SPECTRAL BLIND SPOTS: STRUCTURE, CONTROL, AND FUNCTIONAL EFFECTS IN LANGUAGE MODEL HEADS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model output heads can develop highly concentrated singular spectra. We investigate what shapes this concentration, how it can be controlled, and when spectral intervention improves prediction. Under matched from-scratch training, spectral skew increases as the hidden-dimension-to-vocabulary ratio falls and hidden dimension grows. An empirical relation predicts held-out architectures within this regime, while pretrained-family comparisons expose its transfer limits. We introduce Spectral Skew Regularization (SSR), which penalizes a dominant-to-weaker singular-value ratio above an empirical threshold. Its pure-flow gradient lowers the dominant endpoint and raises the weaker one; with task gradients, we derive a sufficient condition for spectral descent. Experiments demonstrate spectral control across scales. Three-seed continuations at Pythia-410M and 1B improve on task-only training. SSR outperforms spectral-norm penalties on TinyStories at 410M, but SN performs better on both TinyStories and WikiText-2 at 1B. Prediction-preserving transformations and controlled perturbations show why lower skew alone does not imply better prediction. These results establish a conditional role for selective spectral intervention, with functional benefits that depend on the model and continuation setting.

1 INTRODUCTION

The LM head maps hidden states to vocabulary logits, with singular values measuring gain along different input directions. The largest singular value is about forty times the tenth in Pythia-160M, compared with eight in several larger Pythia heads. We call strong concentration *rank collapse*, not exact rank deficiency. This imbalance raises two questions: whether its structure is reproducible, and whether changing it improves prediction. We study the first through matched architectural sweeps. To investigate the second, we introduce selective spectral control and compare prediction under matched continuations.

What is known. The softmax bottleneck bounds logit-matrix rank by hidden dimension (Yang et al., 2018), motivating studies of LM-head expressivity and gradient transmission (Godey et al., 2024; Godey & Artzi, 2026). Output-embedding anisotropy (Gao et al., 2019), deep-linear mode dynamics (Saxe et al., 2014), and attention-layer spectral gaps (Nait Saada et al., 2025) motivate examining geometry beyond rank alone. Geometry still requires functional interpretation: neural collapse describes structured class separation (Papyan et al., 2020), and effective rank need not predict language-model performance (Kulkarni et al., 2026). We therefore examine the final head's spectrum alongside task outcomes.

What is missing. Spectral normalization constrains the largest singular value (Miyato et al., 2018); orthogonal regularization encourages orthogonality (Bansal et al., 2018). Spectrum Control shapes output-embedding singular values toward a slow-decaying prior through an SVD parameterization (Wang et al., 2020). Our objective instead controls a dominant-to-weaker gain ratio above a threshold, without prescribing the full spectrum. We test when this selective control improves prediction over task-only training and alternative spectral constraints. The threshold specifies when to intervene, not whether a head is functionally impaired.

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

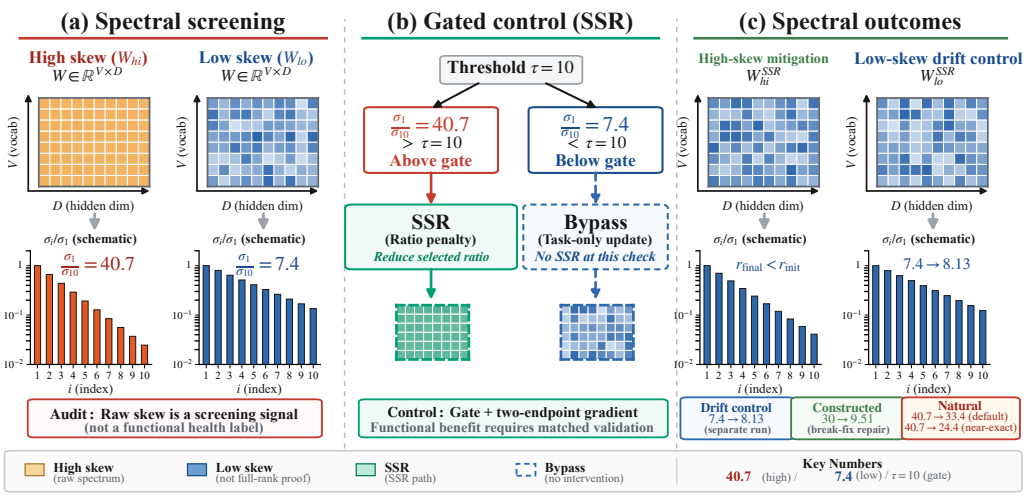


Figure 1: **Spectral screening and gated control.** Matrices and normalized spectra are schematic, not measured entries or full spectral profiles. Gate branches show above-/below-threshold checks, not functional health classes. The numerical outcomes are from separate settings: limiting drift ($7.4 \rightarrow 8.13$), constructed spectral repair ($30 \rightarrow 9.51$), and partial reduction of pretrained skew ($40.7 \rightarrow 33.4/24.4$). Functional benefit requires the paired comparison in Section 4.

Our approach. We characterize concentration under matched training through hidden dimension D and vocabulary size V. Spectral Skew Regularization (SSR) controls the dominant-to-weaker singular-value ratio above a threshold. We analyze its two-endpoint gradient and competition with task gradients. A lower raw ratio need not change prediction: centering removes common-logit shifts without changing softmax probabilities. We report centered spectra and compare continuations from identical starting heads to distinguish spectral change from functional repair relative to task-only training.

Our contributions are threefold:

1. **A two-factor empirical characterization.** Under matched from-scratch training, spectral skew varies with D/V and D. We test a frozen predictor on held-out architectures and document where it fails across pretrained families (Section 2). The relation describes the tested training and preprocessing regime.
2. **Conditional spectral control.** SSR couples an explicit gate to a two-endpoint ratio gradient. We derive conditional finite-time gate hitting under task-gradient competition and test prevention, constructed repair, and estimator-sensitive mitigation of pretrained skew (Section 3).
3. **Matched evidence for conditional repair.** Three-seed continuations at 410M and 1B show gains over task-only training. SSR has lower in-domain NLL than spectral-norm penalties across tested strengths at 410M; SN is better on both corpora at 1B. A below-gate 160M control yields identical paired weights. Controlled perturbations separate ratio reduction from task benefit (Section 4, Appendices E.4, E.5, and F).

Appendix J compares related work on LM-head expressivity, rank collapse, spectral dynamics, and spectrum-control methods.

2 Diagnosing and Characterizing Spectral Concentration

2.1 Spectral Measurements and Their Interpretation

The LM head maps a hidden state $h \in \mathbb{R}^D$ to vocabulary logits $z = Wh$, with $W \in \mathbb{R}^{V \times D}$. In its SVD, $W = \sum_i \sigma_i u_i v_i^\top$, each σ_i is the gain applied to input direction v_i. A concentrated spectrum therefore amplifies some directions much more than others. Whether that imbalance affects predic-

tion depends on the hidden states and the relative logits: adding the same value to every logit leaves softmax unchanged.

Concentration and functional effects. We screen spectral concentration with $r = \sigma_1/\sigma_{10}$ and corroborate it with energy rank. Centering gives $W_c = W - \mathbf{1}\bar{w}^\top$, where $\bar{w}$ is the mean vocabulary row. It removes a common logit shift without changing predictions: pretrained 160M changes from $r = 40.70$ to $r(W_c) = 4.42$. We use *spectral control* for limiting the ratio and *spectral repair* for reducing an established excess. *Functional repair* requires improved task performance relative to a matched task-only control; it need not restore pretrained performance. The gate selects parameter states for control, not models known to have a functional deficit.

Diagnostic measurements. The gate uses $r = \sigma_1/\sigma_{10}$; eff_{90} counts the directions containing 90% of squared singular-value energy. The controlled trends persist for $k \in \{5, 10, 20\}$, with threshold recalibration required for each k (Appendix Table 31). Directional correlation ρ is descriptive, not a gate signal (Appendix D). These are standard measurements, not newly proposed metrics. We place $\tau = 10$ in the observed raw-spectrum gap between pretrained 160M (40.7) and the other five calibration heads (7.4–8.6; Figure 2). This is an operating threshold, not a population classifier. The τ-sweep characterizes intervention sensitivity (Appendix Figure 11); centering and task evaluation address different questions and do not enter the implemented gate.

2.2 An Empirical Relation under Matched Training

Pretrained Pythia heads differ markedly: 160M has σ_1/σ_{10}=40.7 and eff_{90}=1, while 410M, 1B, and 2.8B have skew 7–9 (Table 4). Within 160M, the head is the primary concentration site: backbone layers have median skew 2.1, with only one attention projection exceeding τ (Appendix E). These comparisons mix architectural and training differences. We use matched from-scratch sweeps to examine variation with vocabulary size, hidden dimension, and depth.

Low-rank biases in matrix factorization (Gunasekar et al., 2017; 2018; Arora et al., 2019) and deep networks (Huh et al., 2023) motivate examining spectral modes, but do not derive LM-head dynamics. In our matched depth sweep ($L = 4, 6, 8$), skew is 41.08, 40.87, and 40.33; at higher learning rate it is 64.35, 69.61, and 68.77. Depth is therefore treated as protocol-dependent rather than an independent predictor (Appendix G).

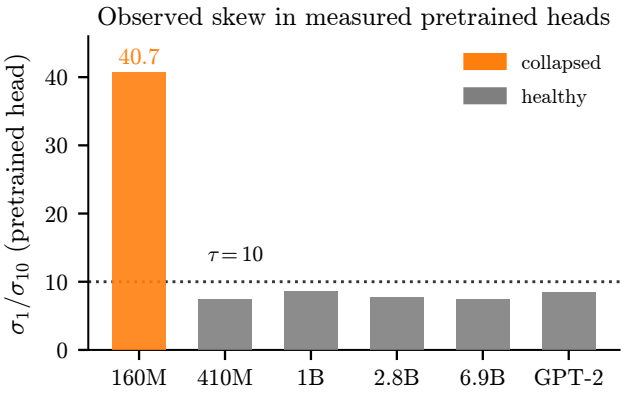


Figure 2: Observed separation in six pretrained heads. Of the measured LM heads (Pythia-160M–6.9B and GPT-2; no fine-tuning), Pythia-160M has σ_1/σ_{10}=40.7 while the other five lie at 7.4–8.6. The operating point τ=10 lies in this sample gap; six heads, with one high-skew case, do not establish population-level bimodality.

Head-to-vocabulary ratio (D/V). Holding D=512 and L=6 fixed, TinyDecoder skew rises from 23.8 to 40.9 across $V \in \{4096, 8192, 16384\}$ as D/V falls. Historical preprocessing also aliases out-of-range GPT-2 IDs into one boundary class at a V-dependent rate (Appendix D), so this is a matched association, not a pure geometric intervention.

Absolute hidden dimension D. The second controlled factor is absolute width. Holding D/V fixed at 3.125% and L=6, skew rises with D: D=256 reaches 28.4, D=512 40.9, and D=1024 62.4 (Figure 3), slope $\beta_2 = +0.569$. Larger D adds directions the head could use, yet the observed skew still deepens at fixed ratio; the fitted relation characterizes this pattern without claiming its mechanism.

The two predictors admit a compact empirical fit:

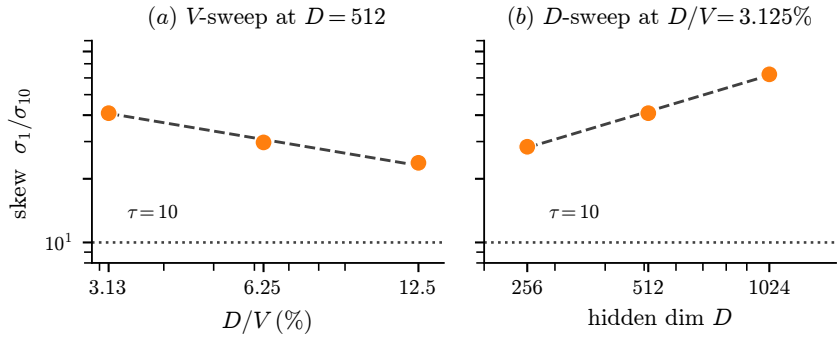


Figure 3: The two-factor relation in log-log space (τ=10 dashed; matched protocol lr 2×10^{-5}, batch 32, seed 42, 30k steps). (a) At fixed D=512, skew falls as D/V grows, slope β_1=−0.409. (b) At fixed D/V=3.125%, skew rises with D, slope $\beta_2 = +0.569$. Both panel slopes match the joint fit (R^2=0.9945).

Empirical Relation 1 (Multiplicative relation of rank collapse)**.** *We summarize skew by* $\log \widehat{r} = \beta_0 + \beta_1 \log(D/V) + \beta_2 \log D$. *The discovery regression additionally includes* $\beta_3 L$ *to test depth on seven configurations:* $\beta_1 = -0.409$, $\beta_2 = +0.569$, *and* $\beta_3 = -0.005$ ($t = -0.39$). *This four-parameter fit gives log-*$R^2 = 0.9945$ *and LOO* $R^2 = 0.955$*; the two architectural signs survive leave-one-out refits. The frozen predictor omits the small depth term without refitting the retained coefficients. Its held-out performance is evaluated separately below.*

More stringently, we freeze the relation on the seven discovery configurations before evaluating six disjoint off-grid architectures ($D \in \{384, 768\}$, $V \in \{8192, 16384, 32768\}$). Without refitting, it predicts held-out log-skew with R^2=0.915, MAE=0.071, and perfect rank ordering (Spearman ρ=1.00); the residuals show a small systematic overprediction rather than a reversal of either architectural trend (Appendix Table 32). The fitted signs are stable within the controlled de-

Table 1: Discovery regression with depth adjustment on seven configurations; coefficient uncertainties are OLS standard errors, not seed variation. The log/LOO scores refer to the four-parameter fit including $\beta_3 L$; the held-out score uses the frozen predictor with the depth term omitted.

coefficient	estimate	interpretation
β_0 (intercept)	-1.22 ± 0.24	–
β_1 (D/V)	-0.409 ± 0.025	larger ratio, lower skew
β_2 (D)	$+0.569 \pm 0.034$	larger hidden dim, more collapse
β_3 (L)	-0.005 ± 0.012 (n.s.)	per-layer factor ($e^{\beta_3} \approx 1.0$)
R^2 (log / LOO CV)	0.9945 / 0.955	–
R^2 (six held-out; no refit)	0.915	MAE 0.071; Spearman ρ=1.00

sign: $\beta_1 < 0$ and $\beta_2 > 0$; at fixed vocabulary the fitted net D-effect is $\beta_1 + \beta_2 = +0.160$. Because $\log(D/V) = \log D - \log V$ shares the $\log D$ coordinate, causal interpretation requires the separate V- and fixed-D/V sweeps. The chosen discovery design makes the two fitted regressors sample-orthogonal (correlation 0.00), but this does not remove protocol or preprocessing confounds.

The fitted exponents are protocol-specific empirical coefficients. A separable spectral scaling cancels in the ratio, so a mechanistic derivation would require mode-dependent scaling beyond the deep-linear motivation of Saxe et al. (2014).

2.3 Transfer Limits across Pretrained Families

Real pretraining introduces differences absent from the matched sweeps. Even at 7B scale, Qwen2.5-7B at D/V=2.36% has skew 10.16, whereas Falcon-7B at 6.99% has skew 3.62. These cross-family observations do not isolate head geometry.

Limits of architectural prediction. GPT-2 (124M) provides the clearest example: it shares the head geometry of Pythia-160M (D/V=1.53%, L=12) and lies in the predicted collapse regime, yet its head has low skew (8.4) while Pythia-160M collapses to 40.7: matching these head dimensions is therefore insufficient to explain the difference. Quantified over the eight pretrained heads of Table 4, predictions run 2–14× above observation (predicted 70–141 vs. observed 7.4–40.7; closest on the collapsed 160M head, 1.7×, 70 vs. 40.7). Six held-out heads from Bloom, SmolLM, and OPT reproduce the overprediction (66–143 vs. 6.6–15.2), with no reliable D/V direction across families (Bloom-560M, the lowest-D/V head, has one of the lowest ratios). The two-factor model captures the trend over controlled architectures, but family variance in real pretraining is unmodeled.

The empirical relation motivates monitoring, but deployment uses the measured ratio rather than an architectural prediction. Our diagnostic is an audit, not a health classifier: raw screening identifies candidate states, centering checks softmax-invisible concentration, and matched validation tests benefit after intervention. Only the raw screening signal enters the implemented gate (Appendix D).

3 Spectral Skew Regularization (SSR)

SSR constrains the gain ratio between a dominant and a weaker spectral direction without prescribing the full spectrum. The ratio is unchanged by uniform rescaling. A threshold suspends the penalty once the measured ratio falls below the operating point.

3.1 Gated Ratio Objective and Spectral Dynamics

SSR is a raw-spectrum-gated regularizer:

$$\mathcal{L}_{\mathrm{SSR}} = \lambda \cdot \mathbf{1}[\sigma_1/\sigma_{10} > \tau] \cdot (\sigma_1/\sigma_{10} - \tau). \tag{1}$$

Why two endpoints? Under pure SSR flow, the dominant singular value decreases and the tenth increases. The tenth singular value is a fixed anchor, not a measure of the entire spectral tail. Matched task evaluations determine whether this spectral change improves prediction. Algorithm 2 gives the training loop.

Proposition 1 (Exact active-stratum gradient)**.** *Let $r(W) = \sigma_1(W)/\sigma_{10}(W)$. Where σ_1 and $\sigma_{10} > 0$ are simple, matrix differential calculus gives exactly (Ionescu et al., 2015) $\nabla r = \sigma_{10}^{-1} u_1 v_1^\top - \sigma_1 \sigma_{10}^{-2} u_{10} v_{10}^\top$ and $\|\nabla r\|_F^2 = (1 + r^2)/\sigma_{10}^2$. Thus pure active SSR decreases σ_1 and increases σ_{10}; the approximation in our implementation comes from randomized SVD, not from differentiating a simple singular value (Halko et al., 2011).*

Conditional spectral descent and gate hitting. Intuitively, skew falls whenever SSR's projected repair force exceeds the opposing component of the task gradient. Formally, write $g = \nabla r$ and $\mu = -\langle \nabla \mathcal{L}_{\mathrm{task}}, g\rangle / \|g\|_F^2$. Exact active gradient flow gives $\dot{r} = -(\lambda - \mu)\|g\|_F^2$, where $\lambda - \mu = \lambda + (\|\nabla \mathcal{L}_{\mathrm{task}}\|_F / \|g\|_F) \cos\theta$. Here θ is the angle between the task and ratio gradients. If this margin is at least $\delta > 0$ and $\sigma_{10} \le M$ until gate hitting, then $t_{\mathrm{hit}} \le M^2[\arctan r(W_0) - \arctan \tau]/\delta$. The result certifies conditional first hitting, not global Transformer convergence or post-hinge invariance; Appendix A gives the derivations and implementation boundaries.

