.736	4.680
saturation-adaptive Shannon $\times 3$	4.701	4.675	4.758	4.710	4.711	4.766	4.739	4.683
learning rate $\times 0.5$	5.359	5.205	5.355	5.377	5.345	5.459	5.276	5.278
untreated	5.819	5.593	5.716	5.720	5.775	5.689	5.499	5.745
healthy control (lr $3\mathrm{e}{-4}$)	4.637	4.639	4.650	4.654	4.637	4.641	4.655	4.640
first validation at or below the target 4.741 *(confirmed at the next one)*								
SPECBREAK $\times 3$	4,250	4,250	n.r.	4,500	4,500	n.r.	4,750	4,250
SPECBREAK $\times 1$ (unit coefficients)	4,250	4,250	n.r.	4,500	4,500	n.r.	n.r.	4,250
gated uniform-CE $\times 1$	4,250	4,250	n.r.	4,500	4,500	n.r.	4,750	4,250
gated Shannon $\times 3$	4,250	4,000	n.r.	4,500	4,250	n.r.	4,750	4,250
saturation-adaptive Shannon $\times 3$	4,250	4,250	n.r.	4,500	4,500	n.r.	n.r.	4,250
learning rate $\times 0.5$	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
untreated	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
normalized area under the post-fork validation-loss curve								
SPECBREAK $\times 3$	5.011	4.986	5.068	5.025	5.030	5.048	5.010	4.997
SPECBREAK $\times 1$ (unit coefficients)	5.018	4.988	5.077	5.035	5.033	5.052	5.020	5.006
gated uniform-CE $\times 1$	5.014	4.987	5.068	5.025	5.031	5.046	5.010	5.000
gated Shannon $\times 3$	5.036	4.986	5.081	5.024	5.031	5.063	5.017	4.995
saturation-adaptive Shannon $\times 3$	5.031	4.986	5.077	5.027	5.034	5.064	5.018	4.998
learning rate $\times 0.5$	5.529	5.367	5.510	5.523	5.499	5.597	5.411	5.439
untreated	5.968	5.731	5.882	5.852	5.905	5.874	5.672	5.875

Recovery times. Table 11 times the secondary outcomes from the released logs (`score_recovery_extras.py`): the first monitor reading after activation with $s_{\max} < 100$ (monitor stream, a reading every 50 updates; the first reading after activation lies 49 updates after it) and the first validation after activation with min-head entropy at or above 0.05 (validation stream, every 250 updates; the first lies 249 after activation). Final $s_{\max}$ is the last monitor reading and final min-head entropy the validation at update $5{,}000$; the per-run table of `score.md` reports the train-stream value of the last train record instead, which differs at the third decimal. The hinge and uniform-CE arms bring $s_{\max}$ below 100 at the first reading in $8/8$ seeds; the Shannon terms at $3\times$ need 150–300 updates. No arm restores the controls' final min-head entropy of 0.19–0.26.

Table 11: **Spectral and entropy recovery after activation.** Updates from activation to the first monitor reading with $s_{\max} < 100$ (readings every 50 updates, the first one 49 updates after activation) and to the first validation with min-head entropy at or above 0.05 (validations every 250 updates, the first one 249 after activation), over 8 seeds; median and range over the seeds that reached the criterion; n.r. means no seed reached it within the horizon. Final $s_{\max}$ is the last monitor reading, final min-head entropy the validation at update 5,000 (`extras.json`; the control's values from `live.json`).

arm	$s_{\max} < 100$		min-head entropy ≥ 0.05		final $s_{\max}$	final min-head entropy	
	reached	median (range)	reached	median (range)	range	range	≥ 0.05
SpecBreak ×3	8/8	49 (49–49)	5/8	1,499 (249–1,749)	36–44	0.018–0.091	4/8
SpecBreak ×1 (unit)	8/8	49 (49–49)	2/8	1,374 (249–2,499)	41–49	0.013–0.051	1/8
uniform-CE ×1	8/8	49 (49–49)	8/8	249 (249–499)	26–35	0.050–0.077	8/8
Shannon ×3	8/8	199 (149–299)	5/8	1,749 (249–1,999)	36–46	0.017–0.080	5/8
adaptive Shannon ×3	8/8	174 (149–249)	7/8	499 (249–1,249)	30–42	0.041–0.160	7/8
learning rate ×0.5	0/8	n.r.	0/8	n.r.	1,512–2,096	0.002–0.004	0/8
untreated	0/8	n.r.	0/8	n.r.	1,888–3,616	0.001–0.003	0/8
healthy control						0.195–0.258	8/8

Figures. Figure 5 shows the validation-loss trajectory of every confirmation seed and Figure 6 the monitor-stream $s_{\max}$ and the train-stream min-head entropy of seed 10.

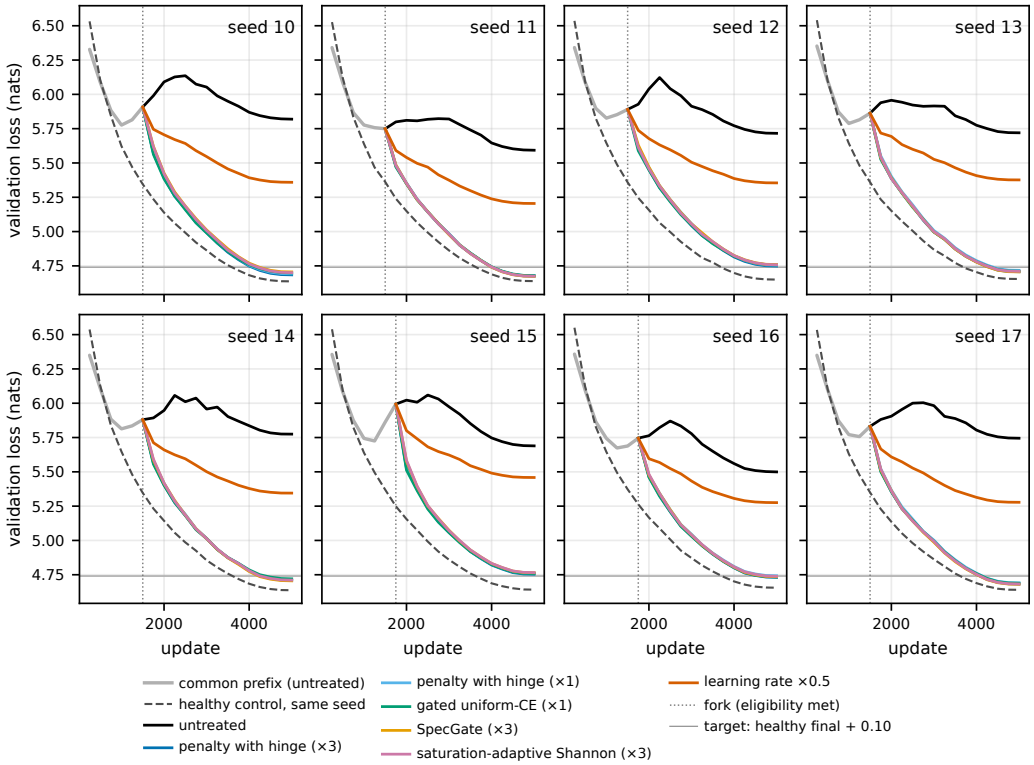

Figure 5: **Recovery from an established collapse, all confirmation seeds.** Validation loss (validation stream, every 250 updates) against update for each of the eight confirmation seeds: the common untreated prefix in grey up to the fork (dotted line, the update after the second label-qualifying validation), one line per arm from the fork, the same-seed healthy control at lr $3e{-4}$ dashed, and the target (reference final plus 0.10) as the horizontal line. Every arm of a seed starts from the same saved state; the untreated arm is the prefix continued.

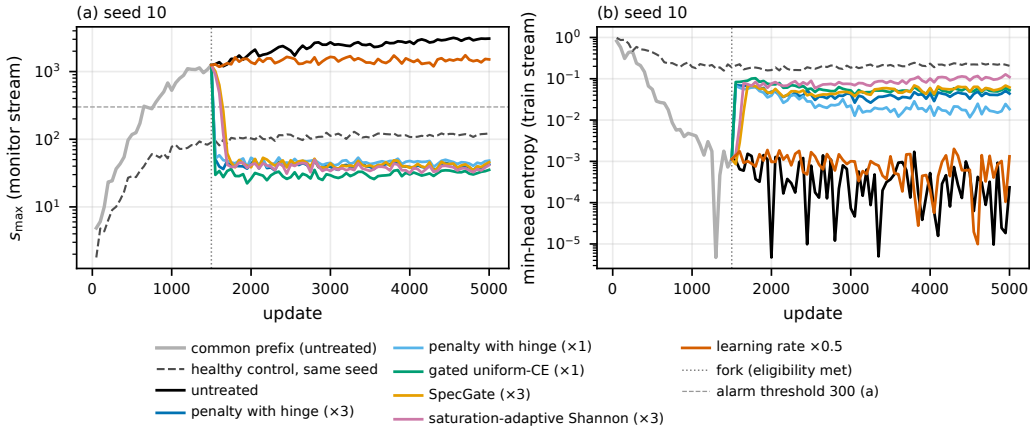

Figure 6: **Secondary outcomes, seed** 10**.** (a) $s_{\max}$ from the monitor stream (a reading every 50 updates, log scale) with the alarm threshold 300; (b) min-head entropy from the train stream (the statistics sub-batch of the logged update, log scale). The fork is the dotted line. The hinge and uniform-CE arms hold $s_{\max}$ near the gate from the first reading after activation, the Shannon terms at 3× reach it after 150–300 updates, and the learning-rate arm and the untreated continuation stay above 1,000; the min-head entropy of every repaired arm stays below the control's.

Released logs. `runs/recovery_v27/` holds the 142 logs of the study by stage: `conf` (56 confirmation branches), `dev` (38 development branches, the six expansion runs included), `live` (16), `bench` (8), `conf_prefix` (8), `conf_controls` (8), `check` (4), `controls` (2) and `prefix` (2); `MANIFEST.json` gives the sha256 of each raw and released log (only the data path and the host name are neutralised; every other field is verbatim). The saved states (1.5 GB each) are not released; the parent hash and the declared overrides of every branch are in its `config` event. The scorers `score_recovery_v27.py`, `score_live_v27.py`, `score_recovery_extras.py` and `score_bench_v27.py` and the generators `gen_recovery_table.py`, `gen_recovery_appendix_tables.py` and `make_recovery_figures.py` regenerate every table and figure of this study from the logs; `experiments/recovery_v27/LOG.md` is the timeline. The study used 17.3 GPU-hours (the sum of the `elapsed` field of each log's `done` event: confirmation branches 6.8, development 4.6, live runs 3.0, prefixes and controls 2.5, checks and benchmark 0.4), reported separately from the 177.6 hours of the corpus.

G Reproducibility and Compute

Optimizer and cell dimensions. All language-model runs use AdamW with β=(0.9, 0.95), weight decay 0.1 on matrices and gradient clipping 1.0, under a cosine schedule unless a constant rate is stated. The main text uses the 124M cell; the 30M cell (6 layers, 8 heads, width 384, context 256) is used in the high learning-rate study, at 4,096 tokens per step; the 124M cell keeps the same tokens per step and the historical 350M cell uses 2,048; new development recipes specify their own token batch in App. H. The supplementary material contains the training scripts (materialized, probe, and matvec variants), the exact command line of every run (regenerable from each log's own config record), all 763 per-run JSONL logs (177.6 logged GPU-hours of single-GPU training; historical numbers are recomputable from them by released scripts), the preregistration artifacts with commit hashes, and the monitor script `scripts/monitor_both_alarms.py`. Model cells: 30M is 6×8×384, 124M 12×12×768 and 350M 24×16×1024 (layers, heads, width), all at context 256. The recovery study adds 142 logs and 17.3 GPU-hours under its own manifest (`runs/recovery_v27/MANIFEST.json`), reported separately from the 763 records and 177.6 hours above.

H COMPLETED DEVELOPMENT PANELS AND FINAL EVIDENCE SCOPE

The new panels below were designed and completed as development studies. They are separate from the historical eight-seed confirmation of repair and alarm start. These development panels did not proceed to a new confirmation panel. The subsequently completed QK-Clip and high-LR comparisons have separate protocols and evidence status (Appendices P and J). Selection on the reported development seeds does not provide an unbiased estimate of a selected policy's future performance. We therefore report all assigned outcomes and descriptive paired differences, without new superiority, equivalence or noninferiority tests. The planned 0.02-nat practical margin is context, not a license to infer equivalence from nonsignificance.

Artifact organization. The final release includes the original reports as portable exports, their original and exported SHA256 digests, all per-seed presentation tables, and a CPU-only renderer. Numerical values, statuses and source digests are unchanged by export; account names and machine locations are anonymized. Embedded original digests authenticate original artifacts, not renamed export files. Historical logs and their scorers remain separate from the new development reports. Raw checkpoints are retained outside the compact supplement. The source snapshots, bindings and amendment records specify the actual execution, including storage changes; source code prepared but never executed is not evidence.

