THE OPTIMIZER IS AN UNCONTROLLED VARIABLE IN SPARSE AUTOENCODER EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoder (SAE) reconstruction quality is routinely compared across differently trained language models. We show it depends on the optimizer that trained the model, and that the sign of this dependence reverses with the depth at which the SAE is attached. Holding all else fixed, we train 13.9M to 210M-parameter transformers on two corpora under Euclidean (SGD), sign-like (Adam) and spectral (Muon) geometries, with per-geometry learning-rate sweeps and each comparison read separately at matched validation loss and matched training compute (3 to 8 seeds, seed-paired intervals). At the final residual stream and fixed sparsity (TopK, $L_0 = 32$), SAEs fitted to the spectral arm reconstruct worse in every run, by 0.03 to 0.07 explained variance, without attenuation to 210M and surviving four times the training; varying only the update norm reproduces it. Refitting the same SAE at every depth reverses the sign: mid-stack, the spectral arm is reconstructed better, by up to 0.087. The model with the best validation loss is reconstructed worst in every run; identical JumpReLU settings yield a $4.08\times$ spread in achieved sparsity. Cross-model SAE comparisons should report the base optimizer and rate, attachment depth and achieved sparsity.

1 INTRODUCTION

Sparse autoencoders (SAEs) have become a widely used instrument for reading features out of a language model's residual stream (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025), and their reconstruction quality (explained variance at a given sparsity) is used as a property of the *model* being read. It is compared across models to argue that pruning makes activations easier to reconstruct (Borobia et al., 2026), that a compressed model keeps its original's features (Gupte et al., 2025), or that one training recipe yields better features than another (Li et al., 2026). Every such comparison assumes that, with the SAE held fixed, a difference in reconstruction is a difference in the representation. This paper tests one variable that sits between the two and is not controlled in any comparison we could find: the optimizer that trained the base model.

Optimizers differ in the norm their update is steepest under (Bernstein & Newhouse, 2024): SGD descends under the Euclidean norm, Adam's per-coordinate normalization makes its step sign-like, and Muon (Jordan et al., 2024) orthogonalizes each matrix update, flattening its singular values towards one. Whether these *geometries* leave different representations behind at equal loss is a question about the model; whether a dictionary fitted afterwards can tell is a question about the instrument. We ask both, under a protocol that controls what such comparisons usually leave free: every arm trains the same architecture on the same data with its own learning-rate sweep, every comparison is read at matched validation loss and again at matched training compute, and every difference must clear an absolute floor as well as a seed-paired interval.

Findings. Across four transformers (13.9M to 210M parameters) on two corpora, three *arms* per model (one per geometry) and up to eight seeds per arm:

1. **The optimizer changes SAE reconstruction quality at the final residual stream.** There, SAEs fitted to the spectral arm explain 0.029–0.057 less variance at matched loss and 0.041–0.072 less at matched compute, in every one of six runs, without attenuation to 210M (§3). Varying only the update norm, in one code path with everything else shared, reproduces it; so does an exact orthogonalizer in place of Muon's iteration. Loss mismatch,

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.

SAE hyperparameters, activation scale and spectrum, weight decay, batch size, activation budget, five SAE architectures and training length are each excluded as the explanation by a dedicated measurement.

2. **The sign depends on where the SAE is attached.** Refitting the identical SAE at every residual site reverses it: mid-stack, the spectral arm is reconstructed better, by up to 0.087, at all four scales (§4). The geometry changes the depth *profile* of reconstructability, not its level: the layer chooses the answer's sign. The reversal has a mechanism: the Euclidean arm's first block overwrites 43% of the embedding where the others overwrite 4–6%, and the final-site deficit does not.
3. **Two consequences for practice.** In every run the arm with the best validation loss is the one reconstructed worst, so ranking models by reconstruction inverts ranking them by the objective they were trained on. And a shared SAE hyperparameter is not a shared setting: identical JumpReLU settings leave the arms at 32.83, 12.35 and 8.05 active latents. This $4.08\times$ spread traces to the units of the sparsity penalty and collapses to $1.07\times$ once the inputs are put on one scale (§5).
4. **Seed and geometry act on different things.** The base model's seed dominates the basis an SAE recovers (two seeds leave dictionaries 0.009 above the random-vector level) but not which features fire; geometry dominates how well they reconstruct. Any claim that two models "find different features" has to be read against a baseline in which two seeds already produce unrelated bases.