Implementation. We check its randomized estimate every $T = 50$ optimizer steps and apply the penalty only at an above-threshold check. Task gradients, estimator error, and AdamW can change the pure-flow behavior; the guarantee concerns the stated projected-margin conditions, while experiments measure realized control and repair.

Computational cost. The penalty uses randomized SVD (ℓ=15, two power iterations), with amortized spectral work $O(DVk/T)$ at detection interval T=50. The retained production logs mix

training, scheduled dense probes, and checkpoint I/O, so they do not support an isolated percentage-overhead claim; Appendix H states this measurement boundary.

3.2 Comparison with Existing Spectral Constraints

Orthogonal regularization acts across singular values, while the SN penalty targets the largest singular value (Bansal et al., 2018; Miyato et al., 2018). SSR instead targets their selected ratio, decreasing its numerator and increasing its denominator under pure flow. Its gate is a separate design choice: applying the same gate to the alternative objectives tests whether selectivity alone explains the results (Section 4).

Weight tying (Inan et al., 2017; Press & Wolf, 2017) couples the input embedding and output head but does not remove spectral concentration: a tied from-scratch head reaches skew 37.98 at $D/V = 6.25\%$, versus 10.02 with SSR. This tests compatibility with tying, not a general ranking of tied and untied models.

4 Experiments

4.1 Experimental Setup

Protocol. Our default direct-pretrained LM comparison freezes the backbone and updates only the head for 5000 steps on TinyStories 10K (lr 5×10^{-5}, effective batch 16, seed 42). Benefits therefore reflect output-layer adaptation, not improved backbone learning. TinyStories (Eldan & Li, 2023) supplies a step-matched intervention budget. The paired continuations below use a separate first stage and multiple seeds; cross-modal protocols are given in Appendix I. Joint and from-scratch experiments train the full model (Table 10).

Models. Pythia-160M through 6.9B (Biderman et al., 2023), GPT-2 (Radford et al., 2019), Qwen2.5-7B (Qwen Team, 2024); Whisper-small and ViT-B for cross-domain checks. Configurations in Table 4.

Interventions and measurement. Gated SSR is compared with no intervention, orthogonal regularization, a top-singular-value penalty (SN-penalty), head-only σReparam (Zhai et al., 2023), and a matched LM-head adaptation of Spectrum Control (Wang et al., 2020). SSR updates use randomized SVD; reported checkpoint spectra and effective ranks use dense SVD remeasurement unless a table explicitly states otherwise. Estimator labels therefore describe the direction used for training, not an incommensurate terminal measurement.

4.2 Spectral Control and Matched Prediction Outcomes

Table 2 separates spectral control from prediction. Blocks (a–b) test control from pretrained heads; block (c) adds constructed perturbations and training checks. Block (d) compares continuations from shared ortho-trained heads, testing whether SSR improves prediction given the same starting point and training history. This comparison measures benefits during continuation, not recovery from naturally occurring pretrained concentration.

Spectral control. In Table 2(a–b), SSR limits all four tested 1B–7B endpoints below τ, while only partially reducing natural 160M skew. The latter retains $\mathrm{eff}_{90} = 2$: ratio reduction is not rank restoration. Block (c) tests constructed repair, joint onset, and full-training prevention; the constructed case lacks a task-only continuation, and overlapping training-loss ranges do not establish functional equivalence. Block (b) is single-seed; block (d) separately tests matched function at 1B.

Prediction under matched continuations. Matched continuations show that SSR can improve prediction, but lower skew alone does not consistently rank prediction outcomes across the tested interventions. Block (d) starts after 5k steps of orthogonal regularization, below the gate. For 410M and 1B, independent stage1 seeds 42–44 each supply four matched 5k-step continuations. The 160M control varies continuation seeds from one fixed stage1 head. Token-weighted NLL uses the same

Table 2: **Spectral control and matched task outcomes.** (a–c) report terminal raw skew. (d) compares continuations after ortho, with three independent pipeline seeds per model; lower NLL is better. Bold marks the lowest mean, not statistical significance.

(a) Head-only fine-tuning (5k steps): raw skew

Model	r_0	Task-only	Ortho	SN	σRep	SC†	SSR (ours)
Pythia-160M	40.70	38.87	13.99	14.26	38.05	5.27	33.38
Pythia-410M	7.37	10.35	3.34	2.10	6.36	7.33	8.13

(b) Scale extension (head-only, 5k steps): raw skew

Model	r_0	Task-only	SSR (ours)	Reduction vs. task-only
Pythia-1B	8.58	11.27	8.86	2.41
Pythia-2.8B	7.66	10.76	9.91	0.85
Pythia-6.9B	7.44	11.31	8.49	2.82
Qwen2.5-7B	10.16	10.70	8.24	2.46

(c) Repair, joint training, and estimator checks: raw skew

Setting	r_0	Task-only	SSR (ours)	Additional measurement
410M constructed (5k)	30.00	n.r.	9.51	$\mathrm{eff}_{90} : 6 \to 6$
410M joint onset (20k)	7.37	29.74	8.81	Loss: .0173 / .0187‡
TinyDecoder (30k)	n.p.	110.2 ± 4.3	10.30 ± 0.11	3 seeds; loss ranges overlap
160M rSVD (5k)	40.70	38.87	33.38	Near-exact skew: 24.37

(d) Matched continuation (5k steps): NLL, mean $\pm$ sample s.d.

Model / corpus	Task-only	SN	Gated SN	SSR (ours)
410M / TinyStories	$1.5651 \pm .0031$	$1.5542 \pm .0049$	$1.5559 \pm .0042$	$\mathbf{1.5503 \pm .0037}$
410M / WikiText-2	$10.1838 \pm .2176$	$\mathbf{7.6691 \pm .0172}$	$8.9539 \pm .1018$	$8.6027 \pm .1542$
1B / TinyStories	$1.4593 \pm .0018$	$\mathbf{1.4458 \pm .0038}$	$1.4505 \pm .0030$	$1.4510 \pm .0031$
1B / WikiText-2	$9.3534 \pm .1795$	$\mathbf{7.5606 \pm .0107}$	$8.4330 \pm .0827$	$8.3780 \pm .0964$

(a–c): seed 42 unless marked; 160M task-only is a three-seed mean (s.d. 0.0015). SSR uses $\lambda = 0.1$ except in (b)'s 1B row (0.01). Ortho: orthogonal regularization; SN: top-singular-value penalty; σRep: head-only σReparam; SC: Spectrum Control. †SC reports three-seed means with a factorization confound (Appendix Table 6). n.r.: task-only endpoint not reported; n.p.: no pretrained initial state. ‡Task-only / SSR training loss. (d): 256 records per corpus, up to 512 tokens each; SSR $\lambda = 0.1$ for both models, SN/gated SN 0.01. Protocols and spectra: Appendices E.4 and F.

256 records per dataset within each model. This tests protection during continuation, not recovery from natural pretrained skew (Appendices E.4 and F).

160M: no additional repair. Every paired final head is identical, as are the validation losses. A low endpoint alone would have misattributed task-only recovery to SSR.

410M: replicated benefit during continuation. Across three independent full pipelines, SSR lowers centered skew from 8.461 to 2.350 and validation NLL from 1.5651 to 1.5503 on TinyStories and from 10.1838 to 8.6027 on WikiText-2 records. Every pair improves on both datasets. However, WikiText-2 NLL remains worse than stage1 (7.6629) and pretrained (3.2968): SSR mitigates additional deterioration rather than restoring generalization. These record-level NLLs differ from the concatenated protocol in Table 9.

Comparison with SN. A matched half/base/double strength sweep tests whether SSR's in-domain benefit depends on one coefficient. Every SSR setting has lower TinyStories NLL than every tested SN or gated-SN setting in each of the three seeds. SSR also retains larger raw and centered σ_{10}, whereas ungated SN consistently gives lower WikiText-2 NLL. This supports a domain-specific advantage across the sampled strengths, not global tuning optimality or a causal role for σ_{10} (Appendix E.4).

1B: benefit transfers, method ranking does not. With the 410M base strengths fixed, SSR improves both NLLs over task-only in every seed, but SN outperforms SSR on both. SSR controls raw skew to $9.987 \pm .007$, versus $42.247 \pm .176$ for task-only. Its advantage over SN is therefore specific to the tested 410M continuation (Appendix F). Starting directly from pretrained 410M, SN also outperforms SSR on HellaSwag (.378 vs. .366) and WikiText-2 perplexity (78.7 vs. 146.7).

Comparison with Spectrum Control. A matched Spectrum Control adaptation reaches three-seed skew 5.27 on 160M and 7.33 on 410M. Its 410M HellaSwag/WikiText-2 results (.355/254.7) show that strong spectral flattening does not automatically preserve either downstream measure. A zero-penalty factorized control explains most of its short-pilot spectral change, so this is an applicability comparison rather than clean objective attribution (Appendix Table 6).

Scope of σReparam. We use $W_{\text{eff}} = \gamma W/\widehat{\sigma}_1(W)$ with learnable γ, initialized to the exact initial σ_1, and one power iteration per forward pass (Zhai et al., 2023). This head-only adaptation controls scale without directly changing the ratio; it does not test the original Transformer-wide stabilization method.

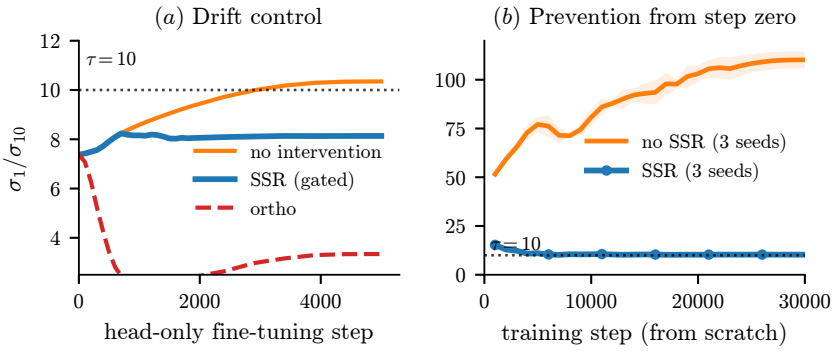


Figure 4: The gate in action. (a) *Drift control* on Pythia-410M (head-only, 5,000 steps): no intervention drifts $7.4 \rightarrow 10.35$, gated SSR limits the increase in skew ($7.4 \rightarrow 8.13$), unconditional ortho reduces the ratio to 3.34. (b) *Prevention from step zero* (TinyDecoder, 30k steps, three seeds): terminal skew is 110.2 ± 4.3 without SSR and 10.30 ± 0.11 with SSR (mean $\pm$ sample s.d.). Curves show seed means; shaded bands span one sample standard deviation on either side.

4.3 Sensitivity and Intervention Analysis

A τ sweep on Pythia-410M (Appendix Figure 11, Table 20) characterizes this operating point: lowering τ from 15 to 8 increases intervention (terminal skew 10.35 to 6.84), while the collapsed 160M head activates at every τ. Within each reported sweep, SSR is insensitive to λ over the tested range, whereas ortho and SN yield lower ratios at the tested strengths. This does not establish universal transfer of one λ: the 1B scale row uses λ=0.01, while the other main SSR rows use 0.1.

Estimator fidelity. Because practical SSR differentiates through randomized SVD, we test whether the approximation budget controls the surrogate gradient; per-checkpoint skew error, gradient cosine, norm, and projected-coefficient statistics are in Appendix A (Table 3). To test whether this approximation error matters end-to-end, we repeat the 5K Pythia-160M run with a near-exact estimator $(p, q) = (40, 4)$. The terminal reference skew falls from 33.38 to 24.37, increasing the repaired fraction of the initial gap to τ from 23.9% to 53.2%, with similar final task losses (2.2553 vs. 2.2554). Estimator fidelity thus limits the realized repair of the default implementation; even near-exact remains substantially above τ=10.

Trajectory-level check. On the default 5k Pythia-160M run, all 99 SSR kicks and 99 task-only drifts decrease reference skew; the total change $40.70 \rightarrow 33.38$ decomposes into -0.4488 during kicks and -6.8737 during intervening updates (25% from σ_1, 75% from σ_{10}). This is pathwise accounting, not a causal decomposition: SSR changes the starting points of subsequent task updates, so their decrease cannot be attributed to training without SSR.

Gate–objective ablation. With the same gate and $T = 50$ schedule, gated ortho and gated SN keep 410M below threshold (8.24/8.21). On 160M they reach 36.03/36.71, versus SSR's 33.38 and no intervention's 38.87 (Table 5). Gating supplies selectivity. At these settings SSR yields deeper repair through σ_{10} growth, while the alternatives mainly reduce σ_1. These single-seed results do not

establish superiority after equal-budget tuning. The separate three-seed continuation comparison above tests functional benefit against SN and gated SN.

Local spectral and task effects. The local probes include top-only, tenth-only, random rank-one, and common-logit-shift controls. Isolated SSR steps on high-skew 410M heads reduce the ratio but can worsen validation NLL. Forks from saved AdamW states combine SSR with task gradients: all 12 checks reduce skew, with small mean NLL gains over task-only updates on both datasets. Top-only controls give larger immediate task gains, while exact SSR produces the largest ratio decrease. Larger spectral changes therefore need not yield larger task gains (Appendix E.5).

4.4 Other Training Settings and Modalities

The geometric association has limited transfer evidence. Low-D/V Whisper heads have skew 36–40 and high-D/V ViT heads 1.4–1.6, but held-out LM families do not preserve reliable ordering (Appendix Figure 12). A ViT-B V-sweep crosses τ near $D/V{=}9.6\%$, with samples per class and task difficulty co-varying (Appendix Figure 14, Table 30).

Control also depends on the training setting. Joint fine-tuning partially reduces pretrained 160M skew ($40.7 \rightarrow 33.6$), while three-seed TinyDecoder training limits emerging concentration at overlapping observed loss ranges (Appendix Tables 11, 23). Constructed break–fix reaches $30 \rightarrow 9.51$; a separate Frobenius-matched skew-40 perturbation reduces diversity. These constructions test specific interventions, not the effects of natural collapse (Appendix Figure 8). Whisper intervention evidence remains single-run (Appendix I).

5 Discussion

What transfers across heads? The relation predicts held-out configurations within the training generator, not pretrained-family rankings. The discovery fit mainly uses seed 42; seed 43 diverges, and vocabulary remapping changes token frequencies. Neither the relation nor $\tau{=}10$ determines whether a new head needs repair. Transfer requires spectral measurement and task evaluation (Appendix D).

When is SSR useful? The gate selects when to penalize skew, not when prediction will improve. After ortho, SSR beats the tested SN penalties on TinyStories at 410M, but not at 1B; SN gives lower WikiText-2 NLL at both scales. Neither restores pretrained WikiText-2 performance. Method choice therefore depends on the continuation setting and evaluation domain (Appendices E.4 and F).

The continuations identify intervention effects, not their spectral cause. SSR retains larger centered σ_{10} across the strength grid, but recovered directions are not identified. Temperature scaling narrows the WikiText-2 gap to SN without reversing it (Appendix F.1); top-only controls give larger immediate gains in optimizer-state-matched probes (Appendix E.5). These checks leave the mechanism of continuation gains unresolved.

Spectral control depends on estimator fidelity. On pretrained Pythia-160M, increasing the randomized-SVD budget lowers terminal skew from 33.38 to 24.37 with almost unchanged task loss. Neither reaches τ, but the default endpoint is not an intrinsic repair limit. Estimator fidelity must be checked before attributing stalled control to task-gradient conflict.

6 Conclusion

LM-head spectral concentration has reproducible structure within our matched training regime. SSR selectively controls a dominant-to-weaker gain ratio and improves validation NLL over task-only continuation after orthogonal regularization at 410M and 1B. Its TinyStories advantage over spectral-norm penalties holds at 410M but reverses at 1B. These results distinguish the presence of concentration, the ability to control it, and the functional benefit of doing so. Selective ratio control can aid adaptation, but the amount of spectral flattening alone does not determine its value or its advantage over other penalties.

STATEMENTS

Ethics. This work uses public pretrained models and benchmarks and no private or personally identifiable data. SSR can alter token probabilities and generation diversity, so behavioral and safety evaluation remains necessary before deployment; spectral health is not a proxy for safety or alignment.

AI use. AI-assisted tools supported drafting, language editing, reviewer-style feedback organization, implementation discussions, and figure/table preparation. The authors reviewed and verified all assisted content and outputs and take full responsibility for the manuscript.

Reproducibility. Appendices G, D, and A provide protocols, pseudocode, and proofs. The prepared reproducibility package includes CPU-only audits, continuation configurations and per-record results through 1B, local-probe code and records (Appendix E.5), and an explicit-input evaluation script. Checkpoint weights are not bundled. Clean-environment end-to-end retraining and GPU validation of the packaged evaluation entry point remain pending.

REFERENCES

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep CNNs? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4792–4808, 2024. doi: 10.18653/v1/2024.acl-long.263.

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

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent English? *arXiv preprint arXiv:2305.07759*, 2023.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *International Conference on Learning Representations (ICLR)*, 2019.

Nathan Godey and Yoav Artzi. Lost in backpropagation: The LM head is a gradient bottleneck. *arXiv preprint arXiv:2603.10145*, 2026.

Nathan Godey, Éric de la Clergerie, and Benoît Sagot. Why do small language models underperform? Studying LM saturation via the softmax bottleneck. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.07647.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nati Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53 (2):217–288, 2011. doi: 10.1137/090771806.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The low-rank simplicity bias in deep networks. In *Transactions on Machine Learning Research (TMLR)*, 2023.

Hakan Inan, Khashayar Khosravi, and Richard Socher. Tying word vectors and word classifiers: A loss framework for language modeling. In *International Conference on Learning Representations (ICLR)*, 2017.

Catalin Ionescu, Orestis Vantzos, and Cristian Sminchisescu. Matrix backpropagation for deep networks with structured layers. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2965–2973, 2015. doi: 10.1109/ICCV.2015.339.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Akhil Kedia, Mohd Abbas Zaidi, Sushil Khyalia, Jungho Jung, Harshith Goka, and Haejun Lee. Transformers get stable: An end-to-end signal propagation theory for language models. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23449–23531, 2024.

Atharva Kulkarni, Jacob Mitchell Springer, Arjun Subramonian, and Swabha Swayamdipta. Disentangling geometry, performance, and training in language models. *arXiv preprint arXiv:2602.20433*, 2026.

Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake Richards. Tracing the representation geometry of language models from pretraining to post-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-averse gradient descent for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

Anand Murugan. Does the LM head create a harmful gradient bottleneck? a causal test. *arXiv preprint arXiv:2608.16671*, 2026.