H.1 OPERATIONAL RECOVERY AT A FIXED REMAINING BUDGET

The model is a 124M decoder with 12 layers, 12 heads, width 768, context 256 and $4{,}096$ tokens/update on WikiText-103. Four development streams (seeds 1000–1003) are replayed to confirmed collapsed states, preserving the full model, optimizer, schedule, data, RNG and controller state. Interventions branch from each stream's saved decision state; rollback uses the latest checkpoint that passes the declared healthy rule using information available at that point, and restart uses the saved original initialization. No future checkpoint is used to choose rollback. The unrepaired continuation shares the collapsed parent.

The ordinary continuations retain the original cosine horizon of $5{,}000$ updates, peak LR 0.0012 and warmup 500. LR rescue multiplies that schedule by 0.1, 0.2 or 0.5, with rewarmup 0, 100 or 500 updates. Rollback and restart use explicitly declared relative cosine schedules of $5{,}000$ updates from their selected origin, peak LR 0.00015, 0.0003 or 0.0006, and warmup 0, 100 or 500. Thus each operational family has nine candidates; the selected family member minimizes the mean loss among candidates with all four valid endpoints. The selected rollback uses LR 0.0006, warmup 0; restart uses 0.0006, warmup 500; LR rescue uses factor 0.5 and 100 rewarmup updates. These are tested alternatives, not an exhaustive optimum over restart or schedule policies.

The three selected loss policies use coefficient multiplier 3, a 250-update treatment window, reactivation after two monitor readings above 300, and at most 750 active updates. The monitor runs every 50 updates; missing readings are ignored by the frozen rule. Selection used the earlier equal-search development policy panel, with all prior information retained in its records. That earlier panel's runner-only budget excluded external staging/archive costs; its endpoints are not substituted into this full-policy comparison. A continuous SPECBREAK arm and untreated continuation complete the 32-candidate panel.

Cost and endpoints. Each assignment has a 300-second one-A40 policy budget. The measured boundary includes state/data staging, restore, setup, training, monitoring, validation, final-state archival and cleanup. The shared prefix cost is excluded from this salvage comparison because it is already spent. Scheduler/container startup, metadata authentication and CPU analysis lie outside the measured policy boundary. This is not a complete service-level latency or whole-training comparison. All four seeds of all 32 candidates have valid within-budget final-state endpoints. Summed policy cost is 10.221 A40 GPU-hours; it is not total sprint compute. Validation runs every 250 updates plus a budget-final evaluation, with 40 batches per validation. Reserved time for finalization leaves slack below the nominal budget; the report retains actual model/evaluation times and total costs, without interpolating between checkpoints or extrapolating to the deadline.

The healthy target is the median final loss of four independent healthy development references (seeds 1007–1010) plus 0.10 nats: 4.743889. Attainment requires two qualifying readings. All 128 runs

remain censored at their last valid observation; none attains the target. Accordingly we make no time-to-target reduction claim. Intermittent policies have follow-up within this budget, not a long-term relapse guarantee. All 128 endpoint/cost/censoring rows and actual validation trajectories are supplied electronically.

Administrative amendments during execution (collection, storage and analysis records) are retained in the study logs; they changed no scientific setting.

H.2 JOINT WHOLE-TRAINING TUNING

The 124M/context-256 model, corpus and effective token batch match the operational shape, but these runs start from initialization. Each family has 12 configurations and three development seeds (1004–1006). Untreated and QK-norm use peak LRs $\{0.0003, 0.0006, 0.0012, 0.0024\}$ crossed with warmups $\{500, 1000, 2000\}$. Each penalty uses LRs $\{0.0006, 0.0012, 0.0024\}$, warmups $\{500, 2000\}$ and coefficient multipliers $\{1, 3\}$. This gives the same number of candidates and the same per-candidate budget, while using prior historical information to choose family-specific grids. The grid is not an exhaustive common hyperparameter domain. The separate repair-policy search supplies no additional whole-training candidates.

All penalties are available continuously from update one. QK-norm is active architecturally from initialization. Monitoring and validation attention diagnostics are omitted for this deployment-cost comparison; unused spectral/entropy graphs are not imposed on the untreated arm. Evaluation uses the same fixed validation protocol every 250 updates and at the budget boundary. Training uses BF16, AdamW with clipping 1, Python 3.13.11, PyTorch 2.10.0 with CUDA 12.8, and one A40 per run. The schedule horizon is 25,000 updates throughout; no arm extends a completed zero-LR schedule or resets its denominator.

The budget is 1,200 allocated seconds per assignment. Each five-arm block assigns measured shared staging and cleanup equally to its five members. Final enclosing block receipts are authoritative for actual allocated cost; all 36 blocks and 180 assignment receipts were terminal before the report read outcomes. Selection requires all three seeds to have finite valid endpoints and total attributed cost within budget. The report separately retains descriptive half-budget readings and the evaluation at update 4,000 where available. The latter is a secondary equal-update comparison; it does not replace the primary cost endpoint. All 180 assignments complete. The 170 strict-valid endpoints leave 50 all-three-valid candidates, ten per family. The ten excluded endpoints are all five arms of each of the two NAS-backed blocks: actual attributed cost is 1,201.653–1,205.122 seconds. Their finite observations remain in the released CSV and are marked with a dagger in the grid; no surviving-seed mean is used for those candidates. Total measured block allocation is 215,012.996 A40 GPU-seconds (59.726 GPU-hours). This is whole-training search cost, not the full sprint cost. All seeds, nonqualifying candidates and actual token counts are released.

H.3 TRANSFER FEASIBILITY AND HEALTHY-ONLY RECALIBRATION

The original conditions change model size (350M), context length (124M/context 1024) or corpus (124M/OpenWebText) from the source recipe, with separate prefixes for every shape. All 12 healthy runs completed; the eight context/corpus references passed and their eight stress prefixes were eligible. All four 350M references failed the absolute minimum-head-entropy admission rule, so their four dependent prefixes were not launched and remain *dependency unavailable* in the original assigned denominator. Improving task losses in these references caution against labeling low minimum-head entropy alone as training failure.

The revised 350M healthy rule was chosen after that failure. It uses matched-update medians from the original four task-healthy development references, not collapsed-run recovery outcomes. From update 1000, two consecutive validations with mean-head entropy below half the matched healthy median fail admission. Two readings of loss more than 0.10 nats above the previous best, or above the matched healthy median, also fail. Minimum-head entropy remains diagnostic. Four new development reference seeds (1087–1090) all pass this rule; their median loss curve then labels four stress prefixes. Their peak learning rates remain 0.0003 and 0.0012, respectively. The stress-prefix eligibility rule still requires minimum-head entropy below 0.05 and a task gap of 0.25 nats; all four

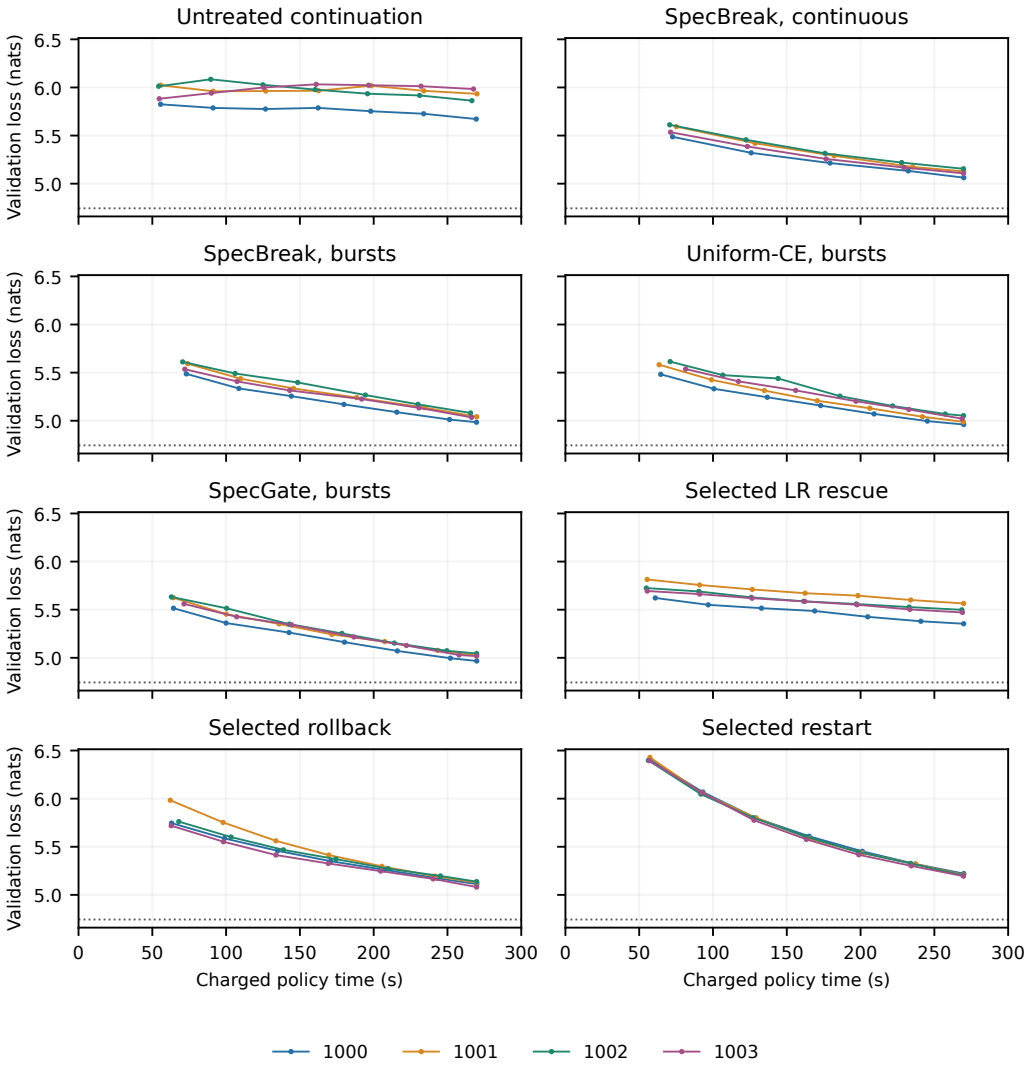

Figure 7: **Observed operational development trajectories.** Four seed traces per policy; lines connect actual validation readings and do not represent interpolated equal-time mean losses. The dotted line is the fixed healthy target. All arms complete within the charged policy budget; none attains the two-reading target. The last three policies are selected within their development grids.

Table 12: All 32 operational candidates and all four development seeds. Every entry is final-model validation loss at the 300-second budget. Names encode the coefficient, treatment policy, LR scale and warmup; `klugate` denotes uniform-CE and `specgate` SPECGATE. No failed, unavailable or target-unreached run is removed.

Candidate	1000	1001	1002	1003	Mean
lr_rescue_s0.1_w0	5.4607	5.6483	5.5571	5.5096	5.5439
lr_rescue_s0.1_w100	5.4691	5.6538	5.5634	5.5145	5.5502
lr_rescue_s0.1_w500	5.4929	5.6695	5.5757	5.5234	5.5654
lr_rescue_s0.2_w0	5.3945	5.6013	5.4775	5.4241	5.4744
lr_rescue_s0.2_w100	5.4059	5.5722	5.4964	5.4461	5.4801
lr_rescue_s0.2_w500	5.4184	5.6077	5.5096	5.4664	5.5005
lr_rescue_s0.5_w0	5.3765	5.5693	5.4820	5.4910	5.4797
lr_rescue_s0.5_w100	5.3539	5.5647	5.4987	5.4712	5.4721
lr_rescue_s0.5_w500	5.3877	5.6049	5.5086	5.4692	5.4926
rollback_lr0.00015_w0	5.2828	5.3125	5.2894	5.2606	5.2863
rollback_lr0.00015_w100	5.2880	5.3129	5.2895	5.2592	5.2874
rollback_lr0.00015_w500	5.3283	5.3746	5.3315	5.3009	5.3338
rollback_lr0.0003_w0	5.1368	5.1531	5.1513	5.1160	5.1393
rollback_lr0.0003_w100	5.1415	5.1618	5.1570	5.1245	5.1462
rollback_lr0.0003_w500	5.1911	5.2152	5.2041	5.1684	5.1947
rollback_lr0.0006_w0	5.1117	5.1279	5.1375	5.0813	5.1146
rollback_lr0.0006_w100	5.1170	5.1327	5.1198	5.1015	5.1178
rollback_lr0.0006_w500	5.1540	5.1625	5.1448	5.1334	5.1486
restart_lr0.00015_w0	5.3804	5.3614	5.3585	5.3513	5.3629
restart_lr0.00015_w100	5.3785	5.3714	5.3502	5.3462	5.3616
restart_lr0.00015_w500	5.3745	5.3749	5.4042	5.3681	5.3804
restart_lr0.0003_w0	5.2677	5.2590	5.2654	5.2486	5.2602
restart_lr0.0003_w100	5.2460	5.2413	5.2245	5.2157	5.2319
restart_lr0.0003_w500	5.2519	5.2493	5.2529	5.2337	5.2469
restart_lr0.0006_w0	5.5211	5.4625	5.4629	5.4740	5.4801
restart_lr0.0006_w100	5.2068	5.2126	5.2211	5.2126	5.2133
restart_lr0.0006_w500	5.2211	5.2082	5.2022	5.1950	5.2066
untreated	5.6716	5.9335	5.8627	5.9836	5.8628
specbreak_c3_continuous	5.0625	5.1278	5.1548	5.1076	5.1132
specbreak_c3_reactivate250cap750	4.9848	5.0416	5.0809	5.0352	5.0356
klugate_c3_reactivate250cap750	4.9609	4.9887	5.0527	5.0212	5.0059
specgate_c3_reactivate250cap750	4.9668	5.0234	5.0449	5.0177	5.0132

Table 13: Whole-training grid, candidates 1–30 in frozen order. Final loss at 1,200 allocated seconds, all development seeds. c is the coefficient multiplier. A dagger denotes a finite descriptive endpoint whose final total cost exceeded the budget; it is ineligible for selection. A dash denotes an inapplicable coefficient or unavailable all-three-valid mean.