What we do not claim. That any optimizer is better: neither wins at every depth. That the effect is specific to SAEs: linear probes find the spectral deficit larger in raw activations than in SAE features, so the dependence belongs to the representation (a confounding variable, not an instrument artifact) and the dictionary is where practitioners meet it. Anything above 210M or near convergence: our scale runs train at 1–5 tokens per parameter, and extending training moves the final-site gap up at one scale and down by half at another, removing it at neither. And a mechanism for the final-site deficit: the one we proposed, our own, was refuted by measurement (§6).

What follows. Section 2 states the protocol, §3 and §4 the effect and its reversal, and §5 the consequences, ending with what a cross-model SAE comparison should report. Every number regenerates from a stored result by the released code; four claims of our own that did not survive re-measurement are reported beside the ones that did.

2 Experimental design

2.1 Three update geometries, one of everything else

Each base model is trained three times from the same initialization on the same data order, under one of three update rules that differ in the norm their step is steepest under (Bernstein & Newhouse, 2024): **Euclidean** (SGD with heavy-ball momentum 0.9), **sign-like** (Adam (Kingma & Ba, 2015), $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$; its accumulation-free form, sign descent, is steepest under a max-of-max norm, hence sign-like rather than Euclidean), and **spectral** (a Muon-style rule (Jordan et al., 2024): update singular values flattened towards one by a quintic Newton–Schulz iteration, momentum 0.9, Nesterov). Headline runs use weight decay 0, so decay does not act on the activation scale we later examine; switching it on leaves the effect unchanged (§3). The rate warms up over the first 5% of steps, then decays by cosine.

The geometry is applied to the MATRIX parameters (attention q, k, v, output projection, both MLP matrices). AUX parameters (embeddings, LayerNorm gains, biases) receive an identical rule at an identical rate in every arm, Adam at 3×10^{-4}, never swept. That makes the Euclidean and spectral arms *hybrids*, with 4–35% of parameters under Adam, and only the sign-like arm homogeneous end to end; tying the auxiliary rule to each arm's geometry would confound the manipulation with embedding learning, and the effect it produces is larger, not smaller (Appendix R). Because the manipulated fraction matters, the vocabulary is capped by training-split frequency ($8{,}192$ tokens on TinyStories at a 0.118% UNK rate), which raises it from 35.4% to 76.5% at our smallest configuration (Appendix A).

As named optimizers these three differ in more than the norm: momentum semantics, adaptivity and per-shape scaling move together, so we also carry a second instantiation of each class in which they do not: one code path whose only branch is the norm's sharp operator (identity, sign, or orthogonalization), with momentum, Nesterov, the shape factor and the step size shared by construction (Appendix J); and a third for the spectral class, exact $UV^\top$ from the SVD, since one orthogonalizer cannot separate a geometry from an optimizer's coefficients.

2.2 Models, corpora and budgets

Base models are decoder-only pre-LayerNorm transformers, context 256, batch 32, bfloat16 (Table 2, Appendix A). Four *scale* runs span 13.9M (6 layers, width 384) to 210M (16 layers, width $1{,}024$) on TinyStories (Li & Eldan, 2024), with the second corpus, WikiText-103 (Merity et al., 2017), read at 60.5M because the vocabulary rule fails at 13.9M there and we enlarged the model rather than relax the rule; they train at 4.71, 1.46, 1.08 and 1.44 tokens per parameter. Two *budget* runs hold the architecture fixed and raise training: 13.9M to 20.04 tokens per parameter, and 60.5M to 4.71, the 13.9M run's own budget. Only $13.9\text{M} \to 44.8\text{M}$ and $44.8\text{M} \to 210\text{M}$ isolate scale, the second at matched tokens per parameter; both also change depth, width and head count, which is why we fit no trend.