Thiziri Nait Saada, Alireza Naderi, and Jared Tanner. Mind the gap: a spectral analysis of rank collapse and signal propagation in attention layers. In *International Conference on Machine Learning (ICML)*, 2025.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences (PNAS)*, 117 (40):24652–24663, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *European Chapter of the Association for Computational Linguistics (EACL)*, pp. 157–163, 2017.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

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

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. Improving neural language generation with spectrum control. In *International Conference on Learning Representations (ICLR)*, 2020.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations (ICLR)*, 2018.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

Emanuele Zangrando, Piero Deidda, Simone Brugiapaglia, Nicola Guglielmi, and Francesco Tudisco. Neural rank collapse: Weight decay and small within-class variability yield low-rank bias. *arXiv preprint arXiv:2402.03991*, 2024.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Joshua M. Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40770–40803, 2023.

A SSR Optimization Theory

This appendix separates exact spectral calculus from the approximations introduced by randomized SVD and a practical optimizer. The analysis concerns the head matrix $W \in \mathbb{R}^{V\times D}$, $k = 10$, and $r(W) = \sigma_1(W)/\sigma_k(W)$ in the active region $r > \tau$.

Setting and smooth active stratum. Let $m = \min\{V, D\}$ and $1 < k < m$. We consider a trajectory on which $\sigma_1(W)$ and $\sigma_k(W) > 0$ are simple. Equivalently, the relevant two-sided spectral separation

$$\gamma(W) := \min\{\sigma_1 - \sigma_2,\ \sigma_{k-1} - \sigma_k,\ \sigma_k - \sigma_{k+1},\ \sigma_k\} \tag{2}$$

is positive. On this open stratum the two singular values and their ratio are differentiable. All flow statements below assume that the solution exists on the stated interval and remains on this stratum; if $\gamma(W)$ reaches zero before the gate is hit, the smooth analysis stops and the corresponding continuation would require subdifferential analysis, which we do not claim. Let $\langle A, B\rangle = \mathrm{tr}(A^\top B)$ and $\|\cdot\|_F$ denote the Frobenius inner product and norm.

Proposition 2 (Exact spectral gradient)**.** *On the smooth active stratum,*

$$g(W) := \nabla r(W) = \frac{u_1 v_1^\top}{\sigma_k} - \frac{\sigma_1}{\sigma_k^2} u_k v_k^\top, \qquad \|g(W)\|_F^2 = \frac{1 + r(W)^2}{\sigma_k(W)^2}. \tag{3}$$

Proof. For a simple singular triplet, the exact differential is $d\sigma_i[H] = u_i^\top H v_i = \langle u_i v_i^\top, H\rangle$ (Ionescu et al., 2015). Applying the quotient rule to σ_1/σ_k gives the first identity. Distinct singular vectors are orthonormal, hence $\langle u_1 v_1^\top, u_k v_k^\top\rangle = (u_1^\top u_k)(v_1^\top v_k) = 0$; summing the squared coefficients gives the second identity. □

Proposition 3 (Pure SSR dynamics and closed hitting time)**.** *Under active pure-SSR gradient flow $\dot{W} = -\lambda g(W)$ with $\lambda > 0$,*

$$\dot{\sigma}_1 = -\frac{\lambda}{\sigma_k}, \quad \dot{\sigma}_k = \frac{\lambda\sigma_1}{\sigma_k^2}, \quad \frac{d}{dt}(\sigma_1^2 + \sigma_k^2) = 0, \quad \dot{r} = -\lambda\|g\|_F^2 < 0. \tag{4}$$

If $r_0 = r(W_0) > \tau$ and the smooth stratum persists until hitting, then, for $E = \sigma_1(W_0)^2 + \sigma_k(W_0)^2$,

$$t_{\text{hit}} = \frac{E}{2\lambda}\left[\arctan s + \frac{s}{1+s^2}\right]_{s=\tau}^{s=r_0}. \tag{5}$$

Proof. Taking the inner product of $\dot{W}$ with $u_i v_i^\top$ and using orthogonality yields the first two equations. They imply conservation of E and, by the chain rule, $\dot{r} = \langle g, \dot{W}\rangle = -\lambda\|g\|_F^2$. Since $E = \sigma_k^2(1 + r^2)$, Equation 3 gives $\dot{r} = -\lambda(1 + r^2)^2/E$. Separating variables and using $\int(1 + s^2)^{-2} ds = \frac{1}{2}(\arctan s + s/(1 + s^2))$ proves Equation 5. □

Collision boundary of the closed dynamics. Before a singular-value collision, the other ordered singular values are constant under pure SSR. Write $a = \sigma_2(W_0)$ and $b = \sigma_{k-1}(W_0)$. The smooth endpoint solution can reach the gate before a collision only if its proposed terminal values satisfy $\tau\sqrt{E/(1 + \tau^2)} > a$ and $\sqrt{E/(1 + \tau^2)} < b$. Equality corresponds to hitting at the stratum boundary and is excluded here. Thus the closed hitting time is not a continuation rule through eigenvalue multiplicities; increasing the selected weak endpoint may exhaust its upper spectral gap before the gate is reached.

Softmax-equivalent heads and diagnostic scope. For any $a \in \mathbb{R}^D$, the head $W' = W + \mathbf{1}a^\top$ produces the same softmax probabilities as W, since $W'h = Wh + \mathbf{1}(a^\top h)$. Its raw singular spectrum need not be the same. A canonical representative for this shift freedom is $W_c = PW$, where $P = I - \mathbf{1}\mathbf{1}^\top/V$. The present experiments and guarantees concern the raw checkpoint parameterization, not an invariant diagnostic of predictive behavior. A checkpoint audit makes this distinction substantive: pretrained Pythia-160M has raw ratio 40.702 but centered ratio 4.417, while 410M changes from 7.375 to 3.626. Thus the raw $\tau = 10$ classification is not preserved under a functionally equivalent reparameterization. These measurements use float64 Gram eigendecomposition on the cached pretrained heads; synthetic hidden-vector checks give maximum softmax differences below 8×10^{-18}. They do not establish centered robustness of the fine-tuned results, whose original natural-head weights were not retained in the inspected archive. If the ratio objective is instead defined on PW, its gradient with respect to W is $P\nabla r(PW)$ on the corresponding smooth stratum, and its threshold and empirical results require fresh calibration.

Proposition 4 (Task–SSR competition and finite hitting)**.** *Assume the active exact flow exists and remains on the smooth stratum until its first gate hit. Let it satisfy $\dot{W} = -\nabla\mathcal{L}_{\text{task}}(W) - \lambda g(W)$ and define*

$$\mu(W) = -\frac{\langle\nabla\mathcal{L}_{\text{task}}(W), g(W)\rangle}{\|g(W)\|_F^2}. \tag{6}$$

Then $\dot{r} = -(\lambda - \mu(W))\|g(W)\|_F^2$. If, until the first gate hit, $\lambda - \mu(W) \geq \delta > 0$ and $\sigma_k(W) \leq M$, then

$$t_{\text{hit}} \leq \frac{M^2}{\delta}\big(\arctan r(W_0) - \arctan\tau\big) \leq \frac{M^2(r(W_0) - \tau)}{\delta}. \tag{7}$$

Proof. The identity follows by the chain rule and Equation 6. Equation 3 and the assumptions give $\dot{r} \leq -\delta(1 + r^2)/M^2$, hence $\frac{d}{dt}\arctan r \leq -\delta/M^2$ while the gate is active. Integrating to the first time $r = \tau$ proves the first inequality; the second follows because $\arctan$ is 1-Lipschitz. This

establishes first hitting only: after the hinge turns off, task dynamics may move the ratio above the gate again. □

Gradient conflict is often described through pairwise gradient inner products and projection-based remedies in multi-task optimization (Yu et al., 2020; Liu et al., 2021). Here we do not modify either gradient; instead, the exact derivative of r yields the following specialization.

Corollary 1 (Gradient-angle compatibility)**.** *Let $h = \nabla\mathcal{L}_{\text{task}}(W) \neq 0$, let θ be the angle between h and g, and set $q = \|h\|_F/\|g\|_F$. Then*

$$\lambda - \mu = \lambda + q\cos\theta. \tag{8}$$

Consequently, active exact flow decreases r exactly when $\lambda + q\cos\theta > 0$. Nonnegative alignment ($\cos\theta \geq 0$) assists SSR; under conflict ($\cos\theta < 0$), the sufficient and necessary pointwise condition is $\lambda > q|\cos\theta|$. A conservative uniform sufficient condition is $q \leq \lambda - \delta$, which implies the margin assumption of Proposition 4. If $h = 0$, the margin is simply λ.

Proposition 5 (Projected randomized-SVD direction)**.** *Randomized range finding approximates the singular subspace (Halko et al., 2011); it need not produce an unbiased gradient. Let $\widehat{g}_t$ be the implemented direction, g_t the exact gradient, and $a_t = \langle g_t, \widehat{g}_t\rangle/\|g_t\|_F^2$. For the continuous direction $\dot{W} = -h_t - \lambda\widehat{g}_t$, where $\mu_t = -\langle h_t, g_t\rangle/\|g_t\|_F^2$, the exact identity is*

$$\dot{r} = -(\lambda a_t - \mu_t)\|g_t\|_F^2. \tag{9}$$

Thus the relevant condition is the projected margin $\lambda a_t - \mu_t > 0$, not unbiasedness or angular accuracy alone.

Proof. The chain rule gives $\dot{r} = \langle g_t, \dot{W}\rangle = -\langle g_t, h_t\rangle - \lambda\langle g_t, \widehat{g}_t\rangle$. Substituting the definitions of μ_t and a_t yields Equation 9; no distributional assumption on $\widehat{g}_t$ is used. □

Randomized-SVD gradient fidelity. The preceding analysis does not require the randomized-SVD (rSVD) gradient to be unbiased; it only requires its projection onto the exact ratio gradient to be favorable. We therefore directly measure how the rSVD approximation budget affects both scalar skew estimation and gradient fidelity.

We evaluate three checkpoints along the Pythia-160M head-only SSR trajectory, corresponding to reference spectral skews $r \approx 40.70$, 36.80, and 33.38. Here p denotes the oversampling dimension beyond rank k, and q the number of power iterations. For each checkpoint and each rSVD configuration, we perform 50 independent randomized trials. A near-machine-precision numerical reference is computed in FP64 from the Gram matrix $W^\top W$, followed by eigendecomposition and analytic construction of

$$g = \nabla r(W) = \frac{1}{\sigma_{10}}u_1 v_1^\top - \frac{\sigma_1}{\sigma_{10}^2}u_{10}v_{10}^\top. \tag{10}$$

For the randomized estimator $\widehat{g}$, we report the relative skew error, gradient cosine similarity, $\|\widehat{g}\|_F/\|g\|_F$, and the projected coefficient

$$a = \frac{\langle g, \widehat{g}\rangle_F}{\|g\|_F^2} = \frac{\|\widehat{g}\|_F}{\|g\|_F}\cos(g, \widehat{g}). \tag{11}$$

Increasing the approximation budget consistently improves both scalar accuracy and gradient fidelity over the tested approximation budgets. At the default setting $(p, q) = (5, 2)$, the median relative skew error is 5.93%, 7.51%, and 7.56% at the three checkpoints, respectively, while the corresponding gradient cosines are only 0.048, 0.020, and 0.014. The randomized gradient norm is substantially inflated ($9.17\times$, $19.39\times$, and $25.64\times$ the exact norm), although its projected coefficient remains positive ($a = 0.552$, 0.516, and 0.493).

In contrast, increasing the sketch size and the number of power iterations rapidly recovers the exact gradient. With $(p, q) = (40, 4)$, the median relative skew errors fall to 0.016%, 0.0033%, and 0.0009%, the gradient cosine similarities rise to 0.9990, 0.9999, and 0.99999, and the norm ratios

Table 3: Randomized-SVD fidelity across three checkpoints of the Pythia-160M SSR trajectory. Entries are medians over 50 independent randomized trials. Each triplet reports start / mid / end.

(p, q)	rel. skew err. (%)	$\cos(g, \widehat{g})$	$\|\widehat{g}\|/\|g\|$	a
$(5, 2)$	5.93 / 7.51 / 7.56	.048 / .020 / .014	9.17 / 19.39 / 25.64	.552 / .516 / .493
$(20, 2)$	1.89 / 1.12 / 0.74	.324 / .428 / .358	2.70 / 2.47 / 2.13	1.025 / 1.062 / .882
$(40, 2)$	0.73 / 0.33 / 0.18	.720 / .890 / .943	1.47 / 1.16 / 1.07	1.057 / 1.031 / 1.010
$(20, 4)$	0.105 / 0.026 / 0.008	.953 / .996 / .9995	1.07 / 1.009 / 1.002	1.016 / 1.005 / 1.001
$(40, 4)$	0.016 / 0.0033 / 0.0009	.9990 / .9999 / .99999	1.004 / 1.001 / 1.0002	1.003 / 1.001 / 1.0002

become 1.004, 1.001, and 1.0002. The corresponding projection coefficients are 1.003, 1.001, and 1.0002.

These results separate two notions of approximation quality: an rSVD estimate may provide a moderately accurate scalar skew while still yielding a poor gradient direction. The discrepancy is induced by the randomized approximation rather than by the exact SSR gradient itself: increasing the randomized-SVD approximation budget recovers the exact spectral gradient. Trajectory dependence is budget-dependent in this probe. At the default $(5, 2)$ budget, angular fidelity decreases and norm inflation grows from start to end; at $(40, 4)$, the estimator remains essentially exact and improves slightly. The data therefore do not support a budget-independent claim that late-stage spectra are equally easy to differentiate.

The fixed-checkpoint fidelity analysis alone does not determine whether this approximation error changes end-to-end training. A matched 5K control does. Replacing the default $(p, q) = (5, 2)$ estimator with the near-exact $(40, 4)$ estimator lowers terminal reference skew from 33.38 to 24.37 under otherwise identical training, with nearly unchanged final task loss. The low-budget randomized approximation is therefore a material contributor to the default 33.4 endpoint under this protocol. It is not the sole source of incomplete repair: the near-exact trajectory still remains above the diagnostic threshold $\tau = 10$.

Projection coefficient. Since $\langle g_t, \widehat{g}_t \rangle = \|g_t\|_F \, \|\widehat{g}_t\|_F \cos(g_t, \widehat{g}_t)$, the projected coefficient factors as

$$a_t = \frac{\|\widehat{g}_t\|_F}{\|g_t\|_F} \cos(g_t, \widehat{g}_t). \tag{12}$$

Hence cosine similarity alone is not the dynamical coefficient governing skew descent: the projected coefficient a_t combines angular alignment with relative gradient magnitude. A positive a_t only certifies a positive projection onto the exact ratio gradient; it does not imply that $\widehat{g}_t$ is an accurate or unbiased estimator of g_t.

Lemma 1 (Discrete projected descent)**.** *Suppose ∇r is locally L_r-Lipschitz on a neighborhood containing the update segment and $W_{t+1} = W_t - \eta_t(h_t + \lambda \widehat{g}_t)$. Then*

$$r(W_{t+1}) \le r(W_t) - \eta_t(\lambda a_t - \mu_t)\|g_t\|_F^2 + \frac{L_r \eta_t^2}{2}\|h_t + \lambda \widehat{g}_t\|_F^2. \tag{13}$$

Hence a positive first-order margin is insufficient by itself at finite step size; it must dominate the quadratic remainder. Summing Equation 13 yields a finite-step gate hit whenever the cumulative certified decrease exceeds $r(W_0) - \tau$ while the iterates remain in the stated neighborhood. More explicitly, with $m_t = \lambda a_t - \mu_t > 0$ and $d_t = h_t + \lambda \widehat{g}_t \neq 0$, the sufficient step-size condition

$$0 < \eta_t < \frac{2 m_t \|g_t\|_F^2}{L_r \|d_t\|_F^2} \tag{14}$$

implies $r(W_{t+1}) < r(W_t)$. The constant L_r is local: it can be bounded on compact subsets of the smooth stratum, but generally deteriorates as relevant singular values approach multiplicity. In particular, the observed 1.17% σ_{10}–σ_{11} gap does not justify assuming a trajectory-wide uniform constant without further measurement.

Proof. Local L_r-smoothness gives $r(W_t + \Delta_t) \le r(W_t) + \langle g_t, \Delta_t \rangle + (L_r/2)\|\Delta_t\|_F^2$. Substitute $\Delta_t = -\eta_t(h_t + \lambda \widehat{g}_t)$ and the definitions of a_t and μ_t to obtain Equation 13. Rearranging its linear and quadratic terms gives Equation 14. □

Corollary 2 (Finite-step gate hitting)**.** *Consider a sequence of pre-hit active updates satisfying Lemma 1. Suppose their update segments remain in a common smooth neighborhood with Lipschitz constant L_r, and define*

$$c_t := \eta_t(\lambda a_t - \mu_t)\|g_t\|_F^2 - \frac{L_r \eta_t^2}{2}\|h_t + \lambda \widehat{g}_t\|_F^2. \tag{15}$$

If, for some finite N, $\sum_{t=0}^{N-1} c_t > r(W_0) - \tau$, then there exists $n \le N$ such that $r(W_n) \le \tau$.

Proof. If no iterate through N hits the gate, every update is pre-hit and Lemma 1 can be summed to give $r(W_N) \le r(W_0) - \sum_{t=0}^{N-1} c_t < \tau$, contradicting $r(W_N) > \tau$. □

Checkpoint-local condition check. At Pythia-160M step 143000, $r = 40.7022$, $\|g\|_F = 2.44162$, and eight TinyStories minibatches give $\mu = -3.54 \times 10^{-4}$ on average (range $[-5.86, -1.38] \times 10^{-4}$). With $\lambda = 0.1$, all eight exact-flow margins are positive (minimum 0.100138), so the task gradient locally assists ratio reduction. This verifies the condition only at one checkpoint, not uniformly along a trajectory. For the default randomized SVD ($p = 5, q = 2$), eight repeats have positive a_t (mean 0.726, range 0.00761–1.098), but low gradient cosine (mean 0.117) and inflated norm (mean ratio 7.05). The logged relative σ_{10}–σ_{11} gap is only 1.17%; because the upper gap $\sigma_9 - \sigma_{10}$ was not logged in this probe, these measurements do not certify the two-sided separation in Equation 2, let alone a trajectory-wide lower bound. Approximation improves substantially with larger sketches and more power iterations: $p = 40, q = 4$ gives cosine 0.99897, norm ratio 1.00348, and ratio error $+0.00564$. These are local sensitivity checks, not evidence that the default estimator is accurate or unbiased.

Periodic application. When SSR is evaluated once every $T = 50$ task steps, no exact λ/T theorem follows. Let $\Delta_j = W_{j+1} - W_j$, let only the active block-start update b include SSR, and assume a common local Lipschitz constant L_r is valid along every update segment in the block. Applying Lemma 1 at every step gives the sufficient block condition

$$\eta_b \lambda \langle g_b, \widehat{g}_b \rangle > -\sum_{j=b}^{b+T-1} \eta_j \langle g_j, h_j \rangle + \frac{L_r}{2} \sum_{j=b}^{b+T-1} \|\Delta_j\|_F^2, \tag{16}$$

under which $r(W_{b+T}) < r(W_b)$. Thus the projected SSR decrease must exceed the accumulated task-driven increase and second-order remainder. Replacing this block condition by strength λ/T is only a slow-variation average-field approximation.