Method	Peak LR	Warmup	c	1004	1005	1006	Mean
Untreated	0.0003	500	–	4.2306	4.2257	4.2304	4.2289
Untreated	0.0003	1000	–	4.2204	4.2110	4.2220	4.2178
Untreated	0.0003	2000	–	4.2378	4.2264	4.2282	4.2308
Untreated	0.0006	500	–	4.2719	4.2808	4.2672	4.2733
Untreated	0.0006	1000	–	4.2888	4.2880	4.2778	4.2849
Untreated	0.0006	2000	–	4.2480†	4.2580	4.2522	–
Untreated	0.0012	500	–	6.7746	7.0190	7.1127	6.9688
Untreated	0.0012	1000	–	6.7005	6.8843	6.9189	6.8346
Untreated	0.0012	2000	–	4.3855	4.3794	4.3655	4.3768
Untreated	0.0024	500	–	7.2346	7.2499	7.3283	7.2709
Untreated	0.0024	1000	–	7.1790	7.3737	7.2719	7.2749
Untreated	0.0024	2000	–	7.3535†	7.2848	7.4042	–
QK-norm	0.0003	500	–	4.2884	4.3123	4.3090	4.3032
QK-norm	0.0003	1000	–	4.2670	4.2772	4.2895	4.2779
QK-norm	0.0003	2000	–	4.2863	4.3040	4.2930	4.2944
QK-norm	0.0006	500	–	4.3327	4.3394	4.3414	4.3378
QK-norm	0.0006	1000	–	4.3250	4.3342	4.3463	4.3352
QK-norm	0.0006	2000	–	4.3426†	4.3592	4.3539	–
QK-norm	0.0012	500	–	4.3950	4.4150	4.4130	4.4077
QK-norm	0.0012	1000	–	4.3797	4.3875	4.3959	4.3877
QK-norm	0.0012	2000	–	4.3875	4.4155	4.4225	4.4085
QK-norm	0.0024	500	–	4.4700	4.4787	4.4938	4.4808
QK-norm	0.0024	1000	–	4.4738	4.4766	4.4919	4.4808
QK-norm	0.0024	2000	–	4.4804†	4.4770	4.4789	–
SpecBreak	0.0006	500	1.0	4.4546	4.4741	4.4622	4.4636
SpecBreak	0.0006	500	3.0	4.4526	4.4668	4.4748	4.4647
SpecBreak	0.0006	2000	1.0	4.4674	4.4774	4.4762	4.4737
SpecBreak	0.0006	2000	3.0	4.4882	4.4868	4.4746	4.4832
SpecBreak	0.0012	500	1.0	4.5120	4.5396	4.5295	4.5270
SpecBreak	0.0012	500	3.0	4.5212†	4.5160	4.5251	–

Table 14: Whole-training grid, candidates 31–60 in frozen order. Same budget, validity rule and development seeds as the preceding table.

Method	Peak LR	Warmup	c	1004	1005	1006	Mean
SpecBreak	0.0012	2000	1.0	4.5451	4.5416	4.5452	4.5440
SpecBreak	0.0012	2000	3.0	4.5387	4.5375	4.5303	4.5355
SpecBreak	0.0024	500	1.0	4.6704	4.6585	4.6566	4.6619
SpecBreak	0.0024	500	3.0	4.5873	4.6097	4.5922	4.5964
SpecBreak	0.0024	2000	1.0	4.6140	4.6050	4.6223	4.6138
SpecBreak	0.0024	2000	3.0	4.6059†	4.5945	4.6100	–
Uniform-CE	0.0006	500	1.0	4.3996	4.4285	4.3940	4.4074
Uniform-CE	0.0006	500	3.0	4.3935	4.4018	4.3944	4.3966
Uniform-CE	0.0006	2000	1.0	4.4570	4.4744	4.4814	4.4709
Uniform-CE	0.0006	2000	3.0	4.4201	4.4538	4.4265	4.4335
Uniform-CE	0.0012	500	1.0	4.4457	4.4749	4.4599	4.4602
Uniform-CE	0.0012	500	3.0	4.4407†	4.4321	4.4345	–
Uniform-CE	0.0012	2000	1.0	4.4825	4.5131	4.5052	4.5002
Uniform-CE	0.0012	2000	3.0	4.4609	4.4676	4.4667	4.4650
Uniform-CE	0.0024	500	1.0	4.8630	4.5840	4.5787	4.6753
Uniform-CE	0.0024	500	3.0	4.5433	4.5664	4.5624	4.5574
Uniform-CE	0.0024	2000	1.0	4.5525	4.5698	4.5623	4.5615
Uniform-CE	0.0024	2000	3.0	4.5514†	4.5561	4.5439	–
SpecGate	0.0006	500	1.0	4.3704	4.3720	4.3623	4.3683
SpecGate	0.0006	500	3.0	4.3775	4.3828	4.3741	4.3781
SpecGate	0.0006	2000	1.0	4.3904	4.3679	4.3584	4.3722
SpecGate	0.0006	2000	3.0	4.3998	4.3958	4.3861	4.3939
SpecGate	0.0012	500	1.0	4.4375	4.4503	4.4366	4.4415
SpecGate	0.0012	500	3.0	4.4342†	4.4336	4.4351	–
SpecGate	0.0012	2000	1.0	4.4488	4.4664	4.4555	4.4569
SpecGate	0.0012	2000	3.0	4.4389	4.4363	4.4266	4.4340
SpecGate	0.0024	500	1.0	4.5723	4.5642	4.5861	4.5742
SpecGate	0.0024	500	3.0	4.5273	4.5344	4.5172	4.5263
SpecGate	0.0024	2000	1.0	4.5265	4.5206	4.5405	4.5292
SpecGate	0.0024	2000	3.0	4.5295†	4.5313	4.5391	–

Table 15: **Feasibility, not repair transfer.** Four healthy references and four assigned stress prefixes per row. "Pass" refers to healthy admission; "eligible" refers to the joint collapse label. Original 350M prefixes were unavailable because their reference dependency failed. The amended row is a separate healthy-only calibration study.

Condition	Healthy rule	Pass	Eligible	Unavailable
350M / WT / 256	Absolute	0/4	0/4	4
124M / WT / 1024	Absolute	4/4	4/4	0
124M / OWT / 256	Absolute	4/4	4/4	0
350M / WT / 256	Healthy-relative	4/4	4/4	0

prefixes qualify. This validates the revised rule only in this recipe, without establishing a model-size or LR scaling law.

No intervention was applied during this feasibility study. The repair study below branches from the eligible prefixes.

H.4 Repair from the transfer prefixes

After all results of the 124M recipe were known and before any repair from a transfer prefix had been run, we branched every eligible prefix of the three configurations above[†] into three arms: the untreated continuation, SpecGate with $\lambda = 3$, $s_0 = 50$ and $w = 10$ as in the 124M recipe, and the entropy penalty at constant weight $\lambda = 3$. No setting was selected anew. Each arm starts from the saved state of the prefix with its weights, optimizer moments, schedule position and data stream, and runs to update $5{,}000$. The protocol was committed before the first run (commit `eaf34971`; runs started on 24 September 2026 at 21:34 UTC). Each configuration is evaluated on its own four prefixes, with 95% paired t intervals, and its target is the final loss of its own healthy reference curve plus 0.10 nats, reached at two consecutive validations. With four pairs per configuration we make no equivalence statements. All 36 runs finished with finite loss.

Table 16: **Repair from transfer prefixes.** Mean unregularized validation loss at update $5{,}000$ over four prefixes per configuration, with the number of runs reaching the target. Healthy: final loss of the healthy reference curve. Δ: paired mean difference of SpecGate minus the untreated continuation (U) or minus the entropy penalty at constant weight (C), with 95% interval. *Prefixes labeled with the healthy rule revised after the original rule failed.

Configuration	Healthy	Untreated	SpecGate	Constant	Δ U	Δ C
124M, context 1024	4.957	5.424 (0/4)	5.248 (0/4)	5.370 (0/4)	-0.177 $[-0.258, -0.095]$	-0.123 $[-0.154, -0.091]$
124M, OpenWebText	5.140	5.601 (0/4)	5.251 (1/4)	5.351 (0/4)	-0.350 $[-0.462, -0.237]$	-0.099 $[-0.107, -0.091]$
350M*	4.595	6.502 (0/4)	4.702 (1/4)	4.825 (0/4)	-1.800 $[-2.012, -1.588]$	-0.123 $[-0.145, -0.102]$

In all three configurations SpecGate ends below the untreated continuation and below the entropy penalty at constant weight, and each of the six 95% intervals, taken separately, lies beyond the ± 0.02-nat margin (Table 16); no simultaneous coverage is claimed. The mean distance to the healthy reference is 0.47 nats untreated and 0.29 with SpecGate at context 1024, 0.46 and 0.11 on OpenWebText, and 1.91 and 0.11 at 350M; the target is reached in 0, 1 and 1 of four runs. The constant weight ends 0.10 to 0.12 nats above SpecGate in each configuration. At context 1024 the lowest head entropy ends at 0.33 to 0.45 with SpecGate and at 0.74 to 0.82 with the constant weight, against 0.007 to 0.009 untreated, while the distance to the healthy reference stays at 0.26 to 0.31 and 0.41 to 0.42 nats.

The saved states load bitwise identical in every configuration. The original prefix runs ran on NVIDIA A40 GPUs and the re-runs in bf16 on A100-SXM4-80GB GPUs. The re-run untreated continuations end between 0.015 nats below and 0.020 nats above the original prefix runs at 124M, and 0.05 to 0.19 nats above them at 350M. For the 350M prefix with the largest difference, the logged

training losses of the original run and the re-run already differ at the first record after the fork, 50 updates in (6.276 against 6.257); the untreated re-run raises its validation loss from 6.25 at update 1,500 to 6.67 at update 5,000. To locate the difference, we continued this saved state three more times for 500 untreated updates with the code of this study, under a protocol committed before the runs (commit `589d8b2e`). Two separate processes on one A100-SXM4-40GB produced bitwise identical model, optimizer and random-number states, and all 22 recorded training, monitoring and validation entries matched the re-run on A100-SXM4-80GB. On an A40, the same code reproduced all 22 entries of the original run. At update 1,750 the validation loss is 6.545 on the A100 GPUs and 6.390 on the A40. The trajectory from this state therefore depends on the execution environment; the test does not separate the GPU from its driver and other execution details, and it covers one state over 500 updates. All arms of this study ran on the same GPU type and branch from the same saved states, so the comparison is paired; the difference to the original runs, up to 0.19 nats, is small against the paired difference of 1.80 nats between SPECGATE and the untreated continuation.

I CONTEXT-LENGTH TRANSFER OF THE SPECTRAL THRESHOLD

We ran matched hinges on cheaper statistics, max row range, RMS (Frobenius) and max |centered logit|, each with its threshold set by the same $1.67\times$-healthy rule. All tie with the spectral hinge at context 256, as our transfer analysis predicted, since at short context every centered-logit hinge prevents and cures collapse. We registered predictions for context-length transfer before running it. The normalized Frobenius statistic dilutes a rank-one collapse spike by $1/\sqrt{T}$, and max-statistics' healthy levels depend on the sampling volume itself (-25–35% when batch shrank $16 \to 4$). At context 1024 with every threshold frozen the spectral hinge holds $s_{\max}$ at 56–62 (three seeds) while Frobenius overshoots to 106–114 (three seeds) and range and max-logit to 111–151 (two seeds each, registered bands), all finishing at or behind the spectral member (Figure 4). We also measured gradient alignment in a synthetic model. In a synthetic ensemble with a planted rank-one spike, our operational definition of the collapse direction, the spectral gradient has cosine 0.92 with the planted direction at the firing threshold, against 0.21 (Frobenius) and 0.02 (range/max), a model-based measurement rather than a training-run one. Thus the spectral threshold transfers better for controlling s in this test; final-loss differences are smaller, and broader calibration robustness remains untested.