Spectral changes along an instrumented training run. As a complement to the local checks above, we record a full 5000-step head-only SSR run on Pythia-160M, with 100 SSR events at steps $0, 50, \ldots, 4950$. The event direction is averaged over the two microbatches that form each optimizer step, matching the implemented gradient accumulation. The step-0 kick is exactly zero because the warm-up learning rate is zero; each of the remaining 99 kicks satisfies $\Delta r_{\text{kick}} < 0$, and all 99 complete task-only drifts satisfy $\Delta r_{\text{drift}} < 0$ (the final event has no subsequent full drift interval). Cumulatively $\sum \Delta r_{\text{kick}} \approx -0.4488$ and $\sum \Delta r_{\text{drift}} \approx -6.8737$, while the numerical reference ratio falls from $r = 40.70$ to $r = 33.38$ (total change -7.3225). Numerically, 6.13% of the observed reduction accumulates during the SSR kicks and 93.87% during the intervening task-only updates. This is an empirical pathwise decomposition, not a mechanistic attribution: SSR can change later task gradients, and AdamW's stateful preconditioning prevents identifying a counterfactual task-only path from this single run.

In endpoint terms, σ_1 decreases from 678.72 to 645.85 (-4.84%) and σ_{10} increases from 16.68 to 19.35 ($+16.04\%$). Since $\log r = \log \sigma_1 - \log \sigma_{10}$, the cumulative log-ratio improvement decomposes as roughly 25% from the σ_1 decrease and 75% from the σ_{10} increase. To confirm that the drift accumulation is not an artifact of the first-order decomposition, we also record the Taylor residual

$e_t = \Delta r_t - \langle \nabla r(W_t), \Delta W_t \rangle$ in absolute value: median $|e_t| = 5.95 \times 10^{-6}$, $p_{95}|e_t| = 1.75 \times 10^{-4}$, and $\max |e_t| = 2.87 \times 10^{-4}$, indicating that the linear term dominates along this trajectory.

Matched near-exact trajectory. We repeat the same instrumented 5K trajectory with $(p, q) = (40, 4)$. The randomized direction remains essentially exact throughout the run (mean cosine 0.99992, mean projected coefficient $a = 1.00021$), and the ratio falls from 40.70 to 24.37. Across the 99 complete inter-event blocks, all post-warm-up SSR kicks and all intervening task-only drifts decrease the numerical reference ratio (the zero-lr step-0 kick is zero). Their cumulative contributions are $\sum \Delta r_{\text{kick}} = -4.257$ and $\sum \Delta r_{\text{drift}} = -12.072$, respectively, compared with -0.449 and -6.874 under the default estimator. Thus increasing estimator fidelity strengthens the direct intervention approximately $9.5\times$ and the subsequent task-only reduction approximately $1.8\times$. As before, this is a pathwise decomposition rather than a causal attribution of the intervening drift.

Reproducibility note. The legacy metrics logger records the `reg_loss` field after the optimizer-step counter is incremented, so its sparse checkpoint rows are not aligned with the SSR intervention steps; the trajectory diagnostics above are therefore logged separately at the actual intervention events, and the absence of per-step `reg_loss` entries at those events should not be read as SSR failing to trigger.

AdamW boundary. Scale invariance gives $r(cW) = r(W)$ for $c > 0$, and differentiating at $c = 1$ yields $\langle \nabla r(W), W \rangle = 0$. Decoupled isotropic weight decay therefore has no direct first-order effect on r. AdamW's diagonal preconditioning, momentum, clipping, finite steps, and optimizer history do alter the projected margin and are outside Propositions 3–4. Appendix E.5 tests single updates with inherited AdamW states; these measurements do not extend the flow theorem. Conditioning on the optimizer state, a realized update ΔW_t has the a posteriori local certificate $\langle g_t, \Delta W_t \rangle + \frac{L_r}{2} \|\Delta W_t\|_F^2 < 0$, which implies $r(W_t + \Delta W_t) < r(W_t)$; this is not a convergence theorem. We therefore claim finite-time gate hitting only under the stated verifiable margin conditions, not global convergence under AdamW. The empirical plateau near 33.4 is not a theoretical equilibrium, the architecture relation remains empirical, and τ is a calibrated diagnostic operating point rather than a theoretically optimal threshold.

B Model Configurations and Additional Comparisons

Measurement note. Unless explicitly marked approximate, reported checkpoint skew and effective rank are dense-SVD remeasurements. The labels "default rSVD" (p=5, q=2) and "near-exact" (p=40, q=4) describe the differentiable direction used by SSR during training, not different terminal measurement scales. Thus the 160M endpoints 33.38 and 24.37 are comparable dense-SVD measurements of trajectories trained with different estimator fidelities. Gate decisions made online use the corresponding randomized estimate. Falcon-7B's skew (3.62 at D/V=6.99%) is a static checkpoint measurement; its head-only fine-tuning was not run because the installed Transformers API removed `get_head_mask`.

B.1 Model Configurations and SSR Baselines

Implementation is based on PyTorch (Paszke et al., 2019). Table 4 lists the architectural specifications of all models studied, together with their LM-head spectral skew (σ_1/σ_{10}) as measured after pretraining. Model architecture parameters (layers, hidden dimension, vocabulary size) are drawn from the published specifications of each model family. SSR baseline values are read from the head-only probe at step 0 of the corresponding baseline experiment in the `ssr_results` directory (prior to any head-only fine-tuning or intervention).

†A low-D/V counterexample: this MoE LM head stays below τ (9.29) at D/V=1.35%, a boundary we hypothesize stems from the sparse-routing gradient dynamics of MoE training, a mechanism we do not isolate here and leave to future work. Its 24 expert-router heads are small softmax heads over 60 experts ($D/V \approx 34$), with measured skew 4.7–6.9. This static observation does not test the fitted relation on routing heads.

Table 4: Model configurations and pretrained LM-head SSR baselines. D/V ratio computed as hidden_size / vocab_size.

Model	Layers	Hidden (D)	Vocab (V)	D/V (%)	Tokens	σ_1/σ_{10}
Pythia-160M	12	768	50304	1.53	∼300B	40.69
Pythia-410M	24	1024	50304	2.04	∼300B	7.37
Pythia-1B	16	2048	50304	4.07	∼300B	8.58
Pythia-2.8B	32	2560	50304	5.09	∼300B	7.66
Pythia-6.9B	32	4096	50304	8.14	∼300B	7.44
GPT-2 (124M)	12	768	50257	1.53	∼10B	8.42
Qwen2.5-7B	28	3584	152064	2.36	∼18T	**10.16**
Qwen1.5-MoE-A2.7B	24	2048	151936	1.35	∼1T	9.29†

Head-only fine-tuning without intervention then drifts some heads further: 410M $7.4 \rightarrow 10.35$ and GPT-2 $8.4 \rightarrow 9.2$ over 5000 steps, and this is precisely the drift that the gated intervention later regulates.

B.2 Gate–Objective Ablation

Table 5: Gated-baseline ablation: gating controls for selective intervention, while the spectral objective determines repair depth. "none" rows are the matched no-intervention endpoints. Gain vs. none is the reduction in final skew for the already-collapsed 160M head. The four gated alternative rows are seed-42 runs; terminal spectra are dense-SVD measurements.

Model	Method	r_0	r_{5000}	σ_1	σ_{10}	Gain vs. none
160M	none	40.69	38.872	—	—	—
160M	gated ortho	40.69	36.027	593.40	16.47	2.85
160M	gated SN	40.69	36.706	603.46	16.44	2.17
160M	SSR	40.69	33.38	645.85	19.35	5.49
410M	none	7.37	10.347	—	—	—
410M	gated ortho	7.37	8.238	62.84	7.63	—
410M	gated SN	7.37	8.207	62.63	7.63	—

B.3 Direct Spectrum Control Comparison

Table 6 reports our matched LM-head adaptation of Wang et al. (2020). The selected prior and orthogonality weights were frozen from the 500-step 160M pilot by held-out TinyStories loss and then applied unchanged to both models and all three seeds. The short matched controls are essential for attribution: replacing the dense head by the factorized parameterization with all Spectrum Control penalties set to zero accounts for most of the observed pilot skew reduction. Moreover, the selected regularizer slightly worsens pilot skew relative to that factorization-only control, although it improves validation loss. The 5k rows therefore demonstrate a stable direct comparator, not an isolated causal effect of the explicit Spectrum Control penalty.

Table 6: Matched Spectrum Control (SC) comparison. Pilot rows use Pythia-160M, seed 42, and 500 steps. Final rows use the frozen selected setting for 5,000 steps and report mean ± sample s.d. over seeds 42–44, including HellaSwag normalized accuracy and WikiText-2 perplexity. Pilot downstream cells are not applicable because those checkpoints were not retained for evaluation. Validation is on 1,000 held-out TinyStories examples. Runtime includes training probes and I/O, so it is descriptive rather than a kernel-level efficiency benchmark.

Setting	Steps	Final skew	Val. loss	HS	Wiki-2 ppl	Time (min)
Dense matched baseline	500	40.299	2.7543	—	—	0.87
Factorized, zero SC penalty	500	25.818	2.5516	—	—	1.21
Selected SC	500	26.685	2.5275	—	—	3.08
Selected SC, 160M	5,000	5.269±0.002	2.0482±0.0003	.2850±.0015	499.1±4.2	30.65±0.33
Selected SC, 410M	5,000	7.327±0.002	1.7833±0.0008	.3551±.0016	254.7±7.3	58.30±0.48

B.4 Directional Correlation (ρ) Cross-Check

For a diagnostic batch, let $g_z = \mathrm{softmax}(z) - e_y$ be the cross-entropy gradient with respect to the logits and $e_i = \mathbb{E}[(g_z^\top u_i)^2]$ its energy along the ith left singular vector. Table 7 reports $\rho = \mathrm{Spearman}(e_i, \sigma_i^2)$. The measured values are strongly negative (≈ -0.7 to -0.9), meaning that weaker singular directions receive larger logit-error energy on average, not that the largest singular directions receive the largest gradients. Its near-constancy across collapsed and healthy heads means it carries no validated threshold information, so we retain it only as a descriptive cross-check and never use it to open the SSR gate.

Table 7: Spearman correlation between per-direction logit-gradient energy e_i and σ_i^2 at the final step. Negative values indicate larger error energy in weaker singular directions. The 1B, 2.8B, and GPT-2 rows pair skew and ρ from the same 5,000-step SSR checkpoint, not from the pretrained head. This statistic is descriptive, not a validated gate.

Condition	σ_1/σ_{10}	ρ
160M baseline (collapsed)	38.87	-0.78
160M SSR	33.38	-0.79
410M baseline	10.35	-0.85
410M SSR	8.13	-0.84
410M ortho	3.34	-0.47
1B SSR	8.86	-0.80
2.8B SSR	9.91	-0.88
GPT-2 SSR	9.17	-0.83
TinyDecoder baseline	113.09	-0.71
TinyDecoder SSR (λ=0.01)	16.04	-0.70

B.5 Downstream Evaluation Across Scale

Head-only fine-tuning on TinyStories modifies the pretrained head; Table 9 reports in-domain perplexity and HellaSwag for the 410M conditions, and Table 8 the scale chain. Downstream does not track spectral health: across Pythia-410M through 6.9B, HellaSwag rises monotonically with scale ($0.341 \to 0.499$) while final head-only skew stays flat (10.4–11.3); Qwen2.5-7B, at $D/V = 2.36\%$, reaches the best score (0.635). Spectral skew thus fails to rank downstream performance across models and families; the probes separate heads that benchmarks rank alike. The same observation in the intervention setting is developed in §3: SN achieves the best downstream score despite the lowest top-ten spectral ratio.

Table 8: Downstream scale chain under the identical head-only protocol (TinyStories, 5,000 steps). HellaSwag spans 0.341–0.635 and out-of-domain WikiText-2 perplexity improves from 600.7 to 107.1, although the corresponding no-intervention head skews occupy the narrow 10.4–11.3 band reported in Table 2.

Model	D/V (%)	HellaSwag	WikiText-2 ppl
Pythia-410M	2.04	0.341	600.7
Pythia-1B	4.07	0.386	226.5
Pythia-2.8B	5.09	0.473	193.1
Pythia-6.9B	8.14	0.499	168.8
Qwen2.5-7B	2.36	**0.635**	107.1

C Supplementary Training Curves & Visualizations

Table 9: Pythia-410M head-only fine-tuning, downstream evaluation. SN achieves the best HellaSwag score among the interventions despite the lowest top-ten spectral ratio; WikiText-2 perplexity exposes complementary distribution shift. SN's 0.378 is the mean of three seeds (0.3786/0.3785/0.3779, s.d. 0.0004). Head-only fine-tuning degrades out-of-domain ppl 21-fold (28.2→600.7), unconditional ortho reaches 3,031, gated SSR 146.7, SN 78.7, and corrected σReparam 321.4. The corrected σReparam row is a single seed and uses the effective reparameterized head.

Condition	σ_1/σ_{10}	ppl (Tiny)	HellaSwag	eff_{90}	Wiki-2 ppl
Stock (pretrained)	7.37	8.21	0.402	825	28.2
Head-only (none)	10.35	6.21	0.341	804	600.7
SSR (gated)	8.13	6.05	0.366	815	146.7
Ortho (unconditional)	3.34	6.40	0.310	787	3,031
SN (unconditional)	2.10	6.05	**0.378**	837	78.7
σReparam	6.36	6.14	0.360	824	321.4

Table 10: Joint head+backbone fine-tuning (TinyStories, 1,000 steps), final skew, aligned with head-only at the same step count. Head-only and joint bases nearly coincide (410M 8.59 vs. 8.74; 160M 40.03 vs. 40.06), so joint adaptation barely shifts the head's spectral trajectory; SSR keeps the 410M head below τ and trims the collapsed 160M head, the pattern of Table 2.

Model	Head-only @1k	Joint base	Joint SSR
Pythia-410M	8.59	8.74	**8.29**
Pythia-160M	40.03	40.06	37.72

Table 11: Joint head+backbone fine-tuning at increasing horizons (TinyStories, 410M head). The base head spontaneously crosses the τ=10 gate and keeps drifting with training length, while SSR holds the ratio just below τ and preserves effective rank (eff_{99}) with the reported training-loss trade-off. The 160M head at 5k (final row) sits outside the onset regime: joint fine-tuning barely moves it (40.7→39.2) and SSR trims it (40.7→33.6); its loss scale is model-specific. Collapse onset is scale-dependent; SSR prevents where onset is active.

Steps	skew σ_1/σ_{10}		train loss		eff_{99}	
	base	SSR	base	SSR	base	SSR
5,000	16.1	8.10	0.097	0.081	953	986
10,000	23.1	8.26	0.0146	0.0148	911	983
20,000	29.7	8.81	0.0173	0.0187	867	976
5,000 (160M)	39.2	33.6	0.549	0.550	591	584

Each row is a separate run; the LR schedule is tied to `total_steps`, so trajectories and loss scales are not comparable across rows (base vs. SSR within a row share a schedule).

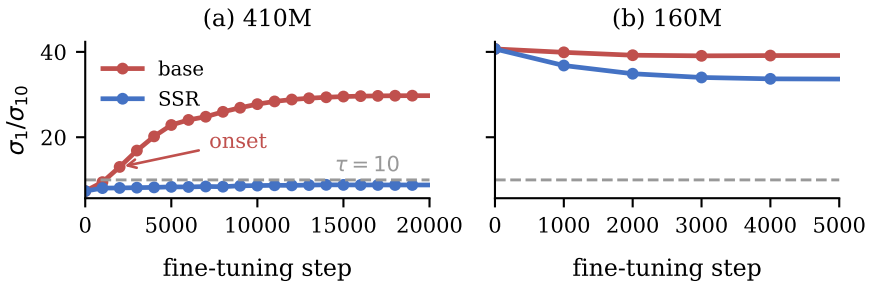


Figure 5: Spontaneous onset under joint fine-tuning (410M 20k steps; 160M 5k steps). At 410M (a) the base head crosses the τ=10 gate near step 2k and drifts (7.4 → 29.7) while SSR holds the ratio at 8.1 → 8.8; at 160M (b) SSR pulls it from 40.7 to 33.6. Onset is scale-dependent; SSR prevents it where onset is active.

Table 12: Step ablation under the identical head-only protocol, the trajectory that motivates the gate. Fine-tuning is itself a progressive spectral-degradation channel for heads near the threshold: the 410M head is already above τ by the first check (10.35 at 5K) and drifts to 18.48 by 20K, while the deeply collapsed 160M head mildly improves (38.87→33.79) as fine-tuning regrows part of a saturated tail. Whether a head at 18.48 warrants intervention is a per-family call, but the drift is what the gate is for: a head monitored as it crosses τ is detected and held before it hardens, and longer training widens the healthy-head gap without repairing the collapsed one.

Model	5,000 steps	10,000 steps	20,000 steps
Pythia-410M	10.35	13.05	18.48
Pythia-160M	38.87	37.22	33.79

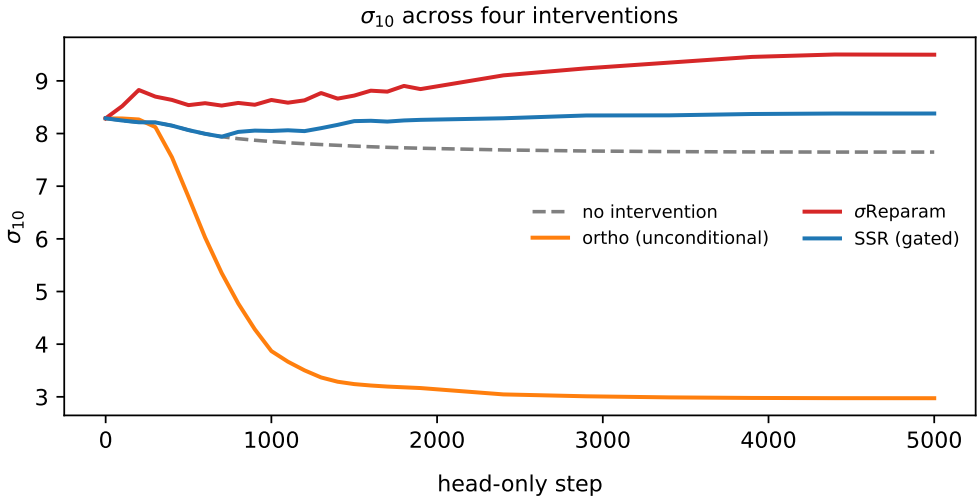


Figure 6: The tenth singular value of the healthy 410M head under four regimes (head-only, TinyStories, 5,000 steps). Baseline drifts mildly, gated SSR holds σ_{10} steady, ortho suppresses it, and the corrected σReparam applicability baseline increases it while ending at skew 6.36. The curves report dense-SVD probes of the effective head.

D EXPERIMENTAL SETTINGS AND IMPLEMENTATION DETAILS

Default hyperparameters. The diagnostic uses the top k=10 singular values with the randomized-SVD estimator (ℓ=15, two power iterations) sampled every T=50 steps. SSR's gate threshold is τ=10; its penalty weight defaults to λ=0.1, a default requiring validation alongside the gate and estimator budget, and the plateau of Table 13 shows the result is insensitive to it.

Algorithm 1 Spectral diagnostic for a single head

Require: head weight $W \in \mathbb{R}^{V \times D}$; diagnostic logits Z and labels Y; top-k=10; threshold τ=10

$(U, \sigma_1, \ldots, \sigma_d) \leftarrow \textsc{DenseSVD}(W)$, where $d = \min(V, D)$ ▷ report-only audit
$s \leftarrow \sigma_1/\sigma_{10}$ ▷ spectral skew, the primary signal
$\mathrm{eff}_{90} \leftarrow \min\{j : \sum_{i \le j} \sigma_i^2 \ge 0.9 \sum_i \sigma_i^2\}$ ▷ energy rank
$g_z \leftarrow \mathrm{softmax}(Z) - \mathrm{onehot}(Y)$
$e_i \leftarrow \mathrm{mean}\left[(g_z^\top u_i)^2\right]$ for $i = 1, \ldots, d$
$\rho \leftarrow \mathrm{Spearman}\big((e_i)_{i=1}^d, (\sigma_i^2)_{i=1}^d\big)$ ▷ descriptive only
if $s > \tau$ **then**
    status $\leftarrow$ ABOVE-THRESHOLD
else
    status $\leftarrow$ BELOW-THRESHOLD
end if
return $(s,\ \mathrm{eff}_{90},\ \rho,\ \text{status})$

Algorithm 2 Spectral Skew Regularization (SSR) at one head

Require: head weight W; penalty λ=0.1; threshold τ=10; interval T=50

for $t = 0, 1, 2, \ldots$ **do** ▷ completed optimizer steps
    Zero accumulated gradients
    for each of the B microbatches in this update **do**
        $\mathcal{L} \leftarrow \mathcal{L}_{\text{train}}$
        if $t \bmod T = 0$ **then**
            $(\sigma_1, \sigma_{10}) \leftarrow \textsc{RandomizedSVD}\,(W,\ \ell$=15, 2 power iterations$)$
            $s \leftarrow \sigma_1/\sigma_{10}$
            if $s > \tau$ **then**
                $\mathcal{L} \leftarrow \mathcal{L} + \lambda\,(s - \tau)$
            end if
        end if
        Accumulate $\nabla_W(\mathcal{L}/B)$
    end for
    Apply global gradient clipping to accumulated gradients
    Update W with AdamW; advance the learning-rate schedule
end for

Unified procedure. Algorithm 1 is the report-only audit and uses the complete spectrum because eff_{90} cannot in general be recovered from only the top ten singular values. Algorithm 2 is the training-time procedure: t counts completed optimizer steps starting at zero, randomized SVD is evaluated only at intervention steps ($t \equiv 0 \pmod T$), and the SSR penalty is included only in that same update. Each accumulated microbatch uses an independent randomized estimate at an intervention step. No spectral penalty is cached through the intervening task-only updates, and no straight-through estimator is used.

D.1 REPRODUCIBILITY: EXPERIMENT DIRECTORY MAPPING

All experiments referenced in this appendix correspond to subdirectories under `results/` (fine-tuning tables under `results/ssr_results/`; the from-scratch prevention runs in Table 23 under `results/br_attack/`). Key directory-to-table mappings:

- Table 4:`exp_pythia160m_baseline_headonly_seed42`, `exp_pythia410m_baseline_headonly_seed42`, `exp_pythia1b_ssr_seed42`, `exp_pythia2_8b_l01_headonly_seed42`, `exp_gpt2_l01_headonly_seed42`.
- Table 13:`exp_pythia160m_l001_tau10_rsvd_seed42` through `l02` (randomized-SVD estimator),plus `exp_pythia160m_ortho_seed42`, `exp_pythia160m_sn_seed42`.
- Table 20:`exp_pythia410m_tau{5,8,10,12,15}_sweep_seed42` and `exp_pythia410m_ungated_seed42`.
- Table 21, 22:`exp_pythia410m_l01_tau10_rsvd_seed42`, `exp_pythia410m_ortho_seed42`, `exp_pythia410m_destroy_seed42`, `exp_pythia410m_repair_rsvd_seed42`.
- Table 23:`br_prevent_full_base_seed42`, `br_prevent_full_base_s43_seed43`, `br_prevent_full_base_s44_seed44` (baseline, three seeds), `br_prevent_full_ssr_l0p1_seed42`, `br_prevent_full_ssr_l0p1_s43_seed43`, `br_prevent_full_ssr_l0p1_s44_seed44` (λ=0.1, three seeds), `br_prevent_full_ssr_l0p01_seed42` (λ=0.01).

Historical spectral tables use archived checkpoint metrics, with duplicate optimizer-step rows deduplicated. Continuation NLLs are recomputed from per-record loss sums and target counts in JSONL files; spectral remeasurements have separate outputs. Configurations and directional-correlation logs are retained where available, but some historical runs lack complete configuration records.

Code reproducibility notes. A fine-tuning run creates a run-specific subdirectory of the configured output directory holding `config.yaml`, `metrics.csv`, `rho_log.csv`, and `model_final.pt`; scripts resolve the data/checkpoint root through a `GA_ROOT` environment variable. The reported from-scratch runs used `TokenClampWrapper`, which maps tokenizer IDs outside the configured vocabulary to the final supported ID. The archived runs did not log this rate. Our released audit script tokenizes a reproducible random sample of 10K TinyStories validation examples (2,125,077 IDs) and finds remapped fractions 18.75%, 11.11%, 5.64%, and 1.35% for support sizes 4096, 8192, 16384, and 32768. Because this aliases thousands of raw IDs into one target class, the language-model V-sweep does not isolate geometry from target-frequency collision. Fixing support would remove varying remapping but leave added output rows without direct target supervision; a causal vocabulary-size study therefore requires a redesigned tokenizer/task, not merely a rerun of the same configs.

E Extra Controlled Experiments

E.1 Intervention Ablation: λ Sweep on Pythia-160M

Table 13 reports the full λ sweep for SSR applied head-only to Pythia-160M (5000 steps; SSR trained with randomized SVD; endpoint values from the archived checkpoint metrics). The baseline head skew is 38.87; SSR at any strength in $\lambda \in [0.01, 0.2]$ reduces it to a narrow plateau of 33.3–33.4. In absolute scales the plateau has nearly fixed σ_1: SSR raises σ_{10} from 16 to 19 while σ_1 stays near 640; on the constructed 410M collapse, where σ_1 is kept healthy, raising σ_{10} to 26 brings skew below τ. We treat 33.4 as a default-estimator empirical plateau rather than a derived equilibrium; it indicates limited head-only repair of this deeply collapsed head, with repair depth sensitive to estimator fidelity (Appendix A).

Orthogonal regularization reduces skew to 13.99 and increases eff_{90} from 2 to 432; SN reaches 14.26 and 404. These objectives flatten more of the spectrum than SSR at the tested settings. The spectral values alone do not establish either overcorrection or functional superiority.

Multi-seed validation of the $\lambda = 0.1$ condition on Pythia-160M (seeds 42, 43, 44, 45) yields final skew values of 33.38, 33.24, 33.41, and 33.47 respectively (mean 33.38 ± 0.10), confirming reproducibility of the default-estimator empirical plateau near 33.4.

Table 13: λ ablation on Pythia-160M (head-only, 5000 steps). Baseline skew at step 5000 = 38.87. SSR uses randomized SVD at $\tau = 10$; entries are archived checkpoint spectral metrics.

Condition	λ	σ_1/σ_{10}	eff_{90}
Baseline (no intervention)	0.0	38.87	2
SSR	0.01	33.34	2
SSR	0.05	33.36	2
SSR	0.10	33.38	2
SSR	0.20	33.38	2
Ortho regularization	—	13.99	432
Spectral normalization	—	14.26	404

E.2 Regularization-Strength Sweep on Pythia-410M

Table 14 reports a finite strength sweep on the initially low-skew 410M head (head-only, 5,000 steps). SSR endpoints remain near 8.2, whereas the tested ortho and SN settings yield ratios 1.5–1.9 and approximately 1.2. This demonstrates sensitivity within the sampled grids, not failure at every possible strength. Different penalty scales also make equal numerical weights an unequal intervention budget; these rows do not establish a matched functional Pareto frontier.

Table 14: Regularization-strength sweep on the initially low-skew 410M head (head-only, 5,000 steps). SSR holds skew near 8.2 across $\lambda \in [0.01, 1]$; ortho and SN produce lower ratios at the sampled strengths.

Condition	λ	σ_1/σ_{10}
SSR	0.01	8.20
SSR	0.5	8.16
SSR	1.0	8.16
Ortho	10^{-4}	1.86
Ortho	10^{-3}	1.51
SN	0.1	1.18
SN	0.5	1.20

E.3 Continuation after Orthogonal Regularization

Stage 1 applies orthogonal regularization with coefficient 10^{-3} for 5,000 steps. Stage 2 loads that head and continues for another 5,000 steps with gated SSR ($\lambda = 0.1$, $\tau = 10$, $T = 50$). Both initial stage-2 ratios are below the gate (1.51 and 2.28), so the endpoints in Table 15 are not repairs starting from the naturally skewed pretrained head. The matched task-only continuations below isolate the additional effect of SSR. The archived `train_loss` includes regularization: the 160M stage-1 value 647.84 contains a 644.87 penalty, leaving task loss 2.97, not evidence of exploding task loss.

Table 15: Sequential orthogonal regularization and gated-SSR continuation (seed 42; 5,000 steps per stage). Each second-stage run starts from the preceding first-stage head, not from the pretrained head. Task loss is the final logged minibatch objective minus its regularization term, not held-out loss. Different training durations and starting points preclude a direct method comparison.

Model	Completed stage	σ_1/σ_{10}	eff_{99}	Task loss
410M	1: ortho	1.51	1014	2.02
410M	2: gated SSR	9.98	986	1.79
160M	1: ortho	2.28	751	2.97
160M	2: gated SSR	4.24	750	2.40

E.4 Matched Continuations and Centered-Spectrum Audit

We first compare task-only and SSR continuations from the same stage-1 head for each model. Both use the preceding 5k-step protocol, with continuation seeds 42–44 and identical settings except for the SSR penalty. In this initial experiment, the stage-1 head is fixed at seed 42; variation across

continuations does not measure variation of the full two-stage pipeline. All twelve runs complete. The independent 410M replication below instead supplies the 410M entries in main Table 2. These experiments complement, rather than replace, the natural-head and constructed-skew studies.

Each saved head is evaluated in FP32 with its unchanged pretrained backbone and native tokenizer on the first 256 eligible records of each validation split: TinyStories and WikiText-2 raw. Blank records and records with fewer than two tokens are excluded; each remaining record is truncated to 512 tokens. NLL is weighted by the number of next-token targets after shifting. The same tokenized records are used for every condition within each model. This independent-record protocol differs from concatenated WikiText evaluation; its values must not be compared directly with Table 9.

Table 16: Initial fixed-stage1 continuation results, mean ± sample s.d. over three continuation seeds. These 410M rows are distinct from the independent full-pipeline results in main Table 2. $r_c = r(W_c)$ is centered skew; validation entries are token-weighted NLL (lower is better). Spectra are remeasured in float64 from saved heads.

Model	Method	Raw r	Centered r_c	TinyStories	WikiText-2
160M	Task-only	$4.249 \pm .013$	$1.876 \pm .002$	$2.1043 \pm .0014$	$9.1818 \pm .1355$
160M	SSR	$4.249 \pm .013$	$1.876 \pm .002$	$2.1043 \pm .0014$	$9.1818 \pm .1355$
410M	Task-only	$40.335 \pm .357$	$8.607 \pm .122$	$1.5547 \pm .0061$	$10.1663 \pm .2183$
410M	SSR	$9.992 \pm .008$	$2.386 \pm .112$	$1.5402 \pm .0051$	$8.6137 \pm .1098$

For 160M, final head hashes match within every pair, as do all evaluated record losses. The low final ratio is therefore not an additional SSR effect. For 410M, the paired SSR-minus-task-only NLL difference is -0.01444 ± 0.00129 on TinyStories and -1.55265 ± 0.10985 on WikiText-2 records. Paired-record bootstrap intervals, conditional on the three seeds, are $[-0.01873, -0.01058]$ and $[-1.61667, -1.49188]$ (2,000 resamples, seed 42). These resample common record indices across seed pairs, assume exchangeable records, and are descriptive rather than population guarantees; WikiText records may be dependent. The pretrained 410M WikiText-2 NLL is 3.2968 under this protocol, so SSR reduces the additional deterioration relative to task-only continuation but does not restore pretrained generalization.

Centering removes the softmax-null component $\mathbf{1}\bar{w}^\top$. In 410M, the centered component accounts for 20.8–21.1% of the squared paired weight difference: the intervention is not merely a change of common offset. This supports a functional effect of the intervention package, not a causal claim that the centered ratio alone mediates that effect. Float64 Gram spectra of the pretrained references agree with an independent QR/SVD computation to relative error below 3×10^{-13} for the top ten singular values. Per-record token hashes, losses, head hashes, and spectral decompositions are retained in the supplementary results.

Independent full-pipeline replication (410M). We retrain the 5k-step ortho stage1 independently with seeds 42–44, then branch each resulting head into task-only and SSR continuations of 5k steps. Each pair shares its initial head, seed, data protocol, and fresh optimizer; only SSR differs. All nine training jobs and the unified evaluation complete successfully, and the three stage1 head hashes differ. The independent units are three full-pipeline pairs, not nine runs. The evaluation protocol is unchanged from above.

For seeds 42, 43, and 44 respectively, paired SSR-minus-task-only validation NLL differences are -0.015494, -0.013097, and -0.015759 on TinyStories, and -1.620192, -1.615357, and -1.507944 on WikiText-2 records. Their means and sample standard deviations are -0.014783 ± 0.001467 and -1.581164 ± 0.063457. Task-only versus SSR NLL is 1.565093 ± 0.003066 versus 1.550309 ± 0.003670 on TinyStories, and 10.183844 ± 0.217585 versus 8.602679 ± 0.154154 on WikiText-2. Raw skew is 39.986835 ± 0.136101 versus 9.989190 ± 0.005915; centered skew is 8.460978 ± 0.012393 versus 2.350485 ± 0.129965.

The benefit thus replicates beyond the single first-stage head. It remains protocol-local: pretrained WikiText-2 NLL is 3.296793, and stage1 NLL is 7.662912 ± 0.004458, both below the SSR continuation. In contrast, TinyStories NLL improves from pretrained 1.863322 and stage1 1.669026 ± 0.006111. SSR mitigates the additional out-of-domain deterioration during continuation, but does not restore pretrained generalization. Three seeds do not establish population-wide

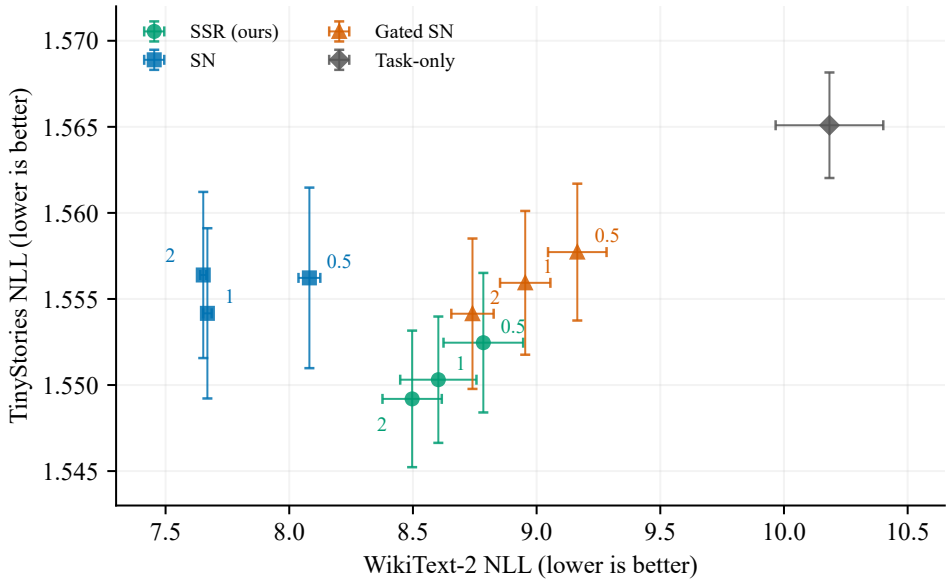


Figure 7: Domain adaptation and cross-domain retention in matched 410M continuations. Each point is a mean over three independent pipeline seeds; bars show marginal sample standard deviations, not confidence intervals. Labels indicate 0.5, 1, and 2 times the method-specific base coefficient (SSR: 0.1; SN and gated SN: 0.01). Lower is better on both axes. All tested settings and the task-only reference are shown, without interpolation or an optimized frontier. The validation records were previously inspected.

efficacy or isolate centered skew as a causal mediator. The initial fixed-stage1 experiment is reported separately, not pooled to inflate the independent sample count.

Matched SN comparisons. We add SN and gated SN continuations from each of the same three independent stage1 heads. Each receives 5k optimizer steps with the matched seed, data, and fresh optimizer used above. Both use the existing SN penalty strength 0.01, without selection on these validation results. SN applies its penalty every update. Gated SN uses $\tau = 10$ and $T = 50$, as in the gate ablation. All six new training runs and the unified evaluation complete. Re-evaluated task-only and SSR record losses match the preceding evaluation exactly. Equal optimizer steps do not imply equal wall-clock cost.

SN reaches raw skew 2.473237 ± 0.048303, centered skew 2.238843 ± 0.038728, TinyStories NLL 1.554164 ± 0.004945, and WikiText-2 record NLL 7.669052 ± 0.017222. Gated SN reaches 26.863828 ± 0.094580, 4.667696 ± 0.003482, 1.555937 ± 0.004176, and 8.953935 ± 0.101775, respectively. All uncertainties are sample standard deviations across the three independent stage1 seeds.