J PROTECTION AT HIGH LEARNING RATES

J.1 DESIGN AND EVIDENCE STATUS

This separate study trains the 30M model on WikiText-103 from initialization: context 256, batch 16, no accumulation, 5,000 updates, zero warmup and treatment from update one. It tests protection, not post-collapse repair. At each cosine rate, two development seeds (6100, 6101) select a SPECBREAK coefficient multiplier from $0.3, 1, 3, 10$ and a QK-Clip threshold from $10, 30, 100, 300$. Selection minimizes mean final task loss among candidates finite on both seeds, with fixed tie rules and no grid expansion. Eight evaluation seeds (6110–6117) are paired across methods in initialization and data streams; selections are frozen before evaluation. Constant-rate arms transfer the corresponding cosine-selected settings unchanged. QK-Clip observes the full effective batch; regularizer statistics sample $8/16$ sequences. Neither search guarantees the optimal setting.

The original eight schedule/rate conditions were extended with cosine and constant LR 0.1 before any evaluation outcomes were inspected. The combined ten-condition analysis replaces the historical eight- and two-comparison adjustments with one Holm family of ten. The extension is preserved as an amendment, not silently included in the original protocol. There are 240 evaluation assignments and 128 development/reference assignments across the original and extreme panels. The same eight seed streams across conditions are not 80 independent replications.

Separate endpoints. Finite completion, absence of the joint collapse event and attainment of a task target answer different questions. The joint event requires two consecutive 250-update validation readings, starting at update 1,000 or later, with minimum-head entropy below 0.05 and task loss at least 0.25 above the same-schedule healthy reference curve. References average two separate untreated development runs at LR 0.0006. The target is their final mean plus 0.10, requiring two readings; nonattainment remains censored. This event-through-training rule differs from the historical final-state label. No joint event can therefore mean that entropy is controlled while task learning still fails. All losses below exclude regularization. All assigned endpoints remain in the denominators; no survivor-only mean is used.

Table 17: **All high-LR protection endpoints.** Eight seeds per row. "No joint" counts runs without the event defined above; "Target" counts two-reading task-target attainment. Seconds are mean child elapsed time on A40, including child setup and evaluation; equal updates do not imply equal time. All 240 evaluation assignments are finite.

Schedule	LR	Method	Finite	No joint	Target	Loss	Seconds
const	0.006	Untreated	8/8	0/8	0/8	7.0415	305.3
const	0.006	SPECBREAK	8/8	8/8	0/8	5.1326	600.3
const	0.006	QK-Clip	8/8	8/8	0/8	5.2282	343.9
const	0.02	Untreated	8/8	0/8	0/8	7.4129	303.0
const	0.02	SPECBREAK	8/8	7/8	0/8	5.5314	622.4
const	0.02	QK-Clip	8/8	8/8	0/8	5.6644	343.7
cosine	0.0006	Untreated	8/8	8/8	8/8	4.7795	298.9
cosine	0.0006	SPECBREAK	8/8	8/8	8/8	4.7866	608.4
cosine	0.0006	QK-Clip	8/8	8/8	8/8	4.7795	333.5
cosine	0.0012	Untreated	8/8	1/8	0/8	4.9635	300.7
cosine	0.0012	SPECBREAK	8/8	8/8	1/8	4.8970	605.9
cosine	0.0012	QK-Clip	8/8	8/8	0/8	4.9265	335.8
cosine	0.003	Untreated	8/8	0/8	0/8	6.0276	302.4
cosine	0.003	SPECBREAK	8/8	8/8	5/8	4.8605	601.0
cosine	0.003	QK-Clip	8/8	8/8	0/8	4.9072	339.5
cosine	0.006	Untreated	8/8	0/8	0/8	6.5086	304.5
cosine	0.006	SPECBREAK	8/8	8/8	5/8	4.8609	608.6
cosine	0.006	QK-Clip	8/8	8/8	1/8	4.9011	340.3
cosine	0.01	Untreated	8/8	0/8	0/8	6.6611	306.2
cosine	0.01	SPECBREAK	8/8	8/8	7/8	4.8398	616.4
cosine	0.01	QK-Clip	8/8	8/8	2/8	4.8938	342.9
cosine	0.02	Untreated	8/8	0/8	0/8	6.7922	303.3
cosine	0.02	SPECBREAK	8/8	8/8	3/8	4.9076	627.2
cosine	0.02	QK-Clip	8/8	8/8	0/8	4.9622	340.4
const	0.1	Untreated	8/8	0/8	0/8	15.5977	294.2
const	0.1	SPECBREAK	8/8	8/8	0/8	7.7304	564.1
const	0.1	QK-Clip	8/8	7/8	0/8	6.6039	344.8
cosine	0.1	Untreated	8/8	0/8	0/8	7.3078	287.9
cosine	0.1	SPECBREAK	8/8	8/8	0/8	7.3279	556.0
cosine	0.1	QK-Clip	8/8	6/8	0/8	5.4619	343.5

Table 18: **High-LR paired task-loss contrasts.** QK-Clip minus SPECBREAK; positive favors SPECBREAK. Intervals are pointwise paired t intervals, not simultaneous intervals. Practical-superiority p values test a 0.02-nat margin and use Holm across all ten contrasts.

Schedule	LR	Contrast	Difference	Pointwise 95% CI	Holm p
const	0.006	QK–SpecBreak	+0.0956	$[+0.0595, +0.1317]$	0.01317
const	0.02	QK–SpecBreak	+0.1330	$[+0.0683, +0.1977]$	0.02974
cosine	0.0006	QK–SpecBreak	-0.0072	$[-0.0141, -0.0002]$	1
cosine	0.0012	QK–SpecBreak	+0.0294	$[+0.0078, +0.0510]$	0.8965
cosine	0.003	QK–SpecBreak	+0.0467	$[+0.0315, +0.0618]$	0.02974
cosine	0.006	QK–SpecBreak	+0.0402	$[+0.0142, +0.0662]$	0.5431
cosine	0.01	QK–SpecBreak	+0.0540	$[+0.0068, +0.1011]$	0.5431
cosine	0.02	QK–SpecBreak	+0.0546	$[-0.0183, +0.1275]$	0.8965
const	0.1	QK–SpecBreak	-1.1265	$[-1.2216, -1.0313]$	2.159e-07
cosine	0.1	QK–SpecBreak	-1.8660	$[-2.1593, -1.5727]$	1.334e-05

SPECBREAK meets the adjusted practical-superiority threshold in three conditions (constant $0.006, 0.02$ and cosine 0.003), QK-Clip in both 0.1 conditions. Four conditions remain unresolved, while the stable cosine 0.0006 interval lies inside the practical margin (a pointwise equivalence description). At cosine 0.1, SPECBREAK avoids the joint label in $8/8$ runs yet ends at 7.328 nats, versus 7.308 untreated and 5.462 for QK-Clip. No arm reaches the healthy task target at 0.1, or at any of the tested constant rates. This is a boundary of useful protection, not evidence that controlled entropy guarantees learning.

Table 19: **Development-selected settings.** A setting is a loss coefficient multiplier or a QK-Clip threshold. Constant-rate evaluations reuse these cosine selections. Full candidate endpoints and boundary selections are retained in the portable JSON reports.

Cosine LR	Method	Selected setting
0.0006	SPECBREAK	0.3
0.0006	QK-Clip	100
0.0012	SPECBREAK	3
0.0012	QK-Clip	30
0.003	SPECBREAK	10
0.003	QK-Clip	30
0.006	SPECBREAK	3
0.006	QK-Clip	30
0.01	SPECBREAK	0.3
0.01	QK-Clip	30
0.02	SPECBREAK	3
0.02	QK-Clip	30
0.1	SPECBREAK	0.3
0.1	QK-Clip	10
0.003	Uniform-CE	10
0.003	SpecGate	3
0.006	Uniform-CE	10
0.006	SpecGate	3
0.02	Uniform-CE	3
0.02	SpecGate	10
0.1	Uniform-CE	3
0.1	SpecGate	10

J.2 LATE-ADDED COMPONENT COMPARISONS

Four conditions were selected after inspecting the SPECBREAK/QK-Clip results: constant $0.006, 0.02$, cosine 0.003 and cosine 0.1. Gated uniform-CE and SPECGATE each receive the same four-coefficient search at the four corresponding cosine rates on two development seeds (64 new runs), then the same eight previously inspected evaluation streams in these four conditions (64 new runs). The 96 earlier comparator endpoints are reused, not counted as additional replications. All assignments are terminal and finite. This extension is exploratory: Holm correction of its 16 contrasts does not account for selecting conditions from their observed outcomes or make the reused seeds fresh confirmation.

Both SPECGATE and gated uniform-CE retain the detached spectral gate. SPECGATE removes the differentiable hinge and retains exactly the same floored entropy arithmetic as SPECBREAK; it is a zero-hinge ablation at equal coefficient, with coefficients selected separately for this comparison. It is not an ungated or spectrally independent baseline. No matched-gradient control isolates the hinge's direction from its effective strength, and these results do not establish superiority over other Shannon implementations.

Table 20: **All exploratory simple-penalty endpoints.** Same endpoint definitions as the preceding protection panel. Earlier untreated, SPECBREAK and QK-Clip rows are reused; times come from temporally separate allocations and do not establish controlled speed ratios or equal-time quality.

Schedule	LR	Method	Finite	No joint	Target	Loss	Seconds
const	0.006	Uniform-CE	8/8	8/8	0/8	5.3486	414.4
const	0.006	SpecGate	8/8	8/8	0/8	5.1347	411.9
const	0.006	Untreated	8/8	0/8	0/8	7.0415	305.3
const	0.006	SPECBREAK	8/8	8/8	0/8	5.1326	600.3
const	0.006	QK-Clip	8/8	8/8	0/8	5.2282	343.9
const	0.02	Uniform-CE	8/8	3/8	0/8	5.6811	411.8
const	0.02	SpecGate	8/8	5/8	0/8	5.8562	415.3
const	0.02	Untreated	8/8	0/8	0/8	7.4129	303.0
const	0.02	SPECBREAK	8/8	7/8	0/8	5.5314	622.4
const	0.02	QK-Clip	8/8	8/8	0/8	5.6644	343.7
cosine	0.003	Uniform-CE	8/8	8/8	2/8	4.9145	419.9
cosine	0.003	SpecGate	8/8	8/8	7/8	4.8415	412.6
cosine	0.003	Untreated	8/8	0/8	0/8	6.0276	302.4
cosine	0.003	SPECBREAK	8/8	8/8	5/8	4.8605	601.0
cosine	0.003	QK-Clip	8/8	8/8	0/8	4.9072	339.5
cosine	0.1	Uniform-CE	8/8	8/8	0/8	7.3481	416.6
cosine	0.1	SpecGate	8/8	0/8	0/8	7.4852	416.3
cosine	0.1	Untreated	8/8	0/8	0/8	7.3078	287.9
cosine	0.1	SPECBREAK	8/8	8/8	0/8	7.3279	556.0
cosine	0.1	QK-Clip	8/8	6/8	0/8	5.4619	343.5

The additional hinge has no established task-loss advantage over SPECGATE in these four conditions. At constant 0.006, SPECGATE minus SPECBREAK is 0.0021 nats, but its interval $[-0.0188, 0.0231]$ does not establish practical equivalence. At cosine 0.003, SPECGATE has a lower mean (4.842 versus 4.861) and more target attainments ($7/8$ versus $5/8$), without an adjusted superiority claim between them. SPECBREAK has lower loss than uniform-CE at constant 0.006 beyond the margin after nominal Holm adjustment; the corresponding constant 0.02 contrast does not pass that adjustment. SPECGATE has lower loss than QK-Clip at constant 0.006 and cosine 0.003 under the same exploratory analysis. At cosine 0.1, QK-Clip has substantially lower loss than both simpler penalties, while uniform-CE, like SPECBREAK, avoids the joint event in all eight runs with poor loss. None of these selected cells establishes a universal method ordering.

Table 21: **All exploratory paired contrasts.** First method minus second method; lower loss is better. Pointwise paired 95% intervals; Holm across all 16 practical-margin tests. These p values remain post-selection descriptions. A favorable interval alone is not an adjusted superiority result.