The paired SN-minus-SSR NLL differences are 0.003855 ± 0.001324 on TinyStories and -0.933627 ± 0.136975 on WikiText-2 records. For gated SN minus SSR, they are 0.005628 ± 0.000857 and 0.351255 ± 0.052382. Each ordering holds for every seed. Thus SSR gives the lowest in-domain NLL among these continuations, whereas ungated SN better preserves out-of-domain performance. Neither result establishes superiority after hyperparameter tuning. The comparison also does not isolate the two-endpoint gradient as the cause, since penalty strength, update direction, and gate activity differ. In particular, SN's lower ratio is not evidence of functional damage.

Sensitivity to regularization strength. Figure 7 displays the complete grid on the same two functional objectives. To test dependence on a single coefficient, we fix a symmetric grid of 0.5, 1, and 2 times each method's preceding coefficient before launching the additional runs. The base coefficients are 0.1 for SSR and 0.01 for SN and gated SN. We reuse the base-strength checkpoints and add 18 continuations: two strengths for each of three methods and three independent stage1 seeds. Initialization, data order within each seed, fresh optimizer, and 5,000-step budget are unchanged. All

runs and the unified evaluation complete. The first nine new runs use one A6000 and the remaining nine another after a GPU-availability interruption. Summed training wall time is 6.79 hours, excluding the interruption; unified evaluation takes 273 seconds. These timings include probes and I/O and do not isolate method overhead.

Evaluation follows the 256-record, 512-token protocol above. The validation records were already inspected in the base comparison: this is exploratory strength sensitivity, not untouched-test confirmation. We report the complete grid without selecting a winning configuration. Re-evaluating the 16 earlier checkpoints alongside the 18 new ones gives identical per-record loss sums; token hashes and target counts align across all 68 evaluation groups. Uncertainty is sample standard deviation over three pipeline seeds, not over the 17,408 head–record evaluations.

Table 17: Complete strength grid for matched 410M continuations. NLL is lower-is-better; spectral entries describe ordered singular values, not functional quality. Multipliers apply to each method's base coefficient. All entries are mean ± sample s.d. across three independent pipeline seeds. The task-only reference is shared with the preceding comparison.

Method	Mult.	TinyStories NLL	WikiText-2 NLL	Raw σ_{10}	Centered σ_{10}
Task-only	—	$1.56509 \pm .00307$	$10.18384 \pm .21758$	$1.935 \pm .006$	$1.896 \pm .003$
SSR	0.5	$1.55246 \pm .00405$	$8.78460 \pm .16103$	$4.616 \pm .104$	$2.645 \pm .072$
SSR	1	$1.55031 \pm .00367$	$8.60268 \pm .15415$	$4.353 \pm .147$	$2.631 \pm .021$
SSR	2	$1.54919 \pm .00397$	$8.49678 \pm .11986$	$4.224 \pm .062$	$2.587 \pm .048$
SN	0.5	$1.55623 \pm .00525$	$8.08142 \pm .04408$	$1.723 \pm .007$	$1.697 \pm .006$
SN	1	$1.55416 \pm .00495$	$7.66905 \pm .01722$	$1.735 \pm .004$	$1.708 \pm .004$
SN	2	$1.55639 \pm .00483$	$7.65236 \pm .01203$	$1.803 \pm .005$	$1.792 \pm .005$
Gated SN	0.5	$1.55772 \pm .00398$	$9.16448 \pm .11830$	$1.888 \pm .005$	$1.852 \pm .002$
Gated SN	1	$1.55594 \pm .00418$	$8.95393 \pm .10178$	$1.873 \pm .005$	$1.835 \pm .002$
Gated SN	2	$1.55414 \pm .00437$	$8.74074 \pm .08560$	$1.854 \pm .004$	$1.813 \pm .002$

Every SSR strength yields lower TinyStories NLL than every sampled SN or gated-SN strength in each matched seed, not only in the aggregate means. Conversely, every ungated SN strength yields lower WikiText-2 NLL than every SSR strength in each seed. The strongest gated SN also improves mean WikiText-2 NLL over the weakest SSR, so dominance over gated SN does not hold across all strength pairs. The base-strength paired contrasts reported above remain unchanged.

The spectral endpoints distinguish the interventions. SSR retains larger raw and centered tenth singular values throughout the grid (Table 17); the difference is therefore not solely a common-logit shift. SSR's mean raw ratio stays within 9.986–9.995, while its validation losses change with strength. SN's mean raw ratio falls from 9.424 to 2.473 to 1.145, but its in-domain loss is lowest at the middle sampled strength. Thus neither threshold attainment nor a lower ratio orders task performance. Larger ordered σ_{10} does not establish restoration of the same singular vectors, recovery of the entire tail, or causal mediation of the loss gains. This finite grid supports a protocol-local in-domain advantage, not a matched-loss frontier or optimality under unrestricted tuning.

E.5 Local Spectral Interventions and Joint Optimizer Updates

The continuation results establish an effect of the training intervention, but do not identify ratio reduction as its cause. We use two local controls to examine this distinction. The first changes only the head along prescribed spectral directions. The second combines task and regularizer gradients in a single AdamW update from a saved optimizer state. Neither experiment is a substitute for the matched training runs above.

Isolated directions. We first evaluate three independent 410M stage1 heads and the pretrained 160M and 410M references. The stage1 heads have raw ratios near 1.5, below the gate, so forcing an SSR step on them tests unnecessary intervention rather than active repair. We then apply the unchanged direction and step-size grid to the three saved task-only continuation endpoints, whose raw ratios are 39.9868 ± 0.1361. This second phase is a follow-up after checking the initial heads' spectral state; both phases are retained. The checkpoints are read-only and the backbone is frozen.

For each head, let A be either W or its centered version W_c, with singular triplets (u_i, s_i, v_i). The three directions are $-u_1 v_1^\top$, $u_{10} v_{10}^\top$, and $-\nabla r(A)$. Each is normalized to the same Frobenius perturbation budget $b \in \{0.1, 0.5\}\sigma_{10}(W)$ and added to the original head W. Thus the centered directions are additional diagnostic controls, not the implemented SSR objective. We include three centered random rank-one directions and a common-logit-shift control at each budget. Random directions are fixed by seeds 142–144 and are not additional model seeds. This design matches parameter-change magnitude, not ratio reduction or logit change. Singular values are re-sorted after the spectral perturbation; increasing the original tenth direction need not increase the new tenth singular value by the same amount.

Evaluation uses eligible validation records 257–384 from each of TinyStories and WikiText-2 raw, with the native tokenizer and truncation to 256 tokens. Blank and shorter-than-two-token records are excluded. All variants share the same hidden states and next-token targets. We record token-weighted NLL, token hashes, head hashes, and hidden-state projection energies. Low-rank logit updates agree with direct modified-head multiplication; common-shift controls change per-token NLL by less than 2×10^{-5}. These records and sequence lengths differ from the continuation table above, so absolute NLL values are not directly comparable.

Table 18: Isolated spectral steps on the three high-skew 410M task-only endpoints. Entries are changes in NLL relative to the unmodified head, mean ± sample s.d. across independent pipeline seeds. Negative is better. The budget is divided by the original raw σ_{10}. All six spectral directions and both budgets are reported.

Basis	Direction	Budget	TinyStories ΔNLL	WikiText-2 ΔNLL
Raw	Top decrease	0.1	$-.000157 \pm .000011$	$-.008786 \pm .000277$
Raw	Tenth increase	0.1	$-.000219 \pm .000019$	$+.000790 \pm .000214$
Raw	SSR	0.1	$-.000223 \pm .000020$	$+.000570 \pm .000214$
Raw	Top decrease	0.5	$-.000704 \pm .000055$	$-.043839 \pm .001384$
Raw	Tenth increase	0.5	$+.000952 \pm .000258$	$+.005080 \pm .000994$
Raw	SSR	0.5	$+.000935 \pm .000260$	$+.003980 \pm .000991$
Centered	Top decrease	0.1	$-.000585 \pm .000059$	$-.046325 \pm .001911$
Centered	Tenth increase	0.1	$-.000181 \pm .000038$	$+.003594 \pm .000454$
Centered	SSR	0.1	$-.000257 \pm .000031$	$-.001883 \pm .000672$
Centered	Top decrease	0.5	$-.001455 \pm .000263$	$-.229505 \pm .009505$
Centered	Tenth increase	0.5	$+.000666 \pm .000314$	$+.019880 \pm .002300$
Centered	SSR	0.5	$+.000278 \pm .000290$	$-.007556 \pm .003402$

Table 18 shows that a raw SSR step can reduce the ratio while improving TinyStories NLL and worsening WikiText-2 NLL. The larger step worsens both. At raw skew near 40, the tenth-direction coefficient in $-\nabla r$ is about 40 times the top-direction coefficient, which explains the close local agreement with the tenth-only intervention. Top-only contraction gives a more favorable domain trade-off here. Centering changes the direction and its functional outcome; it does not validate a centered gate or a new training method.

The random controls have much smaller effects in these high-skew heads. Across the three random directions, seed-mean NLL changes range from -1.36×10^{-5} to 2.58×10^{-5} on TinyStories and -7.30×10^{-5} to 2.32×10^{-5} on WikiText-2 at the smaller budget. At the larger budget, the corresponding ranges are $[-4.72\times10^{-5}, 1.64\times10^{-4}]$ and $[-3.19\times10^{-4}, 1.71\times10^{-4}]$. These ranges describe direction variability, not confidence intervals. Equal parameter norms do not equalize their effects on the hidden-state distribution.

The low-skew stage1 controls give a different result: raw SSR steps increase TinyStories NLL by 0.002031 ± 0.000088 and 0.026422 ± 0.000786 at the two budgets, while decreasing WikiText-2 NLL by 0.020216 ± 0.000458 and 0.080549 ± 0.002520. Both pretrained references also show domain- and direction-dependent effects. These forced, ungated perturbations cannot validate the gate as a classifier of functional harm. A 32-record pilot checks numerical correctness only and is excluded from all reported comparisons.

Joint steps from saved AdamW states. We recover the task-only 410M checkpoints at continuation step 3500 for seeds 42–44, including first and second moments and the optimizer counter. Their

raw ratios are 39.8824, 39.7655, and 40.0398. For each checkpoint we construct four fixed TinyStories training minibatches, each containing two microbatches of eight packed 256-token sequences. Each minibatch is independently forked into five single-step updates from the same weights and optimizer state: task-only, task plus implemented rSVD SSR, task plus exact SSR, and two task plus exact top-singular-value controls.

The SSR coefficient is 0.1, with $\tau = 10$, $p = 5$, and $q = 2$ for the implemented estimator. Step 3500 is an active check under the $T = 50$ schedule. Independent randomized gradients from the two microbatches are averaged as in the training implementation. The top-only controls use coefficient 0.01 or a coefficient chosen to match the Frobenius norm of the added rSVD-SSR gradient; they use the exact top singular direction, not the finite-power SN implementation. We preserve the saved head-group learning rate (1.25×10^{-5}), weight decay (0.1), AdamW moments, betas $(0.9, 0.95)$, epsilon 10^{-8}, and gradient clipping at norm 1. Each fork advances the counter to 3501 and is then discarded. The data iterator and random state were not stored, so this is an optimizer-state-matched counterfactual step, not a replay of the historical trajectory.

Validation uses eligible records 385–512 of the same two datasets, with truncation to 256 tokens. We recompute NLL from the per-record sums and token counts. The four minibatch contrasts are averaged within each seed before computing the mean and sample s.d. over the three independent seeds. Thus the 60 updates are repeated local probes, not 60 independent training runs.

Table 19: Joint AdamW steps from saved optimizer states. NLL contrasts are relative to the task-only update; ratio changes are relative to the common pre-update head. Entries are means ± sample s.d. across three seeds after averaging four minibatch forks within each seed. NLL columns are scaled by 10^{-5}.

Added term	TinyStories ΔNLL ($\times 10^{-5}$)	WikiText-2 ΔNLL ($\times 10^{-5}$)	Δr
None	0	0	$+.000154 \pm .000052$
rSVD SSR	$-1.512 \pm .504$	-23.800 ± 4.235	$-.007682 \pm .000408$
Exact SSR	$-3.262 \pm .354$	-70.198 ± 1.765	$-.210186 \pm .000988$
Top-only, norm-matched	$-8.379 \pm .941$	-181.886 ± 9.936	$-.016287 \pm .000091$
Top-only, coefficient .01	$-3.061 \pm .129$	-136.356 ± 9.786	$-.015851 \pm .000086$

All 12 implemented-SSR forks reduce raw skew, and their mean NLL contrasts favor SSR over task-only on both datasets (Table 19). The unpreconditioned projected margin $\lambda a - \mu$ is positive in these probes, ranging from 0.001746 to 0.071925, but the observed AdamW updates are the relevant check: the rSVD gradient norm is 8.14–15.12 times the exact norm, and combined-gradient clipping scales it by factors of 0.0315–0.0579. Exact SSR produces a much larger spectral change. Both top-only controls yield larger immediate task gains than implemented SSR, despite smaller ratio changes than exact SSR.

The joint and isolated probes establish different local outcomes. They also differ in checkpoint time, step magnitude, and optimizer history, so their contrast cannot isolate the cause of that difference. These small one-step effects do not establish long-run gains or statistical significance. They show that SSR can remain spectrally effective in a joint update with inherited optimizer history, while the magnitude of ratio reduction does not order functional benefit. Per-record outputs and source code are retained for both phases; the current public continuation supplement does not yet include these local probes.

E.6 Low-Skew Behavior of the Gated Head

Figure 4(a) shows ratio control on the initially low-skew Pythia-410M head. Over 5,000 steps, SSR ends at 8.13, compared with 10.35 without intervention and 3.34 with unconditional ortho. These spectral endpoints alone do not establish functional safety.

E.7 Additional Scale and Seed Checks

At 1B, joint fine-tuning (1k steps) ends at skew 9.57 without SSR and 8.30 with SSR, with similar observed losses (1.073 vs. 1.082; no equivalence test) and effective rank near 1986. Across seeds

the 410M joint endpoints are consistent (base 8.75/8.76, SSR 8.27/8.28 for seeds 43/44, compared with seed 42's 8.74/8.29). On a ViT-B head (high D/V), the final skew is 2.79 after 1k-step fine-tuning, below τ, and validation accuracy is 77.07% on 50,000 images. This endpoint is consistent with the intended healthy-head behavior, but alone does not establish that the gate remained inactive throughout training.

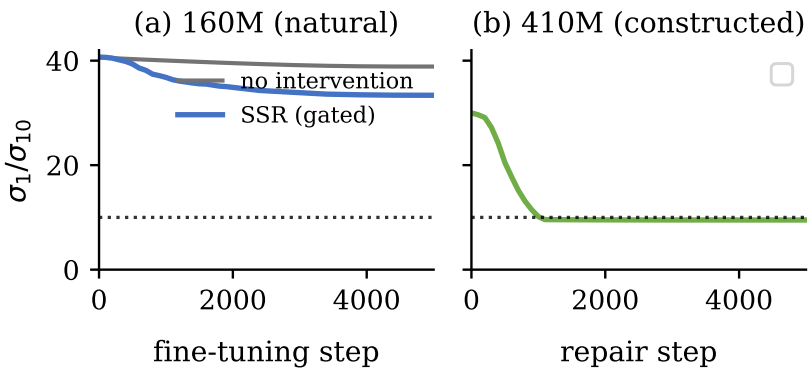


Figure 8: Spectral outcomes in two settings. (a) SSR reduces natural 160M skew to 33.38, below the no-intervention endpoint. (b) After a constructed perturbation, skew falls from 30 to 9.51 under SSR. These curves measure spectral change, not functional recovery.

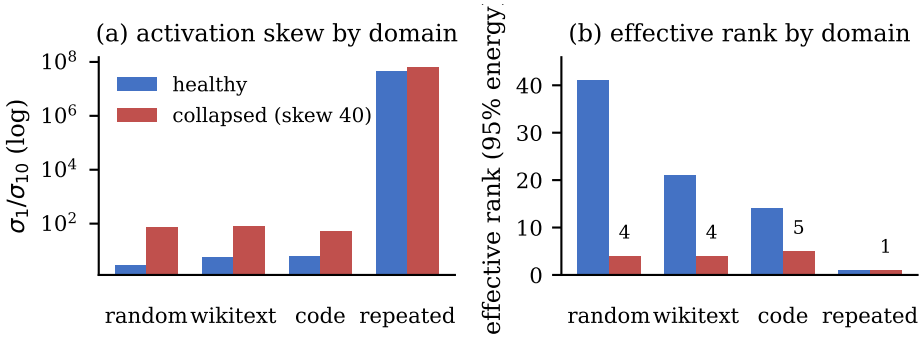


Figure 9: Input dependence of activation spectra in an internal attention head (L5H2), distinct from the LM-head weight diagnostic. With fixed perturbed weights (weight skew 40), activation skew ranges from 52 to 80 across random / Wikitext / code inputs (unperturbed 2.9–6.4, log scale), and effective-rank retention ranges from 36% (code) to 19% (Wikitext) to 10% (random). Repeated sequences are intrinsically degenerate in both conditions. The weight spectrum itself does not change with input.

E.8 Threshold Sensitivity: τ Sweep on Pythia-410M

The 160M head is deeply collapsed ($\sigma_1/\sigma_{10} \approx 33$) throughout head-only training, so every τ in $\{6, \ldots, 16\}$ leaves the gate open and produces identical training dynamics; τ sensitivity is not measurable far from the threshold (a methodological boundary, Section 5). On 410M, whose skew hovers around the threshold, the sweep discriminates (Figure 11):

The gate opens progressively as τ falls: at τ=15 the hinge never activates and the head drifts to the no-intervention baseline (10.35), while τ=8 and below hold the head near 6.8. τ=10 sits within the informative band $\tau \in [8, 15]$; this sweep measures sensitivity within the 410M protocol, not cross-model threshold validity.

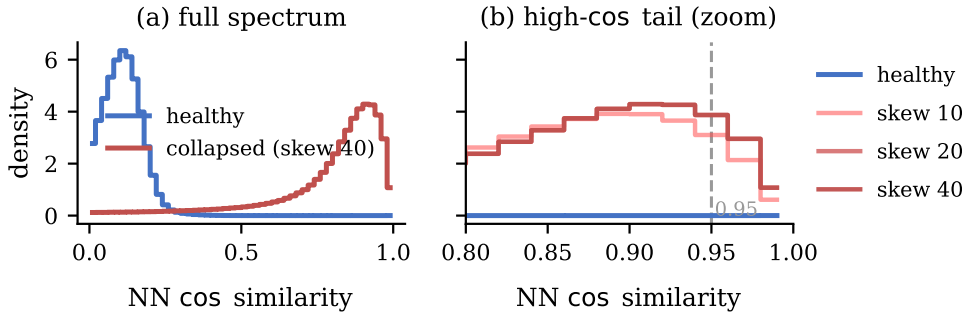


Figure 10: Pairwise head-output cosine distribution in the illustrated healthy (blue) and rank-collapsed (skew≥10, red) conditions. Healthy pairs cluster near orthogonality and only 0.5% of tokens have a cos >0.95 nearest-neighbour twin; in the collapsed condition, the bulk shifts to cos ≈0.9 and over 99% find such a twin. Holding the top singular value fixed does not fix σ_{10} or the ratio. The histogram provides complementary output-geometry evidence, not a demonstration that the ratio remained unchanged.

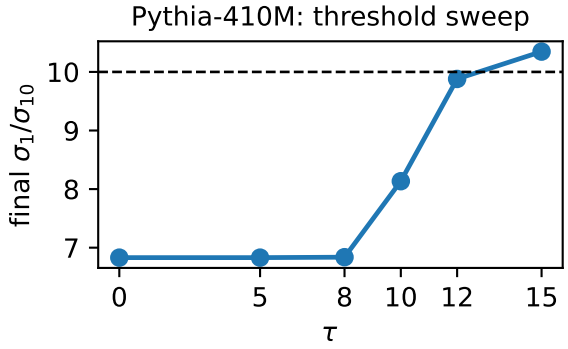


Figure 11: τ sweep on Pythia-410M: final skew rises monotonically with τ; $\tau \in [8, 15]$ is the informative band, and τ=10 sits within it.

E.9 Pythia-410M: Orthogonal Regularization and Constructed Spectral Repair

Ortho on an initially low-skew head. Table 21 shows three spectral endpoints: no intervention 10.35, SSR 8.13, and ortho 3.34. The functional concern about ortho comes from its worse HellaSwag and WikiText-2 results (Table 9), not from the lower ratio itself. SN reaches an even lower ratio with better downstream scores, ruling out a general inference that greater flattening is destructive.

Break-fix experiment. Table 22 tracks the constructed spectral perturbation and subsequent SSR training. Starting from the pretrained low-skew 410M head ($\sigma_1/\sigma_{10} = 7.37$ at step 0), collapse is constructed by SVD-scaling the spectrum to $\sigma_1/\sigma_{10} = 30.00$ (step 0 of the repair run reflects the constructed state). Head-only SSR at $\lambda = 0.1$ repairs the head to 9.51 over 5000 steps, crossing below $\tau = 10$. The repair trajectory is monotonic: skew declines from 30.00 to 29.64 (step 100), 29.12 (step 200), 27.18 (step 300), and eventually 9.51 (step 5000).

GPT-2 break-fix (not shown in full) follows the same pipeline: healthy skew 8.42 (construction baseline), constructed to 30.00, repaired to 11.17 after 5000 steps of head-only SSR.

Table 20: τ sweep on Pythia-410M (head-only, λ=0.1, 5000 steps). τ=15 never opens the gate (skew stays below τ); lower τ opens it progressively.

τ	σ_1/σ_{10}	eff_{90}
ungated (τ=0)	6.83	821
5	6.83	821
8	6.84	821
10	8.13	815
12	9.88	807
15	10.35	804

Table 21: Pythia-410M head-only interventions from an initially low-skew head. Spectral endpoints are compared here; downstream outcomes are in Table 9.

Condition	σ_1/σ_{10}	eff_{90}
Baseline	10.35	804
SSR (λ=0.1)	8.13	815
Ortho	3.34	787

E.10 Pythia-160M Layer-6 Attention Projection

The Pythia-160M layer-6 QKV attention projection was identified as the only backbone layer exceeding $\tau = 10$ ($\sigma_1/\sigma_{10} = 18.8$ in the main-text layer-wise probe). We apply SSR ($\lambda = 0.1$, $\tau = 10$) specifically to this projection weight (`ssr_layer_idx=6`) while leaving the head unregularized. The head spectrum is monitored throughout to inspect possible changes in head skew. Over 900 steps, the head skew remains essentially unchanged (40.69 to 39.97), showing no large head-skew change in this run; without a matched control this does not exclude indirect effects. The layer-6 QKV projection's own spectral trajectory is not captured in the head-level metrics.csv; the values reported in the main text ($18.8 \rightarrow 15.2$) are from direct spectral measurement at those projection weights.

E.11 TinyDecoder From-Scratch Prevention

Table 23 reports full-model TinyDecoder training ($L = 14$, $D = 512$, $V = 16384$, 30k steps, three seeds). Without SSR, terminal skew is $108.5/106.9/115.1$ and eff_{99} is 27–89. With λ=0.1, skew is $10.35/10.18/10.38$ and eff_{99} is 273–282. SSR therefore limits spectral concentration in all three seeds. Observed loss ranges overlap (1.32–1.44 versus 1.31–1.45); this is not a loss-equivalence test. A single weaker-penalty run (λ=0.01) ends at skew 16.4 and eff_{99}=148.

Preventive SSR's upper bound depends on what is trainable. Head-only from-scratch training at λ=0.1 diverges (skew 13–15 already at steps 200–600, loss NaN at step 700), but the same penalty is stable in full training, where the backbone absorbs the early spectral transient and the head is pinned at 10.3 (Table 23). Reactive-mode SSR is stable across the full $\lambda \in [0.01, 0.2]$ range on the collapsed 160M head, whereas from-scratch prevention tolerates weaker penalties head-only. The preventive operating window is therefore narrower than the reactive one, a concrete guidance difference beyond the τ=10 threshold.

E.12 Detection-Interval Sensitivity during From-Scratch Training

Two regimes of from-scratch collapse bound where diagnostic-triggered intervention applies. Pythia-style collapse is *gradual*: the spectrum degrades over hundreds of steps and periodic detection reliably triggers, whereas TinyNeoX trained from scratch collapses *fast*: skew reaches 13–15 within 200–800 steps under interval-100 detection. With every-step detection the early spike is suppressed (8.2 vs. 13.3 at step 200), and at step 30,000 the head has σ_1/σ_{10}=9.9997, close to τ, but effective rank falls to eff_{90}=4: dense detection contains the explosion yet cannot preserve rank. Fast from-scratch collapse is interval-sensitive.

Starting SSR at step zero can also limit emerging concentration: TinyDecoder from-scratch training at λ=0.01 holds the head at skew 16 against a baseline of 113.

Table 22: Pythia-410M break-fix experiment. Constructed collapse repaired below τ.

Stage	σ_1/σ_{10}	eff_{90}
Healthy (step 0)	7.37	824
Constructed (step 0)	30.00	6
After SSR (step 5000)	9.51	6

Table 23: TinyDecoder from-scratch prevention ($L = 14$, $D = 512$, $V = 16384$, 30k steps, full training). The no-SSR and $\lambda = 0.1$ rows contain three seeds; $\lambda = 0.01$ is one seed. SSR limits skew and raises terminal energy rank, with overlapping observed training-loss ranges.

Condition	σ_1/σ_{10}	eff_{99}	eff_{90}	loss
No SSR (baseline)	108.5/106.9/115.1	48/89/27	1	1.31–1.45
SSR λ=0.1	10.35/10.18/10.38	276/282/273	2–3	1.32–1.44
SSR λ=0.01	16.4	148	1	1.31

E.13 β_2: THE HIDDEN-DIMENSION EXPONENT

The two-factor model assigns hidden dimension D an exponent $\beta_2 = +0.569$. At fixed D/V=3.125% (V=32D) and L=6, skew is 28.35, 40.87, and 62.38 for D=256, 512, 1024, respectively (Figure 3). The log-log slope 0.57 agrees with the joint fit. This is variation at fixed ratio, not variation of D alone: V and its token-remapping rate also change (Appendix I.6).

F CROSS-SCALE FUNCTIONAL COMPARISON

We transfer the 410M base-strength protocol to Pythia-1B without retuning. For each seed in $\{42, 43, 44\}$, a separate 5,000-step orthogonal stage ($\lambda = 0.001$) supplies a head for four fresh-optimizer, 5,000-step branches: task-only, SSR ($\lambda = 0.1$), SN (0.01), and gated SN (0.01). Gates use $\tau = 10$ and interval 50. SSR uses $k = 10$, oversampling 5, and two power iterations. The direct-pretrained 1B row in Table 2(b) instead uses SSR $\lambda = 0.01$.

The backbone is frozen. Training uses the first 10K TinyStories examples, sequence length 256, effective batch 16, learning rate 5×10^{-5}, 500 warmup steps, AdamW betas $(0.9, 0.95)$, weight decay 0.1, and gradient clipping at 1.0. Runs use an RTX A6000 and float32. Evaluation uses the first 256 eligible validation records per corpus, truncated to 512 native-tokenizer tokens, with token-weighted NLL, not concatenated WikiText-2 perplexity. Sample indices, token hashes, and predicted-token counts agree across methods and seeds. Recomputed per-record loss sums reproduce all summaries. Independent units are seeds, not records.

Table 24: Pythia-1B per-seed validation NLL. Each seed has an independent first stage. Lower is better.

Corpus	Method	Seed 42	Seed 43	Seed 44
TinyStories	Task-only	1.457365	1.460863	1.459681
	SSR	1.447619	1.453574	1.451858
	SN	1.441654	1.449041	1.446703
	Gated SN	1.447251	1.453278	1.451054
WikiText-2	Task-only	9.478830	9.433585	9.147781
	SSR	8.464338	8.395587	8.273972
	SN	7.563434	7.569618	7.548811
	Gated SN	8.494439	8.465475	8.338989

SSR improves both NLLs over task-only in every seed, but SN improves both over SSR. Relative to gated SN, SSR achieves lower raw skew and better WikiText-2 NLL, yet slightly worse TinyStories NLL. Benefit over task-only transfers to 1B, whereas the in-domain ranking against SN does not. Three seeds in one family do not establish broad transfer or a globally tuned optimum. The comparison estimates intervention effects under this protocol, not causal mediation by ratio reduction.

Table 25: Pythia-1B spectra and paired NLL differences, mean ± sample s.d. over three seeds. Negative SSR-minus-comparator values favor SSR. These are descriptive summaries, not significance tests.

Method	Raw skew	Centered skew
Task-only	$42.2472 \pm .1757$	$8.9059 \pm .0674$
SSR	$9.9875 \pm .0069$	$2.4798 \pm .0352$
SN	$2.5272 \pm .0157$	$2.2415 \pm .0068$
Gated SN	$26.1499 \pm .1750$	$5.6008 \pm .0048$
Paired difference	**TinyStories NLL**	**WikiText-2 NLL**
SSR − task-only	$-.008286 \pm .001293$	$-.975433 \pm .088790$
SSR − SN	$.005218 \pm .000718$	$.817345 \pm .088188$
SSR − gated SN	$.000489 \pm .000275$	$-.055002 \pm .021702$

For completeness, 410M centered skew accompanying Table 2(d) is $8.461 \pm .012$ (task-only), $2.239 \pm .039$ (SN), $4.668 \pm .003$ (gated SN), and $2.350 \pm .130$ (SSR). The separate 160M control has identical paired weights and NLLs (TinyStories 2.1043, WikiText-2 9.1818).

F.1 Post-hoc Temperature Control

Different interventions can change confidence as well as prediction rankings. Following temperature scaling (Guo et al., 2017), we evaluate $\mathrm{softmax}(z/T)$ on the completed task-only, SSR, and SN heads at 410M and 1B, with three independent pipeline seeds per method. Positive scalar temperature preserves token rankings and, for these bias-free heads, is equivalent to scaling W uniformly, which leaves σ_1/σ_{10} unchanged. This control therefore tests a functional change that does not require changing the spectral ratio; it is not another training intervention.

Protocol. We fix the nine-point grid $T \in \{2^{j/2} : j = -4, \ldots, 4\}$ before evaluation. Each head and corpus selects its temperature by minimum token-weighted NLL on eligible validation documents 513–640. We then evaluate on documents 1–256, the same records used in the continuation comparisons. Both subsets use the native tokenizer, no added special tokens, and truncation to 512 tokens. They have disjoint document indices, but the evaluation subset was inspected in earlier experiments; this is an exploratory control, not a new untouched-test result. Per-corpus selection also assumes access to calibration data from that corpus and is not a zero-shot cross-domain deployment protocol. Gated SN and non-base training strengths are not included.

Checkpoint hashes match the preceding experiments. Evaluation token hashes and counts match all 9,216 corresponding document/head records. Recomputed grid losses agree with the saved summaries; raw per-document NLL differs from the earlier evaluator by at most 1.5×10^{-6} per token due to numerical evaluation differences. We aggregate over three training seeds, not over documents or temperatures.

Results. On WikiText-2 at 410M, task-only NLL changes from 10.1838 ± 0.2176 to 8.5481 ± 0.1063, SSR from 8.6027 ± 0.1542 to 7.8663 ± 0.0916, and SN remains at 7.6691 ± 0.0172. Selected temperatures are respectively 2, $\sqrt{2}$, and 1 in every seed. At 1B, task-only changes from 9.3534 ± 0.1795 to 8.3681 ± 0.1224, SSR from 8.3780 ± 0.0964 to 7.8152 ± 0.0643, and SN remains at 7.5606 ± 0.0107. Task-only and SSR select $\sqrt{2}$, and SN selects 1, in every seed. Uncertainty is sample standard deviation across seeds. The mean SSR-minus-SN gap narrows from 0.9336 to 0.1972 at 410M and from 0.8173 to 0.2546 at 1B, without a ranking reversal.

On TinyStories, all heads select $T = 1$ from this grid, leaving the reported NLLs and rankings unchanged. The grid is too coarse to rule out improvements near $T = 1$, particularly relative to the small SSR–SN differences. No selected temperature is a grid endpoint. These results show that confidence adjustment can substantially narrow the cross-domain NLL gap without altering skew. They neither exhaust calibration alternatives nor establish that the remaining differences are caused by particular spectral directions.

G Protocol Dependence of the Skew Relation

The two-factor relation uses lr 2×10^{-5}, batch 32, seed 42, and 30,000 steps on RTX 3090 with PyTorch 2.6; the $D{=}1024$ point exceeds 24 GB at batch 32 and was run on A6000. Table 26 holds batch and step count fixed while varying depth at two learning rates. The low-lr values are nearly flat; the high-lr values are higher and nonmonotonic. These observations establish protocol dependence but do not identify a layer-wise contraction or amplification mechanism.

Table 26: Skew under two optimization protocols (batch 32, seed 42, 30,000 steps; the 2×10^{-5} column is the matched-lr protocol of the two-factor relation). Skew is nearly flat across the sampled depths at low learning rate and higher but nonmonotonic at high learning rate; these observations do not identify a depth mechanism.

L	4	6	8
lr 2×10^{-5} (matched)	41.08	40.87	40.33
lr 6×10^{-4}	64.35	69.61	68.77

The depth range is limited to $L{=}4, 6, 8$ because $L{=}12$ and $L{=}16$ run out of memory at batch 32 on the RTX 3090 (24 GB) used for the matched-lr protocol. The earlier three-factor depth sweep of Section 2.2 (batch-16 protocol at 6×10^{-4}) is retained here as regime evidence: real but protocol-specific.

H Computational Cost and Timing Limitations

The SSR penalty adds a randomized SVD of the head ($V\times D$) every T steps. The dominant spectral work is $O(DVk)$ for subspace projection plus the small projected SVD, giving amortized cost $O(DVk/T)$. A closed gate removes the penalty gradient but not the scheduled spectral estimate used to decide whether it is closed. The retained production timings include scheduled dense-SVD probes, checkpoint serialization, data loading, and training, and the historical SSR runner also executed a dense debug SVD at intervention steps. They therefore cannot isolate randomized-SSR overhead, and we withdraw the earlier percentage estimate rather than infer one from incomparable logs. A controlled timing benchmark should hold sequence length, batch, probe schedule, checkpoint I/O, hardware, and estimator fidelity fixed. Fine-tuning and intervention runs use NVIDIA RTX A6000 GPUs (48 GB); the matched-lr runs of Section 2.2 use RTX 3090 (24 GB, PyTorch 2.6). Peak-memory and paired throughput measurements remain unreported limitations.

I Transfer Probes and Supplementary Interventions

I.1 Held-Out Pretrained-Family Probe

Table 27: Held-out transfer probe: six pretrained heads from Bloom, SmolLM, and OPT, none in Table 4. The two-factor relation overpredicts the absolute level on these held-out heads (pred 66–143 vs. observed 6.6–15.2, 4–21$\times$ above); the direction does not transfer reliably across families (Bloom-560M, the lowest-D/V head, has one of the lowest ratios). D is the head's own input dimension (D_{head}; e.g. OPT uses its word-embedding projection).

Model	D_{head}	D/V (%)	pred	obs	obs/pred
Bloom-560M	1024	0.41	142.6	6.63	0.047
SmolLM-135M	576	1.17	66.4	15.19	0.229
OPT-125M	768	1.53	70.1	11.03	0.157
SmolLM-360M	960	1.95	71.9	12.61	0.176
OPT-350M	512	1.02	65.8	7.70	0.117
OPT-1.3B	2048	4.07	81.5	10.43	0.128

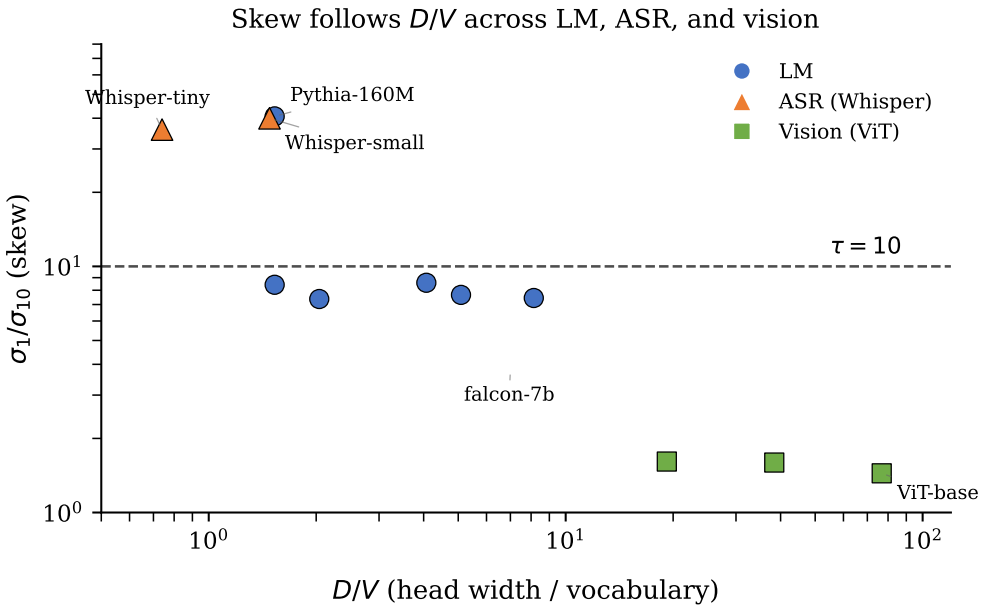


Figure 12: Head skew vs. D/V across LM, ASR, and vision heads. The measured low-ratio ASR heads have high skew, while the measured high-ratio ViT heads have low skew. This cross-domain association does not establish functional health.