Schedule	LR	Contrast	Difference	Pointwise 95% CI	Holm p
const	0.006	Uniform-CE–SPECBREAK	+0.2160	$[+0.1289, +0.3031]$	0.01316
const	0.006	Uniform-CE–QK-Clip	+0.1203	$[+0.0347, +0.2060]$	0.2769
const	0.006	SpecGate–SPECBREAK	+0.0021	$[-0.0188, +0.0231]$	1
const	0.006	SpecGate–QK-Clip	-0.0935	$[-0.1156, -0.0714]$	0.001427
const	0.02	Uniform-CE–SPECBREAK	+0.1497	$[+0.0553, +0.2441]$	0.1549
const	0.02	Uniform-CE–QK-Clip	+0.0167	$[-0.1170, +0.1505]$	1
const	0.02	SpecGate–SPECBREAK	+0.3248	$[-0.1123, +0.7618]$	1
const	0.02	SpecGate–QK-Clip	+0.1918	$[-0.2391, +0.6227]$	1
cosine	0.003	Uniform-CE–SPECBREAK	+0.0540	$[-0.0001, +0.1082]$	1
cosine	0.003	Uniform-CE–QK-Clip	+0.0074	$[-0.0399, +0.0546]$	1
cosine	0.003	SpecGate–SPECBREAK	-0.0190	$[-0.0391, +0.0012]$	1
cosine	0.003	SpecGate–QK-Clip	-0.0657	$[-0.0818, -0.0495]$	0.003663
cosine	0.1	Uniform-CE–SPECBREAK	+0.0202	$[-0.3937, +0.4341]$	1
cosine	0.1	Uniform-CE–QK-Clip	+1.8862	$[+1.6274, +2.1450]$	8.789e-06
cosine	0.1	SpecGate–SPECBREAK	+0.1573	$[-0.1414, +0.4559]$	1
cosine	0.1	SpecGate–QK-Clip	+2.0233	$[+1.9862, +2.0604]$	7.624e-12

Reproduction and uncertainty. The portable reports retain every seed, all paired differences, 256-sign-flip diagnostics under sign symmetry, leave-one-seed-out intervals, censored targets, development candidates and source hashes. The deterministic renderer checks means, denominators, paired intervals and Holm families before emitting the manuscript tables. With eight seeds, binary endpoint counts remain imprecise; they are descriptive and do not establish differences in success probability. In particular, $8/8$ is not a guarantee of protection. The fixed-update experiment provides no equal-time efficiency ranking. Full SPECBREAK takes about 556–622 seconds in the four selected conditions, versus about 412–420 for the simpler penalties; allocation timing and overhead remain part of the practical question rather than grounds for extrapolated speedups.

K CALIBRATION OF THE GATE

Threshold. It was hand-set once from the calibration rule ($1.67\times$ the healthy level $\Rightarrow$ s_0=50, w=10) and then *fixed in subsequent from-scratch experiments except the sensitivity checks*; among the tested settings, s_0 from 25 to 100 changes final loss by at most 0.05 nats at the extreme rate; w from 3 to 30 has little effect, and a $0.3\times$ gated-entropy weight lowers finals by 0.02. These checks do not make s_0=50 universal; they make it auditable.

L DIRECT SHANNON GATE ABLATION AT 124M

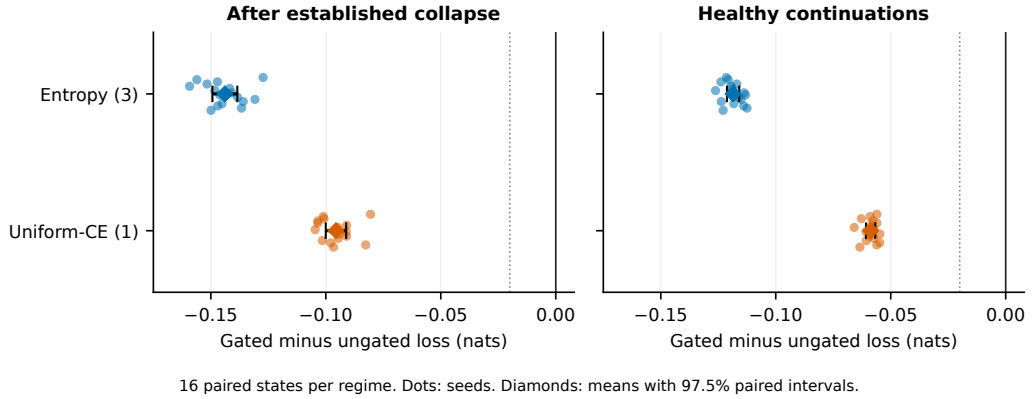

Figure 8: **Gated minus ungated penalty on the panel set.** Sixteen states per regime; dots show per-seed differences in final loss at equal coefficient (entropy 3, uniform cross-entropy 1), diamonds and bars the paired means with 97.5% intervals. Negative values favor the gated penalty; the dotted line marks the prespecified margin of -0.02 nats. Seed-level values for the entropy penalty follow below.

Question and intervention. The ablation replaces the detached spectral gate in Eq. (2) by one. Both methods use the historical Shannon implementation: the same probability floor 10^{-8}, causal mask, omission of singleton rows, normalization by $\log m$, and per-head averaging. Both compute the spectrum and advance its RNG stream, even when U does not use its value. Thus the fixed-coefficient contrast isolates the complete gating rule, including the change in total regularization strength. It does not compare an optimized ungated runtime implementation.

Recipe and execution. The model has 12 layers, 12 heads, width 768, context 256 and 123,849,984 parameters. The study retains the WikiText-103 data identity, batch 16, accumulation 1, statistics subbatch 4, five power iterations, BF16, AdamW and gradient clipping at 1. Both gates use $s_0 = 50$, $w = 10$; the fixed entropy coefficient is 3. The 5,000-update cosine horizon and 500-update warmup retain their original global positions at the fork. Validation uses 40 batches every 250 updates and monitoring every 50. Task losses exclude regularization.

All reported branches run on NVIDIA A100-SXM4-80GB GPUs with Torch 2.10.0+cu128, CUDA 12.8 and Python 3.13.11. Earlier A40 scheduling and handover records are retained as amendments; they do not replace the executed A100 configuration. Both methods were rerun on the same hardware class. Preflight records check value/gradient identities, unchanged historical G behavior, and save/resume behavior. The fixed diagnostic probe restores production RNG states. Parent hashes and full optimizer, schedule and data/RNG state are preserved.

Historical evaluation and tuning. E1-fixed uses the eight original parents (seeds 10–17), with G and U at coefficient 3. These known states support an exploratory comparison, not a new independent confirmation. E1-tuned uses the same parents after separate selection on seeds $3, 4$, with candidates $\{0.3, 1, 3, 10\}$ for each method. The smallest mean final loss among candidates finite on both seeds selects $G = 3$ and $U = 1$, neither at a grid boundary. Selection was frozen before

the tuned evaluation. The eight $G = 3$ logs are reused across E1 panels and counted once. The respective mean endpoints are $G = 4.71641$, $U = 4.86158$ (fixed) and $U = 4.84069$ (tuned). G reaches the task target in $6/8$ cases; neither ungated setting reaches it in any of the eight cases.

New-seed assignment and eligibility. E2 assigns all 16 seeds 83300–83315 to both regimes. The registry reports no overlap with the prior retained run records. The E2 rule and seed list were frozen before any E2 prefix executed, as a local timestamped record rather than a public preregistration. High-rate prefixes use LR 0.0012 and select the first full checkpoint satisfying the existing two-validation collapse criterion by update $2{,}500$. All 16 qualify: 11 at update $1{,}500$, four at $1{,}750$, and one at $2{,}000$. Prefix execution continuing to $2{,}500$ does not change the selected earlier parent. No seed is replaced. Healthy-rate prefixes use LR 0.0003 and are taken at update $1{,}750$ regardless of their diagnostics. Each parent supplies untreated, $U = 3$ and $G = 3$ branches. All 96 branches finish at update $5{,}000$ with finite losses.

There are 32 fresh prefixes, 16 development continuations, 16 E1-fixed continuations, eight additional E1-tuned continuations and 96 E2 continuations: 168 unique runs in this completed execution. The endpoint export has 144 rows for 136 unique continuations because it includes the eight reused G rows. These numbers do not include preflight or superseded execution attempts. All receipts and amendments remain available.

Analysis. The primary outcome is unregularized validation NLL at global update $5{,}000$, paired by common parent. E1 uses descriptive 95% paired t intervals. E2 prespecifies two primary $G - U$ contrasts and uses 97.5% intervals separately, giving at least 95% joint coverage by Bonferroni. The practical margin is ± 0.02 nats; an interval entirely below -0.02 indicates a practical G advantage. This holds in all four reported panels; G has lower loss in every paired seed. Panels and regimes are not pooled. Untreated comparisons below are supporting analyses with ordinary 95% intervals, not extra members of the two-contrast confirmatory family.

On fresh collapsed states, $G - U = -0.14391$ with 97.5% interval $[-0.14932, -0.13850]$. Both penalties improve relative to untreated continuation: G ends at 4.70718, U at 4.85109 and untreated at 5.61354. The supporting G minus untreated difference is -0.90636 (95% interval $[-0.98227, -0.83044]$). G reaches the frozen task target in $13/16$ runs, versus $0/16$ for U and untreated.

On fresh healthy-rate states, $G - U = -0.11856$ with 97.5% interval $[-0.12123, -0.11589]$. G ends at 4.63967, U at 4.75823 and untreated at 4.63579. Thus G is not a task-loss improvement over healthy untreated training: its supporting difference is $+0.00388$ (95% interval $[0.00315, 0.00461]$), while U costs 0.12244 nats. Both untreated and G reach the task target in $16/16$ runs; U in none. These are same-rate comparisons, separate from the historical low-rate reference used to quantify incomplete repair.

Target and missing outcomes. The frozen target is 4.740661001205444 nats, the original development reference final plus 0.10, attained at two consecutive validations. The tables show the confirming update; an unreached target remains right-censored. Numerical failures are outcomes and are not retried; infrastructure retries require an identical configuration and preserved attempt records. The completed panels have no missing or nonfinite pairs, and all fresh prefixes produce their assigned parents.

Mechanistic scope. The archived initial fixed-validation probes have mean gate values 0.933 in fresh collapsed states and 0.105 in healthy-rate states. Final training-batch probes in the G branches are near 0.026 in both regimes. These use different probe batches and are descriptive, not a controlled longitudinal measurement. The weighting diminishes as the observed spectra change, consistent with less persistent regularization; the mean-gate follow-up in App. N tests a scalar spectral control. It does not fix a common coefficient trajectory or compare a nonspectral schedule, so causal attribution remains limited. The study does not alter the asymptotic Shannon-gradient result, establish an added-hinge benefit, rank G against QK-Clip on the new seeds, or test new model sizes or corpora.

Reproduction and audit. The portable evidence export is `paper/evidence_v33`: `endpoints.csv`, `results.json`, the selection and E2 freeze records, eligibility, source

snapshot and compressed raw logs. `scripts/audit_gate_paper_v33.py` independently reconstructs all four intervals, coefficient selection, target attainment and fresh prefix eligibility from those logs. It checks logged parent and data identities, executed source hashes and the E2 freeze timestamp. This audit does not reload or freshly hash the multi-GB checkpoint payloads; their execution-time identities remain in the run records.

Table 22: **Fresh collapse-regime gate ablation.** All assigned seeds qualified; loss at global update 5,000. No untreated or ungated branch reached the two-reading task target. A dash means G also did not reach it.

Seed	Untreated	Ungated	SPECGATE	$G - U$	G target update
83300	5.70477	4.85639	4.70643	-0.14996	4750
83301	5.30236	4.87409	4.73742	-0.13667	5000
83302	5.68755	4.84095	4.69393	-0.14702	4500
83303	5.72132	4.83039	4.68525	-0.14514	4500
83304	5.56169	4.87739	4.74147	-0.13592	–
83305	5.32937	4.82483	4.69402	-0.13081	4500
83306	5.68064	4.85698	4.71859	-0.13839	4750
83307	5.70042	4.86266	4.71656	-0.14609	4750
83308	5.63705	4.84639	4.70546	-0.14094	4500
83309	5.62848	4.82811	4.67981	-0.14830	4500
83310	5.52730	4.88370	4.74188	-0.14182	–
83311	5.52902	4.82325	4.66399	-0.15927	4250
83312	5.67085	4.84946	4.69779	-0.15167	4500
83313	5.75779	4.82779	4.68064	-0.14715	4500
83314	5.71465	4.85185	4.69577	-0.15608	4500
83315	5.66335	4.88321	4.75591	-0.12729	–

Table 23: **Fresh healthy-regime gate ablation.** Common parents at update 1,750; final loss at 5,000. All untreated and G branches reach the target, versus none of the ungated branches. Lower is better.