I.2 Spectral Intervention and Recognition Error in Whisper

intervention	skew σ_1/σ_{10}	σ_{10}	eff_{90}	WER
baseline	40.1	2.92	125	**0.165**
ortho (λ=10^{-10})	34.5	2.96	205	0.170
SSR (λ=0.1, gated)	30.2	**3.84**	136	0.193
SN (λ=0.01)	21.9	2.98	**382**	0.264

Table 28: Whisper-small head, head-only fine-tuning on LibriSpeech train-clean-100 (3000 steps, lr 5×10^{-6}), WER on 500 test-clean utterances. The output head is untied without introducing a bias because the pretrained projection has none. Gated SSR grows σ_{10} but its WER is 0.193 versus 0.165 for baseline; SN more than triples eff_{90} ($125 \rightarrow 382$) while worsening WER to 0.264. Each condition is one seed; no equivalence or significance claim is made.

I.3 Prediction and Generation under Constructed Spectral Perturbations

We perturb the Pythia-410M head by SVD-rescaling to skew 40 at fixed Frobenius norm and evaluate the same frozen hidden states. This differs from the break-fix construction, which keeps σ_1 fixed. The frozen-state probe reports next-token accuracy falling from 0.34 to 0.034, entropy rising from 2.56 to 5.18, and confidence falling from 0.46 to 0.13. The intervention demonstrates a functional consequence of this particular spectral perturbation. Fixed weight norm does not fix the scale of Wh for a non-isotropic hidden-state distribution, however, and the construction does not establish that naturally occurring raw skew has the same consequence.

Table 29 reports generation under the original, perturbed, and separately SSR-fine-tuned heads. Severe perturbation reduces distinct-1 ($0.340 \rightarrow 0.038$) and increases self-BLEU ($0.044 \rightarrow 0.432$). The SSR-fine-tuned head has intermediate values but was not trained as a matched continuation from this perturbed head; these rows therefore do not estimate a causal repair fraction. High repetition despite a low ratio also shows why generation must be checked separately. The matched functional comparison is in Appendix E.4.

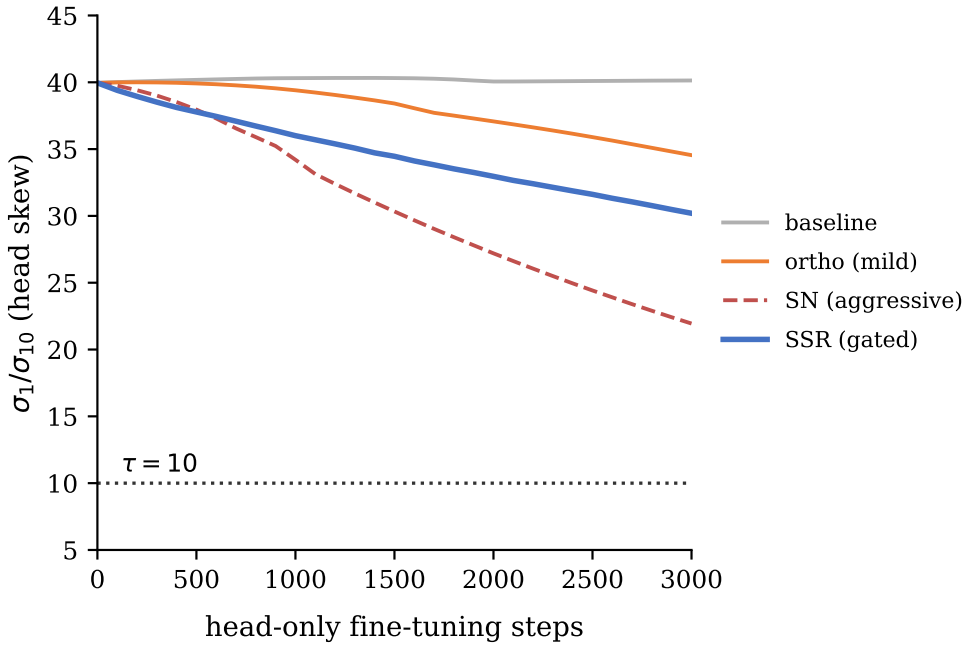


Figure 13: Whisper head-skew trajectories under head-only fine-tuning. Baseline drift leaves the head above threshold; gated SSR lowers skew by growing σ_{10}, with WER changing from 0.165 to 0.193; the stronger SN flattening reaches lower skew but WER 0.264. All comparisons are single-run.

Table 29: Generation under original, SVD-perturbed, and separately SSR-fine-tuned 410M heads; matched prefixes, frozen backbone, 10 sequences. The mild perturbation has lower skew than the original (7.1 versus 7.4), so these rows do not establish monotonicity in skew. The SSR row is not a matched repair of the severe perturbation. Entropy is measured on generated sequences, distinct from the frozen-hidden-state probe.

Head	skew	distinct-1	self-BLEU	bigram rep.	entropy
Original	7.4	0.340	0.044	0.265	3.16
Induced (mild)	7.1	0.314	0.040	0.372	2.67
Induced (severe)	40.0	0.038	0.432	0.740	3.27
SSR fine-tuned	8.1	0.122	0.158	0.708	1.23

I.4 Random-Label Class-Count Sweep on Frozen Vision Features

D	V	D/V	final skew σ_1/σ_{10}	eff_{90}
200	200 (1 sample/class)	100%	2.07	151
768	768	100%	1.72	390
768	2000	38.4%	3.68	537
768	4000	19.2%	6.29	590
768	8000	9.6%	**10.28**	608
768	12000	6.4%	13.46	603
768	16000	4.8%	15.86	593

Table 30: Random-label sweep on frozen ImageNet features; random-init head, head-only AdamW (5×10^{-5}), 5000 steps, batch 16. The primary ViT-B curve fixes D=768 and varies V, but class count, samples per class, and task difficulty co-vary with D/V. The D=V=200 one-sample-per-class run is a separate boundary control, not a point on the fixed-D curve and not a causal isolation of geometry from sparsity.

Dependence on the feature setting. On ViT-Tiny features (D=192, ImageNet-1K cache), the recorded training loss is 9.22 at V=8000, compared with chance loss $\ln V$=8.99. This condition

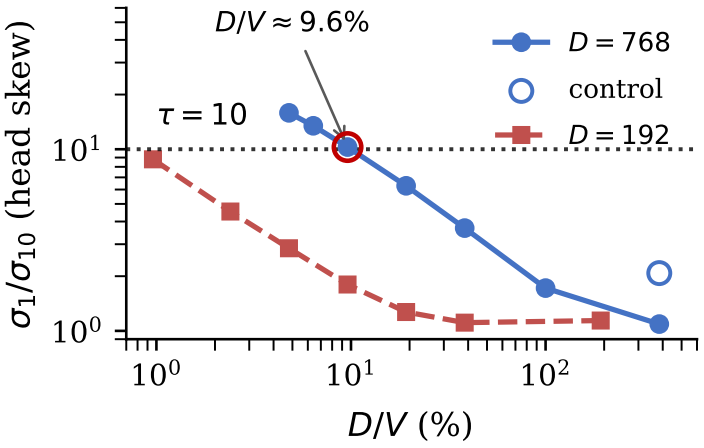


Figure 14: Random-label V-sweep at fixed D=768: skew rises as V increases and crosses τ=10 near D/V=9.6%. Separate D=200 one-sample-per-class and D=192 weak-feature runs illustrate learnability boundaries but do not isolate D/V causally.

is not matched in D/V to the ViT-B condition with skew 15.9, so their difference cannot isolate geometry. At V=20000, D=192, and N=50000 (2.5 samples per class), ViT-Tiny skew reaches 8.8. These observations motivate reporting the feature setting and fit quality alongside the random-label sweep; they do not establish a capacity bound.

I.5 Sensitivity to the Weak-Spectrum Anchor (k)

k	β_1	β_2	R^2	LOO R^2	healthy max	collapsed min
5	−0.415	+0.642	0.995	0.967	8.36	19.90
10	−0.409	+0.569	0.994	0.958	10.16	23.80
20	−0.361	+0.388	0.998	0.984	12.08	30.53

Table 31: Sensitivity to the weak-spectrum anchor (k). Exact full-SVD remeasurement of the seven controlled from-scratch configs (log-skew fit $\log r_k = \beta_0 + \beta_1 \log(D/V) + \beta_2 \log D$) and of the pretrained heads of Table 4; "healthy max" and "collapsed min" are the extreme measured skews within the healthy and collapsed groups at each k. This analysis remeasures saved checkpoints and refits a two-factor model, whereas Table 1 includes a depth term in its discovery fit. Their scores therefore differ in both measurement and model specification; the architectural coefficient signs agree.

Across $k \in \{5, 10, 20\}$, the controlled relation preserves the same coefficient signs, with $R^2 \geq 0.994$ and LOO $R^2 \geq 0.958$. Pretrained-head rankings are likewise stable (pairwise Spearman $\rho \geq 0.959$), and every k retains a non-empty separation between the measured healthy and collapsed groups.

Thus the measured trends are not specific to the tenth singular value. We use $k = 10$ as a fixed anchor, not a theoretically privileged index. This sensitivity check concerns checkpoint spectra, not gradient-estimation accuracy, which is evaluated in Appendix A. Thresholds require recalibration when k changes.

I.6 Held-Out Architecture Validation within the Matched Training Regime

The fixed-D/V sweep changes V together with D, so its preprocessing collision rate also changes. Consequently both fitted architectural coefficients can absorb token-remapping effects. Off-grid validation under the same preprocessing tests prediction within that generator and does not remove this confound. A clean architectural attribution requires a redesigned vocabulary/task control.

Table 32 reports an exact full-SVD remeasurement of six held-out off-grid architectures ($D \in \{384, 768\}$, $V \in \{8192, 16384, 32768\}$; seed 42, matched protocol), evaluated with the two-factor relation frozen at the discovery-stage coefficients ($\beta_0{=}-1.22$, $\beta_1{=}-0.409$, $\beta_2{=}+0.569$). No fitting is performed on these points; the frozen relation is applied unchanged.

D	V	D/V	predicted σ_1/σ_{10}	observed σ_1/σ_{10}	log residual
384	8192	4.6875%	30.495	27.696	-0.096
384	16384	2.34375%	40.490	36.680	-0.099
384	32768	1.1719%	53.762	51.808	-0.037
768	8192	9.3750%	34.072	30.579	-0.108
768	16384	4.6875%	45.239	42.336	-0.066
768	32768	2.34375%	60.067	59.004	-0.018

Table 32: Held-out validation of the frozen two-factor relation (exact full-SVD remeasurement, seed 42, matched protocol). Predicted values use the frozen coefficients $\beta_0{=}-1.22$, $\beta_1{=}-0.409$, $\beta_2{=}+0.569$; no refitting is performed on the held-out points.

The $2{\times}3$ off-grid grid deliberately pairs configurations with the same D/V at different absolute D: $(384, 8192)$ and $(768, 16384)$ both have $D/V{=}4.6875\%$, with observed skew rising $27.70 \to 42.34$; $(384, 16384)$ and $(768, 32768)$ both have $D/V{=}2.34375\%$, with observed skew rising $36.68 \to 59.00$. The absolute-D effect therefore reappears on previously unseen architectures at fixed D/V, as predicted by the frozen relation rather than by D/V alone.

Auxiliary reruns at two endpoint configurations were close for the two finite seeds (42 and 44): $D{=}384, V{=}8192$ gave 27.696 (s42) and 28.123 (s44); $D{=}768, V{=}32768$ gave 59.004 (s42) and 58.139 (s44). Seed 43 diverged to NaN at both configurations and is reported as an optimization failure; no value was imputed.

All six held-out points lie far above $\tau{=}10$, so this exercise validates magnitude and rank generalization of the frozen relation, not threshold classification. The threshold evidence continues to rest on the pretrained separation (Section 2.3), the τ-sweep (Appendix Figure 11), and the k-sensitivity analysis (Table 31).

J Related Work

J.1 Softmax Bottleneck and LM Head Expressivity

Yang et al. (2018) formalized the softmax bottleneck through a rank constraint and proposed mixture-of-softmaxes to increase expressivity. Godey & Artzi (2026) subsequently studied the LM head as a gradient bottleneck. These results motivate examining the output layer, but neither identifies our selected singular-value ratio as a cause of task degradation. We study its measured concentration and test explicit head interventions. Our matched head-only continuations evaluate output-layer adaptation, not improved gradient transmission to a trainable backbone.

J.2 Gradient Compression and Rank Collapse

Godey & Artzi (2026) reported an LM-head gradient bottleneck, with 95–99% compression of the backpropagated gradient norm in their measurements. Their 2B-parameter controlled experiments showed that smaller hidden dimensions produce worse convergence—a $16\times$ gap between $D = 512$ and $D = 4096$. However, their analysis focuses on aggregate compression ratios across the full vocabulary, leaving two questions open: whether compression is structurally uniform or asymmetric across SVD directions, and how model architecture parameters beyond D interact with the bottleneck. Our directional-energy correlation provides a descriptive cross-check, not a causal account of gradient compression. Our matched D/V sweep characterizes architectural associations while retaining its token-aliasing confound. Recent causal tests further suggest that geometric gradient compression alone does not establish the magnitude of its optimization harm (Murugan, 2026), reinforcing our focus on directly measured head geometry and controlled interventions rather than treating gradient compression itself as a sufficient causal explanation.

Dong et al. (2021) showed that pure attention loses rank doubly exponentially in depth, making rank collapse a known failure mode of attention stacks; we localize a distinct form in the final projection head, where we measure protocol-dependent associations with head geometry. Closest to our layer-wise probe, Nait Saada et al. (2025) identify rank collapse in width of the attention matrix, a spectral gap between its two largest singular values that widens with context length, alongside the standard rank collapse in depth; they analyze signal propagation with random-matrix theory and mitigate collapse by removing the outlier eigenvalue. We instead target the final head's weight spectrum, quantify its $D/V \times D$ dependence, and turn the diagnosis into a conditional intervention. Kulkarni et al. (2026) systematically varied hyperparameters across 108 OLMo models and found that effective rank is primarily a reflection of training choices (batch size, weight decay) rather than a causal driver of performance. Their work establishes that low rank does not imply saturation: a critical boundary condition for any study claiming collapse is harmful. Both studies examine training-dependent geometry. We vary D/V and D within a fixed protocol, then test an explicit ratio-control objective and matched functional outcomes; this adds an intervention study rather than a general causal account of rank (Appendix G).

End-to-end signal-propagation theory provides a broader comparison. Kedia et al. (2024) derive forward- and backward-moment recursions for Transformers and use them to diagnose rank collapse and instability throughout the network. Their remedy changes initialization and residual scaling globally; our narrower question concerns the trained LM head, a measured gate, and whether a local head intervention should activate at all.

J.3 Representation Quality and Downstream Behavior

Several recent works have characterized output-representation geometry. Wu & Papyan (2024) documented neural collapse in LLM output representations, showing that token class means approach a simplex equiangular tight frame as model scale increases. That structured class-separation geometry is not evidence of pathological LM-head rank collapse. Our contribution instead studies concentration in the head's weight spectrum and tests which architectural variables predict it under a matched training protocol.

J.4 Spectral Learning Dynamics

Our SVD-based analysis is motivated by theoretical work on spectral learning dynamics. Saxe et al. (2014) showed that, for deep linear networks under specific assumptions, individual singular modes can evolve at different rates and weak modes can exhibit long learning delays. This supplies a qualitative precedent for measuring modes separately, but it does not derive our D/V relation: the algebraic rank ceiling when $V > D$ does not determine the relative ratio σ_1/σ_{10}. Jacot et al. (2018) further showed that, in the infinite-width kernel regime, convergence rates depend on kernel eigendirections. Our correlation statistic ρ is only a measured alignment diagnostic for the LM-head update; we do not identify it with an NTK eigendecomposition.

Hu et al. (2022) showed that low-rank parameterizations can capture effective adaptation updates in large language models. This concerns the intrinsic dimension of a constrained update across model layers, whereas we measure the spectrum of the pretrained LM-head weight itself. The rank of the current head does not by itself constrain the rank of its unconstrained gradient, so our observations should not be read as a mechanism for LoRA. Papyan et al. (2020) documented neural collapse (NC), the convergence of class means and classifiers to a simplex equiangular tight frame, during the terminal phase of classification training. We note that our spectral collapse is fundamentally distinct from NC: NC is a *structured* collapse to a maximal-separation simplex (rank at most the number of classes minus one), while ours is an *unstructured* concentration toward a dominant direction. This distinction arises from the different objects and training regimes being measured.

Zangrando et al. (2024) analyze low-rank bias in nonlinear networks, showing that weight decay and small within-class variability drive neural rank collapse toward low-rank solutions during optimization. We localize this phenomenon in the LM output head rather than the hidden layers, quantify its dependence on D/V and D, and couple the diagnostic with a conditional intervention instead of relying on implicit regularization alone.

J.5 Output-Embedding Geometry and Spectrum Control

A separate line of work shapes or probes the output embedding's spectrum. Gao et al. (2019) identified the representation degeneration problem: pretrained output embeddings become anisotropic and occupy a narrow cone. Most directly, Wang et al. (2020) reparameterize the output embedding as $W = U\Sigma V^\top$ and guide Σ toward a slow-decaying singular-value prior. Their method is therefore the closest prior spectrum-control objective, not merely background: it shapes the spectrum unconditionally, whereas SSR optimizes a diagnostic ratio only when a calibrated gate opens. We include a matched LM-head adaptation of Spectrum Control. It produces low terminal skew, but a zero-regularization factorized control explains most of the short-pilot spectral change; we therefore treat it as a direct applicability comparison with a parameterization confound, not as clean attribution to the prior penalty. Cancedda (2024) further show that tail directions of embedding and unembedding spectra carry functionally specific signals, reinforcing why changing weak singular directions requires behavioral evaluation rather than a spectral-only success criterion. Li et al. (2025) trace representation geometry across pretraining and find low-rank, anisotropic phases associated with the $d \ll |V|$ regime; they characterize representation trajectories, whereas we measure the final head and use its ratio as an intervention trigger. Across these comparisons, SSR does not guarantee task preservation: its empirical contribution is selective spectral control, whose functional effect is evaluated separately by task and domain.

J.6 Positioning

Existing work has established that substantial gradient compression is observed across model scales (Godey & Artzi, 2026), while rank alone does not predict performance (Kulkarni et al., 2026). What is missing is a predictive understanding of *what determines collapse severity*. We provide a matched-regime two-factor relation of collapse severity: a multiplicative association with the head-to-vocabulary ratio D/V and the absolute dimension D (R^2=0.9945, §2), with depth effects assessed separately under two optimization protocols (Appendix G). Because the language-model vocabulary sweep also changes token-aliasing frequency, this relation is predictive within the tested generator rather than a causal identification of geometry. The two geometric predictors cannot be tuned independently: increasing D moves D/V for fixed V, so width and ratio trade off against each other.