Seed	Untreated	Ungated	SPECGATE	$G - U$	G target update
83300	4.63374	4.76025	4.63737	-0.12288	4000
83301	4.64531	4.76274	4.65023	-0.11251	4000
83302	4.63545	4.75074	4.63679	-0.11395	4000
83303	4.65133	4.77224	4.65398	-0.11826	4000
83304	4.62799	4.75530	4.63161	-0.12368	4000
83305	4.63755	4.75485	4.63977	-0.11507	4000
83306	4.63677	4.75737	4.64128	-0.11609	4000
83307	4.64227	4.75817	4.64503	-0.11314	4000
83308	4.63364	4.75281	4.63902	-0.11379	4000
83309	4.62143	4.75438	4.62819	-0.12620	4000
83310	4.63808	4.76148	4.64252	-0.11897	4000
83311	4.63459	4.75827	4.63886	-0.11941	4000
83312	4.63101	4.75161	4.63463	-0.11698	4000
83313	4.63196	4.75833	4.63448	-0.12385	4000
83314	4.63650	4.76220	4.64153	-0.12067	4000
83315	4.63500	4.76090	4.63944	-0.12146	4000

M SELECTED DEVELOPMENT RESULTS AT FIXED COST

These tables retain the development summaries discussed in Appendices H.1 and H.2; full grids and cost accounting remain in App. H.

Table 24: **Salvage at a fixed remaining time budget.** Final validation loss at $B = 300$ charged policy seconds on A40; lower is better. Four *development* seeds per row. Bursts last 250 updates with reactivation and a 750-update active cap. Operational alternatives are selected from nine candidates per family. Target: two readings at or below 4.744 nats. Selection and evaluation use development data; no confirmatory ranking.

Policy	Mean loss	Seed range	Target
Untreated continuation	5.863	5.672–5.984	0/4
SpecBreak, continuous	5.113	5.063–5.155	0/4
SpecBreak, bursts	5.036	4.985–5.081	0/4
Uniform-CE, bursts	5.006	4.961–5.053	0/4
SpecGate, bursts	5.013	4.967–5.045	0/4
Selected LR rescue	5.472	5.354–5.565	0/4
Selected rollback	5.115	5.081–5.138	0/4
Selected restart	5.207	5.195–5.221	0/4

Table 25: **Whole-training development selection at equal cost.** Each family receives 36 runs: 12 configurations and three seeds. The lowest mean loss among candidates with all three valid total-cost endpoints is selected. Updates differ because the 1,200-second budget is shared, not the token count. Selected means are development estimates.

Method	Mean loss	Peak LR	Warmup	Updates
Untreated	4.218	0.0003	1000	8,500–8,580
SpecBreak	4.464	0.0006	500	5,721–5,750
Uniform-CE	4.397	0.0006	500	7,009–7,077
SpecGate	4.368	0.0006	500	6,810–6,906
QK-norm	4.278	0.0003	1000	7,349–7,395

N PER-HEAD VERSUS MEAN SPECTRAL GATING

Question and control. The direct $G-U$ ablation establishes a benefit of the complete gating rule. E5 asks whether applying different weights to different heads is needed for a comparable final task loss. Write $d_{\ell,h} = 1 - H_{\ell,h}$ and let $\bar{g} = \mathrm{mean}_{\ell,h}\, g_{\ell,h}$, averaged in FP32 over all layers and heads. At the fixed entropy coefficient 3, the two penalties are

$$R_G = 3 \operatorname*{mean}_{\ell,h}(g_{\ell,h} d_{\ell,h}), \qquad R_M = 3\bar{g} \operatorname*{mean}_{\ell,h} d_{\ell,h}.$$

The original per-head gate values, detached computation, entropy floor, masking, aggregation, spectral estimator and RNG path remain unchanged. The control applies the scalar $\bar{g}$ to every head; it does not round that mean back to BF16. Both methods still measure the spectral statistic.

On the same state and batch the sums of weights match, and in real arithmetic $R_G - R_M = 3\,\mathrm{Cov}_{\ell,h}(g, d)$, with population covariance. This identity does not match loss magnitudes, gradient norms, clipped Adam updates or cumulative treatment. Each arm measures its own current spectrum. Once their trajectories diverge, their scalar gate schedules can differ as well. E5 therefore compares two adaptive policies, not head allocation under an externally fixed common schedule.

Design and provenance. E5 reuses every E2 parent in App. L: 16 collapsed and 16 healthy-rate states, seeds 83300–83315. The control was designed after inspecting the E2 G/U outcomes. Its protocol was fixed before the new mean-gate branches; it is a subsequent control on known states, not fresh independent replication. No coefficient selection, eligibility filter or replacement seed is used. The model, dataset, batches, AdamW/BF16, thresholds, cadence and update-5,000 endpoint match E2.

Two full per-head replays on A100-SXM4-80GB, seed 83300 in each regime, agree exactly with the historical scientific records under the frozen administrative-field exclusions (146 healthy and 157 collapse records). The fallback requiring all G arms to be rerun was not used. The primary comparison therefore pairs 32 new M branches with the original G branches, with matched parent,

data and core-source identities. A separate 200-update profile uses a historical development parent. All 35 processes complete without numerical failures or retries. The completed allocation takes 37.6 minutes and 5.017 allocated A100 GPU-hours including preflight, profiling and idle reserved slots; this is not a comparative method-throughput measurement.

Results and analysis. Each regime uses a 97.5% paired t interval for $G - M$, with the predeclared practical margin ± 0.02 nats. Both intervals exclude zero in favor of G, but lie entirely within that margin (Table 26). Thus the small mean advantage and the practical-equivalence classification coexist. The conclusion concerns the mean final task loss, not every individual seed or attention metric. Every leave-one-seed-out interval retains the same equivalence classification. Exact two-sided sign-flip tests of a zero-mean null give $p = 0.0004883$ after collapse and $p = 0.00003052$ in healthy continuations; these are sensitivity checks under sign-exchangeability and do not test equivalence. No primary pair is missing or nonfinite.

Table 26: **Mean spectral gating retains most of the observed endpoint benefit.** Task NLL at global update $5{,}000$; lower is better. G uses per-head gates and M their global mean. Each regime has 16 finite pairs. Target counts use the original two-reading threshold 4.740661. Intervals are adjusted for the two contrasts; E5 remains a subsequent control on already inspected E2 parents.

Regime	G	M	$G - M$	97.5% interval	Target G; M
Collapse	4.70718	4.71371	-0.00652	$[-0.01061, -0.00244]$	13/16; 12/16
Healthy	4.63967	4.64263	-0.00296	$[-0.00383, -0.00209]$	16/16; 16/16

For context, ungated Shannon ends at 4.85109 after collapse and 4.75823 in the healthy regime. Mean gating retains most of the observed improvement over these endpoints. Both adaptive penalties still sit slightly above healthy untreated training (4.63579): $+0.00388$ for G and $+0.00684$ for M. After collapse, G and M attain the task target in $13/16$ and $12/16$ cases, respectively, versus $0/16$ for ungated and untreated continuations. Healthy-regime target counts are $16/16$ for G, M and untreated, versus $0/16$ ungated. These secondary comparisons reuse E2 context; no new family-wide superiority claim is based on them.

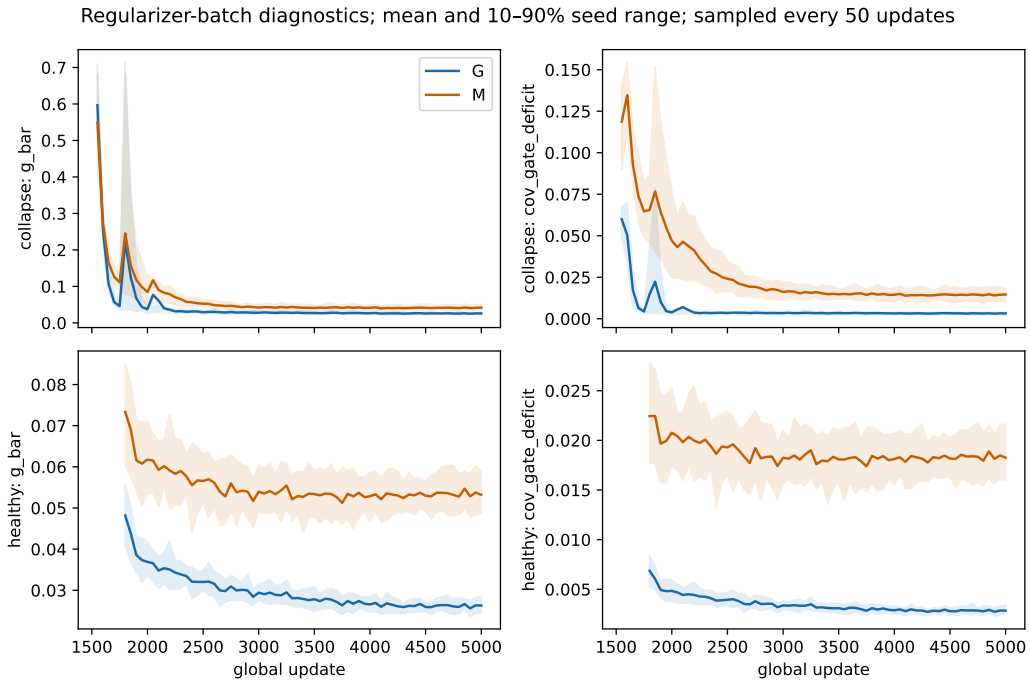

Figure 9: **Gate trajectories and final loss of per-head and mean gating.** Regularizer-batch mean spectral gate and population gate–deficit covariance for per-head G and mean-gate M. Curves show means and 10–90% seed ranges at observed global updates. Both are shown at a 50-update cadence: historical G has no dense intervening trace; M's every-update diagnostics are sampled to that cadence. Early collapse-regime updates contain only branches already activated, because parents occur at different updates. The scalar mean and covariance are computed within each method's own evolving state. These are descriptive trajectories, not an optimizer-space mediation analysis.

What E5 establishes. The spectral adaptive scalar is sufficient to retain most of the observed final-loss benefit in these recipes. Per-head weighting adds a small mean advantage below the chosen practical relevance threshold. This does not establish sufficiency of an arbitrary or nonspectral schedule, interchangeability on other endpoints, or equality of whole-trajectory dose. Figure 9 explicitly shows that the two adaptive trajectories differ. The covariance identity is a loss decomposition, not evidence that covariance mediates task recovery.

Artifacts and independent check. `paper/evidence_e5` contains the frozen protocol, endpoint and trace exports, new raw logs, replay receipts and the executed adapter. Historical G/U/untreated logs remain in `paper/evidence_v33`. `scripts/audit_gate_control_e5.py` independently reconstructs the two primary intervals, all leave-one-out classifications, sign-flip checks and target counts from the raw logs. It also checks paired configuration identities, uniform applied mean-gate weights and the covariance identity on the dense M diagnostics. Checkpoint payloads were not reloaded during this paper audit.

O Panel set and added arms

States. The panel set (called E2 in the released artifacts) holds sixteen states per regime, seeds 83300 to 83315, assigned together with the protocol before any run. The collapsed and the healthy regime use different states despite the same seed identifiers; within a regime all arms branch from the same parent. The direct gate comparison of Section 5.2 and its seed-level values are in Appendix L.

Added arms. Three follow-up studies were added after the first results: treatment started at the alarm (D01; eight paired seeds 10 to 17 of the confirmation set), query/key clipping (D02) and uniform cross-entropy with and without the gate, including the smaller constant dose (D03; sixteen parents per regime of the panel set). They reuse previously inspected seed streams and saved states and are marked † in the main text; they are not an independent confirmation of the recipe. D01 contains sixteen new runs, D02 sixty-four and D03 ninety-six; the late SPECGATE, untreated and panel-set references are reused. All 176 runs completed at update 5,000 without numerical failures or retries. The D01 hardware amendment used 80-GB A100s, recomputed the live SPECBREAK references, and verified pre-intervention identity and alarm timing.

Execution history. The original design assigned sixteen seeds per regime (commit `fceb7dac`, 23.09.2026, 16:28 UTC). After cost-only profiling the assignment was temporarily reduced to eight (`9bd325a0`, 19:23 UTC), and all original seeds were reinstated while the first half was running (`bdeee7d1`, 20:05 UTC), before comparative outcomes were inspected in this analysis workflow. The remaining seeds ran in a second block without changes to methods, coefficients, thresholds or endpoints. The full sixteen-seed analysis is primary; the intermediate eight-seed report is not an independent confirmation.

Intervals and multiplicity. The committed rules apply: D01 uses paired 95% t intervals; D02 uses 97.5% intervals within each two-contrast family of a regime; D03 uses 97.5% intervals for its two regime comparisons. The direct threshold contrasts of D02 and the dose contrasts of D03 form their own two-contrast families with 97.5% intervals. Secondary contrasts use 95%. A Holm correction over eleven prespecified directional tests against the 0.02-nat margin protects joint superiority statements; it does not replace the interval decisions or the equivalence criterion. The four D02 intervals do not give joint 95% coverage; that would have required 98.75% intervals, which were not frozen in the executed protocol. Exact sign-permutation tests (256 for D01, 65,536 for D02 and D03) and leave-one-seed-out intervals are in the study files; the permutation tests are not equivalence tests.

Table 27: **Contrasts of the added arms.** Unregularized validation loss at update 5,000, paired differences in nats with the committed interval level. Directions as written; SG: SPECGATE; uCE: uniform cross-entropy with coefficient.

Contrast	Regime	Δ	Interval
Alarm start minus late start (SG)	collapse	-0.21976	$[-0.25819, -0.18132]$, 95%
SG minus SPECBREAK, both at alarm	collapse	-0.01266	$[-0.01723, -0.00809]$, 95%
QK-Clip $\tau = 100$ minus SG	collapse	$+0.02943$	$[+0.02369, +0.03517]$, 97.5%
QK-Clip $\tau = 100$ minus SG	healthy	-0.00388	$[-0.00474, -0.00303]$, 97.5%
QK-Clip $\tau = 30$ minus SG	collapse	-0.01697	$[-0.02104, -0.01289]$, 97.5%
QK-Clip $\tau = 30$ minus SG	healthy	-0.00191	$[-0.00247, -0.00134]$, 97.5%
QK-Clip $\tau = 30$ minus $\tau = 100$	collapse	-0.04640	$[-0.05012, -0.04267]$, 97.5%
QK-Clip $\tau = 30$ minus $\tau = 100$	healthy	$+0.00198$	$[+0.00130, +0.00265]$, 97.5%
uCE 1 gated minus ungated	collapse	-0.09558	$[-0.10003, -0.09114]$, 97.5%
uCE 1 gated minus ungated	healthy	-0.05874	$[-0.06075, -0.05673]$, 97.5%
uCE 1 gated minus uCE 0.3 ungated	collapse	-0.03104	$[-0.03395, -0.02813]$, 97.5%
uCE 1 gated minus uCE 0.3 ungated	healthy	-0.01737	$[-0.01920, -0.01554]$, 97.5%
uCE 1 gated minus untreated	healthy	$+0.01506$	$[+0.01345, +0.01667]$, 95%
QK-Clip $\tau = 30$ minus untreated	healthy	$+0.00197$	$[+0.00140, +0.00255]$, 95%

Holm-protected at the margin: alarm against late start, both gated-minus-ungated uCE contrasts, the dose contrast after collapse, and $\tau = 30$ against $\tau = 100$ after collapse. For QK-Clip after collapse the prespecified Holm direction was QK-Clip better than SPECGATE; the positive $\tau = 100$ interval is therefore reported as a measured higher loss, not as a confirmed Holm result for SPECGATE.

P QUERY/KEY CLIPPING WITH A SELECTED THRESHOLD

Our QK-Clip implementation uses a threshold τ on each head's largest valid attention logit over the full effective training batch. After the optimizer step it rescales that head's query and key projections

when the recorded maximum exceeds τ, leaving Adam moments unchanged. The main comparison uses the published threshold $\tau = 100$ (Kimi Team, 2025). The selected threshold was obtained on the collapsed development states for seeds 3 and 4, using final unregularized validation loss at update 5,000. The initial grid was $\{30, 100, 300\}$. Its winner, 30, triggered the prespecified single boundary expansion to 10; 30 remained the winner. Candidates had to finish with finite loss on both seeds. Exact ties were resolved by proximity to 100 in log distance, then by the larger threshold.

The coefficient of SPECGATE was also selected by mean final loss on seeds 3 and 4, from $\{0.3, 1, 3\}$ with the predeclared expansion by one step to 10, and with its own tie rule (Appendix F). The two selections thus share the development seeds, the objective and a single boundary expansion, not the full search procedure. The selected $\tau = 30$ was transferred unchanged to both regimes of the panel set. Both thresholds branch from all sixteen panel-set states per regime used for the SPECGATE comparisons. These QK-Clip arms were added after the SPECGATE outcomes on the panel set had been inspected. They are subsequent controls on the panel set, not an independent confirmation on unseen states.

Table 28: **Query/key clipping against SPECGATE on the panel set.** Sixteen paired states per regime; mean unregularized validation loss at update 5,000. Δ: QK-Clip minus SPECGATE, with prespecified 97.5% paired t intervals for the two contrasts within each regime. These are separate families, not simultaneous 95% coverage across all four intervals. Equivalence margin ± 0.02 nats. Target: two consecutive validation readings at or below 4.740661; counts are descriptive. $\tau = 100$: published threshold; $\tau = 30$: selected on development seeds.

	After collapse			Healthy		
Arm	Loss	Target	Δ	Loss	Target	Δ
SPECGATE, $\lambda = 3$	4.707	13/16		4.640	16/16	
			$+0.0294$			-0.0039
QK-Clip, $\tau = 100$	4.737	10/16	$[+0.0237, +0.0352]$	4.636	16/16	$[-0.0047, -0.0030]$
			-0.0170			-0.0019
QK-Clip, $\tau = 30$	4.690	16/16	$[-0.0210, -0.0129]$	4.638	16/16	$[-0.0025, -0.0013]$

With $\tau = 30$, QK-Clip ends 0.017 nats below SPECGATE after collapse and reaches the target in all sixteen runs, compared with thirteen (Table 28). The prespecified equivalence criterion is not satisfied: the 97.5% interval $[-0.02104, -0.01289]$ extends below -0.02. It also does not establish a benefit exceeding 0.02 nats. In a supplementary analysis at the 90% level, chosen after the results were known, the paired interval is $[-0.01983606, -0.01409690]$ and lies within the margin. This post hoc result does not change the prespecified comparison, which remains unresolved. In healthy continuations, both thresholds are equivalent to SPECGATE at the prespecified level.

The two thresholds give different observed rankings after collapse. Lowering τ from 100 to 30 reduces the final loss of QK-Clip by 0.0464 nats (paired 97.5% interval for $\tau = 30$ minus $\tau = 100$: $[-0.05012, -0.04267]$); this contrast also passes the prespecified Holm correction across studies. At $\tau = 100$, QK-Clip ends above SPECGATE, and at the development-selected $\tau = 30$ it ends below. These mean rankings do not by themselves establish the corresponding directional claims beyond the margin. In healthy continuations the threshold difference is small ($+0.00198$ nats for $\tau = 30$ minus $\tau = 100$; 97.5% interval $[+0.00130, +0.00265]$) and within the margin.

For a fixed measured head maximum, a lower threshold triggers rescaling at a smaller value and rescales more strongly when clipping is active. Both arms start immediately after the same parent; the experiment does not vary their start time independently. The maximum is measured before the optimizer step, so the rule does not guarantee a strict bound on τ in a fresh forward pass after the update.

Confirmation seeds. The same selection rule was applied earlier to the eight confirmation seeds, branching from the same established collapses as Table 2; it also chose $\tau = 30$, and all eight runs reach the target. There QK-Clip ends 0.020 nats below SPECGATE $[-0.029, -0.010]$, and it is equivalent to the penalty with hinge within the margin (difference -0.015 $[-0.019, -0.011]$; 95% paired intervals, the level fixed for that study). At the published threshold it ends at 4.743, against 4.716 for SPECGATE. These QK-Clip arms were also added after the penalty outcomes were known.

Equal remaining time. On six new collapsed states with the same wall-clock budget for every arm, QK-Clip at $\tau = 30$ ends 0.068 nats below SPECGATE $[-0.084, -0.052]$ and reaches the target in five of six runs (Appendix T).

Q ENTROPY AS THE SIGNAL FOR A GATED PENALTY

A penalty can also be gated by the quantity it penalizes. Jha & Reagen (2025) apply their entropy penalty only where the entropy of a head deviates from a learnable threshold by more than a tolerance, in either direction. We ran this regularizer[†] from the sixteen collapsed and sixteen healthy states of the panel set (Appendix O), alongside SPECGATE at $\lambda = 3$ and the untreated continuation.

Implementation and selection. The loss function of the released code (commit `728944d`) is called unchanged and added to the task loss; the rest of the training step is the same as for all other arms. The tolerance factor is 0.2 and the coefficient 10^{-5}, as in the released library. The thresholds, one per layer and head, start at 0.5, which is our choice and not a published setting, and are trained with the model at the same learning rate with fresh optimizer moments. A second coefficient was selected on the collapsed development states of seeds 3 and 4 from $\{0.3, 1, 3, 10\} \times 10^{-5}$ by the lowest mean final loss, without grid expansion; mean final losses were 5.658, 5.462, 5.088 and 4.780, and the selected coefficient 10^{-4} lies at the upper edge of the grid. Arms, endpoints, intervals and decision rules were frozen before the first run (commit `9e8f2b6a`). The SPECGATE results on these states were known at that time.

Table 29: **Entropy-gated regularizer on the panel set.** Mean unregularized validation loss at update 5,000 over sixteen paired states per regime. Target: two consecutive validation readings at or below 4.740661; counts are descriptive. All 64 runs finished with finite loss.

	After collapse		Healthy	
Arm	Loss	Target	Loss	Target
Untreated	5.614	0/16	4.636	16/16
SPECGATE, $\lambda = 3$	4.707	13/16	4.640	16/16
Entropy-gated, coefficient 10^{-5}	5.517	0/16	4.635	16/16
Entropy-gated, coefficient 10^{-4} (selected)	4.769	1/16	4.644	16/16

After collapse, the entropy-gated regularizer ends 0.810 nats above SPECGATE at the library coefficient $[0.731, 0.889]$ and 0.061 nats above it at the selected coefficient $[0.055, 0.067]$, with 1 of 16 runs reaching the target against 13 (Table 29; 95% paired intervals, the level fixed for this study). Both differences exceed the ± 0.02-nat margin and pass the prespecified Holm correction over the four primary contrasts. In healthy continuations both coefficients are equivalent to SPECGATE within the margin (differences -0.004 $[-0.005, -0.004]$ and $+0.005$ $[+0.004, +0.005]$). At the selected coefficient the regularizer also lowers the final loss after collapse from 5.614 untreated to 4.769.

The two regularizers differ in more than their signal: the entropy-gated penalty acts on low and high entropy, trains its thresholds and has a different functional form. The comparison therefore concerns the two complete constructions under the stated adaptation and grid, not the spectral signal in isolation; Appendix R changes only the signal.

Entropy as a signal of collapse. Low entropy in a head is common in runs that train normally. Of the 227 runs in the corpus, 48 end with a head below normalized entropy 0.05 and still reach near-healthy loss (Appendix E). As an alarm, entropy separates collapsed from non-collapsed runs less well than spectral sharpness. The entropy criterion detects 106 of 108 collapsed runs and the spectral alarm 104; the entropy criterion fires in 69 of 119 non-collapsed runs, the spectral alarm in 25 (Appendix E). At context length 1024, a hinge on the spectral sharpness with its threshold unchanged held the sharpness near that threshold, where hinges on cheaper statistics overshot (Appendix I); detection rates at that length were not measured.

R SIGNAL AND FEEDBACK OF THE GATE

Two controls† keep the construction of SPECGATE, the penalty $1 - \bar{H}$, the detached gate, the average over layers and heads, $\lambda = 3$ and the probability floor, and change only what sets the gate. No parameter of either control was selected on the outcomes of this study. The full spectral computation runs in every arm, so the study compares the behaviour of the interventions, not their cost; Appendix T compares the replayed schedule without the spectral computation at equal remaining time.

Entropy-deficit gate (E). The gate of each head reads its entropy deficit $1 - \bar{H}$ in place of the spectral sharpness, $g = \mathrm{sg}[\sigma((d - d_0)/w_d)]$, with d_0 and w_d set by a rule on healthy runs only: the mean deficit over layers and heads of two untreated healthy development runs (seeds 3 and 4) at update 1,750 is 0.404, d_0 is 1.67 times this value, 0.674, and $w_d = 0.2\, d_0 = 0.135$, for all heads and both regimes. This rule mirrors the ratios of the spectral gate but is not a reconstruction of how $s_0 = 50$ was set. It is compared with SPECGATE.

Replayed schedule (R). The gate is a fixed function of the number of updates since the fork, the same for every head and every run, in both regimes. It is built from the two development runs of SPECGATE after collapse (seeds 3 and 4, not part of the panel set): the gate applied to each head is averaged over heads every 50 updates, the two runs are averaged at equal age since their fork, and the value is held between readings; the first 49 updates use the value at age 50, and past the end of the shorter run its last value is held. The schedule falls from 0.72 at the start to below 0.05 within 250 updates and stays near 0.03 afterwards. It carries no information about the run it is applied to, but it is derived from treated development runs, so it is not a control designed without repair data. Because it gives every head the same weight, it is compared with the mean gate of Appendix N, which does the same but follows the current run.

Both controls branch from the sixteen collapsed and sixteen healthy states of the panel set (Appendix O). The protocol was frozen before the first run (commit `b668ed66`, 24 September 2026, 22:27 UTC; runs started at 22:30 UTC); the SPECGATE and mean-gate results on these states were known. The reference arms were taken from the earlier studies after a 250-update replay on the same GPU type reproduced their logged records exactly, and every branch loaded its parent bitwise identical. For each of the four primary contrasts we test three statements against the margin: lower by more than 0.02 nats, higher by more than 0.02 nats, and equivalent within ± 0.02 nats (two one-sided t tests), and apply one Holm correction to all twelve. An equivalence statement also requires the whole 95% paired interval inside the margin. After correction the four equivalence statements have $p \leq 4.2 \times 10^{-7}$ and all eight directional statements $p = 1$. Secondary contrasts are descriptive. All 64 runs finished with finite loss.

Table 30: **Signal and feedback of the gate on the panel set.** Mean unregularized validation loss at update 5,000 over sixteen paired states per regime, with runs reaching the target. Δ: primary contrast with 95% paired interval; all four are equivalent within ± 0.02 nats after Holm correction.

	After collapse		Healthy	
Arm	Loss (target)	Δ	Loss (target)	Δ
SPECGATE (G)	4.707 (13/16)		4.640 (16/16)	
		E−G: $+0.004$		E−G: $+0.002$
Entropy-deficit gate (E)	4.711 (12/16)	$[+0.002, +0.006]$	4.642 (16/16)	$[+0.002, +0.003]$
Mean gate (M)	4.714 (12/16)		4.643 (16/16)	
		R−M: -0.002		R−M: $+0.011$
Replayed schedule (R)	4.712 (12/16)	$[-0.006, +0.002]$	4.653 (16/16)	$[+0.010, +0.012]$

After collapse, both controls end within 0.005 nats of SPECGATE and of the mean gate: the entropy-deficit gate 0.004 above SPECGATE, the replayed schedule 0.002 below the mean gate and 0.005 above SPECGATE $[+0.002, +0.007]$ (Table 30). In healthy continuations the entropy-deficit gate ends 0.002 above SPECGATE and 0.006 above the untreated run, and the replayed schedule 0.011 above the mean gate and 0.018 above the untreated run $[+0.017, +0.018]$, against 0.004 for SPECGATE.

The entropy-gated regularizer of Appendix Q ended 0.061 nats above SPECGATE on the same states, whereas the entropy-deficit gate, which reads entropy inside the construction of SPECGATE, ends 0.004 above it. Reading entropy in place of the spectrum within this construction therefore accounts for a small part of that difference; the experiments do not decompose the remaining differences between the two regularizers.

S TIME COURSE OF REPAIR

We read the logged training statistics and validation losses of the confirmation set for the three penalties at their confirmation settings: SPECGATE at $\lambda = 3$, the penalty with hinge at multiplier 3 and gated uniform cross-entropy at multiplier 1. This analysis was made after all results were known and is descriptive.

The largest spectral sharpness on the statistics sub-batch, logged every 50 updates, lies between 984 and 1,904 in the untreated continuation 250 updates after the fork. It falls to at most 50, the gate center, within 150 to 350 updates with SPECGATE and within 50 to 100 updates with the two other penalties.

The loss advantage over the untreated continuation builds up more slowly and at nearly the same pace for all three penalties (Table 31). About a third of the final advantage is present 250 updates after the fork, and nearly all of it after 2,000 updates. The earlier spectral recovery of the penalty with hinge and of uniform cross-entropy did not come with an earlier loss recovery in these runs.

Table 31: **Share of the final loss advantage over the untreated continuation reached after a given number of updates since the fork.** Validation loss; mean over the eight seeds of the confirmation set, range in brackets.

Penalty	250	500	1,000	2,000
SPECGATE	0.36 [0.29, 0.48]	0.57 [0.50, 0.66]	0.83 [0.73, 0.94]	0.97 [0.96, 0.99]
With hinge	0.38 [0.31, 0.53]	0.58 [0.50, 0.68]	0.83 [0.73, 0.96]	0.97 [0.96, 0.99]
Gated uniform-CE	0.39 [0.32, 0.55]	0.59 [0.51, 0.69]	0.84 [0.74, 0.97]	0.98 [0.96, 0.99]

T EQUAL REMAINING TIME AND A SHORT TREATMENT

Two studies[†] were added after all other 124M results were known. They use six new seeds, 20 to 25, assigned with the protocol before any run; their states had not been inspected. The protocol was frozen before the first training run (commit `0fc83e42`, 25 September 2026, 06:54 UTC). It replaces an earlier freeze whose input record named two data files wrongly; the workers stopped in a preflight check before any training, and the corrected names point to files with the frozen hashes. Each seed trains the 124M recipe at learning rate 1.2×10^{-3} until the first state that meets the collapse rule of the confirmation set; all six qualified, four at update 1,500 and two at 1,750. All arms of a seed branch from this state, and all runs use one A100-SXM4-40GB GPU each. No setting was selected on these seeds.

Equal remaining time. Four arms receive the same wall-clock budget from the saved state: untreated training, SPECGATE ($\lambda = 3$, optimized implementation), the replayed schedule of Appendix R, unchanged, with $\lambda = 3$ and without power iterations or spectral monitoring, and QK-Clip at the threshold $\tau = 30$[‡] selected on development seeds of this task for the final loss (Appendix P). The budget is 302 seconds, the time an untreated development run (seed 3) needed on this GPU type from its collapsed state at update 1,750 to update 5,000, including start-up, loading, validation and the final checkpoint. The clock starts before the training process is launched and charges every step. Updates stop when 25 seconds remain; the model at that point is validated, and this validation must finish within the budget. The endpoint is therefore the validated model within the budget, not a saved checkpoint: all 24 validations finished between 279 and 281 seconds, and saving the evaluated state as a reusable checkpoint finished within the budget in 7 of 24 runs and up to 17 seconds after it in the others. This time is recorded separately and cannot change the endpoint. The learning-rate schedule stays indexed by updates with its horizon at 5,000, as in the recipe, so an arm that completes fewer updates ends at a higher learning rate; no arm reached update 5,000. The comparison

therefore measures repair within the budget under the unchanged schedule, not a difference caused by implementation cost alone. Validation omits the spectral computation in every arm. The primary contrasts are SPECGATE minus replayed schedule and SPECGATE minus QK-Clip, with paired t intervals at 95% over the six seeds. A difference beyond the ± 0.02-nat margin is established when the interval clears the margin and the Holm-adjusted p over the four one-sided statements (two directions for each contrast) is below 0.05. With six seeds no equivalence test was planned. The target is the one of Section 5: two consecutive validations at or below 4.741, both finished within the budget. Validation runs every 250 updates and once more on the final model, which can serve as the second of the two; in three of the five QK-Clip runs that reach the target it does, 153 to 213 updates after the first. All 24 runs ended with a validation inside the budget.

Table 32: **Repair at equal remaining time.** Six new collapsed states; every arm branches from the same state and receives the same 302-second wall-clock budget; updates stop when 25 seconds remain, and the last model is validated within the budget. Updates: completed after the branch, mean and range. Loss: validation loss of the last model completed within the budget. Target: two consecutive validations within the budget at or below 4.741. Δ: arm minus SPECGATE, 95% paired interval. ‡: threshold tuned on development seeds of this task for the final loss, not the published default.

Arm	Updates	Loss	Target	Δ
Untreated	3,223 (3,206–3,247)	5.649	0/6	$+0.870$ $[+0.796, +0.943]$
SPECGATE	2,472 (2,427–2,500)	4.779	0/6	
Replayed schedule	2,676 (2,636–2,696)	4.757	0/6	-0.022 $[-0.027, -0.017]$
QK-Clip, $\tau = 30^{\ddagger}$	2,925 (2,843–3,000)	4.711	5/6	-0.068 $[-0.084, -0.052]$

QK-Clip at the tuned threshold ends 0.068 nats below SPECGATE, lower in all six seeds and beyond the margin (Holm-adjusted $p = 0.001$), and reaches the target in five of six runs (Table 32). The replayed schedule ends 0.022 nats below SPECGATE, lower in all six seeds; a difference beyond the margin is not established (Holm-adjusted $p = 0.65$). The study measures repair from a collapse within a fixed remaining time on one GPU type and one recipe, not the cost of a whole training run.

A short treatment. Three arms train from the same states to update 5,000 without a time budget: untreated, SPECGATE throughout, and SPECGATE for the first 250 updates after the branch with its weight set to zero afterwards; the monitor keeps running. By a change recorded before any run, the short arm recomputes its first 250 updates from the saved state instead of branching from the continuous arm, because the runner does not accept a changed setting on resume; its treatment during these updates is the same. All 66 logged entries of the two arms up to the switch, eleven per seed with training loss, gradient norm, spectral and entropy statistics, batch hashes and the validation at the switch, are identical; no checksum of the full state was written at that point. The primary contrast is short minus continuous treatment, with a 95% paired interval and the same margin. A return of the spectral sharpness is recorded when two consecutive monitor readings of $s_{\max}$ after update 250 of the branch exceed the gate center 50.

Table 33: **Continuous and short treatment.** Six new collapsed states, final loss at update 5,000. Δ: arm minus continuous SPECGATE, 95% paired interval. $s_{\max}$: largest spectral sharpness over layers and heads at the last monitor reading.

Arm	Loss	Target	Δ	$s_{\max}$ at 5,000
Untreated	5.647	0/6	$+0.928$ $[+0.844, +1.013]$	2,008–2,480
SPECGATE, continuous	4.719	4/6		34–39
SPECGATE, first 250 updates	4.821	0/6	$+0.102$ $[+0.061, +0.143]$	172–536

Stopping after 250 updates ends 0.102 nats above continuous treatment, higher in all six seeds and beyond the margin, and 0.826 nats below the untreated continuation $[-0.900, -0.753]$ (Table 33, Figure 1). When the weight was set to zero, $s_{\max}$ was between 38.5 and 61.2, below the gate center in four of six runs and at or above it in two. At the next reading, 50 updates later, it was between 106 and 183 in all six runs and stayed above the gate center through update 5,000. The study shows that stopping this treatment after 250 updates raised the final loss and let the spectral sharpness

rise; it does not show whether a repair that had completed would need a continued weight. Under continuous treatment $s_{\max}$ stayed near the gate center; in one run it exceeded it at two consecutive readings, with a maximum of 54. Running to update 5,000, a continuous SPECGATE run took 372 to 408 seconds of process time against 288 to 313 seconds untreated.

U NUMERICAL IMPLEMENTATION

Probability floor. Proposition 2 concerns the exact Shannon entropy. The implementation replaces each probability p by $q = \max(p, 10^{-8})$ and evaluates $-\sum q \log q$ without renormalizing, so the floor enters both factors. We checked its effect on three fixed development states, one healthy, one at the alarm and one after an established collapse, with 144 heads each. Removing the floor changed the per-head logit gradient by at most 1.37×10^{-6}, 2.56×10^{-6} and 3.49×10^{-6} in relative norm, and the smallest cosine between the gradients with and without the floor was 0.9999999999946. The criterion fixed before the check, a relative change above 0.01 or a cosine below 0.99, flagged no head in FP32 or FP64 entropy calculations. The check uses fixed FP32 logits; it does not establish equality of BF16 optimizer updates or of training trajectories.

Spectral sharpness. We estimate $\|L(S)\|_2$ with five power iterations on the materialized centered score matrix of each sampled sequence and head.

V PARAMETER STABILITY

The coefficient of SPECGATE was selected on development seeds 3 and 4 of the 124M repair task from $\{0.3, 1, 3\}$ with the predeclared expansion by one step to 10 (Table 8): final loss falls from 5.388 at 0.3 to 4.832 at 1 and 4.720 at 3, and is 4.727 at 10. A separate 30M study of protection from initialization, run later with development seeds 6100 and 6101 and the grid $\{0.3, 1, 3, 10\}$, selected 3 at learning rates 0.003 and 0.006 and 10 at 0.02 and 0.1 (Table 19); its runs use the value selected at each learning rate. The 30M study therefore supports a default between 3 and 10 but does not validate the fixed value 3 at every rate. We use $\lambda = 3$ as the default of the method and keep it unchanged in the 124M repair and panel-set comparisons; analyses that vary the coefficient, such as the development grid above, say so where they are reported. The same 30M study selected a QK-Clip threshold of 30 at five of seven learning rates; the main text uses the published threshold of 100 and, marked ‡, the threshold of 30 selected for the 124M repair task (Appendix P).

W COST

In a separate A100-SXM4-40GB benchmark with three blocks per method and backend, each of 200 measured updates after 50 warm-up updates and a monitor reading every 50 updates, active SPECGATE added 35.9% per update with the optimized implementation and 60.9% with the reference implementation, relative to each backend's no-monitor baseline. Monitoring alone added 0.34% and 0.91%, respectively. The panels of the main text ran on 80-GB A100s, the studies of Appendix T on 40-GB A100s. A closed gate does not by itself save the computation of the statistic in this implementation, and the benchmark does not compare alarm-started with always-on treatment at equal final loss. Repair at equal remaining time, which includes this cost, is compared in Appendix T.