2.3 The readout: one SAE recipe, refitted

Every headline number is the explained variance (EV) of a **TopK** sparse autoencoder (Gao et al., 2025) with $k = 32$ active latents and an $8\times$ overcomplete dictionary, fitted for $3{,}000$ Adam steps at 3×10^{-4} on $163{,}840$ residual-stream activations drawn from frozen validation batches (the identical token windows for every arm) and evaluated on a held-out slice, with the SAE seed pinned. Only the activations differ between cells. The recipe is refitted, never re-tuned, at every site and scale; four further architectures (BatchTopK, Matryoshka, Gated, JumpReLU) are fitted the same way where stated (Bussmann et al., 2024; 2025; Rajamanoharan et al., 2024; 2025). A comparison is made only on admitted sites: if any arm's dictionary leaves more than 2% of its latents dead at a site, the site is excluded rather than read, because a comparison between a live dictionary and a degenerate one measures the degeneracy. EV is a proxy metric (Karvonen et al., 2025), and §3 tests whether the deficit reaches what the model computes.

2.4 Learning rate is not a nuisance parameter

Every cross-geometry comparison uses per-geometry learning-rate sweeps; no comparison is made at a shared rate (Choi et al., 2020). Optimal rates differ by orders of magnitude and move along the axes we vary: selected rates differ across arms by $398\times$ to $1778\times$. Each arm's sweep runs five to seven log-spaced points at half duration and one seed, with an audit that fails any optimum at a grid edge; the final stage retrains three rates per arm: the arm's own best plus the two whose loss lies nearest the common target, at full duration and 3–8 seeds; where that grid brackets an arm's best from below only (the sign-like arm at three scales), the next rate up was also run at full duration, and the optimum held (Appendix A).

2.5 Two readouts, because "matched" is a choice

The arms do not reach equal loss (the worst-to-best loss ratio, or parity, is 1.07–1.09 across the scale runs), so any comparison must choose what to hold fixed; we report every headline under both choices. Matched loss reads each arm at the trained rate whose final validation loss lies nearest the worst arm's best, using converged endpoints only: an unrestricted scan over (rate, step) selects mid-training checkpoints that hit the target exactly, substituting "not trained yet" for "trained worse" (in the toy model, before we imposed the restriction, only 2.2% of loss-matched selections sat at the final step). Matched compute reads each arm at its own best rate for identical steps; spectral's orthogonalization costs 1.14–$1.27\times$ the FLOPs, worth under 0.005 of a gap (Appendix AK). Neither is correct alone: the first handicaps arms, the second confounds geometry with achieved loss, and where they disagree we say so.

Table 1: SAE explained variance per arm at the final residual stream (the matched-loss cells), and the Euclidean−spectral difference under both readouts. Four *scale* runs above the rule, two *budget* runs below it. Every difference clears the 0.0100 floor with a seed-paired interval off zero; the intervals are in Appendix H. The 13.9M row is $n = 8$; the others are $n = 3$ and their intervals are read as ranges.

			SAE explained variance			**ΔEV** (Eucl. − spectral)	
Corpus	**Params**	**tok/par**	Eucl.	sign	spectral	**matched loss**	**matched compute**
TinyStories	13.9M	4.71	0.8196	0.8209	0.7896	+0.0300	+0.0414
TinyStories	44.8M	1.46	0.8204	0.8157	0.7743	+0.0461	+0.0692
WikiText-103	60.5M	1.08	0.8360	0.8217	0.7788	+0.0573	+0.0663
TinyStories	210M	1.44	0.7919	0.7874	0.7375	+0.0544	+0.0722
TinyStories	13.9M	**20.04**	0.8092	0.7993	0.7671	**+0.0422**	+0.0414
WikiText-103	60.5M	**4.71**	0.7809	0.7707	0.7523	**+0.0286**	+0.0475

2.6 Significance is not sufficient

Seeds collapse onto near-identical solutions (seed spread in effective rank is $\approx 0.03\%$), so an interval that excludes zero is easy to obtain and says little. Every reported difference must therefore also clear an absolute floor of 1% of the metric's bounded range (0.0100 in EV) built into the bootstrap routine so no call site can omit it; a *verdict* is whether a comparison clears the floor with its interval off zero, and Appendix AI derives each floor and shows that every spectral verdict and the whole depth profile survive 0.5% and 2% as well. Intervals resample base-model seed pairs, one draw carrying every arm's value at that seed, since the arms share initialization and data order; at $n = 3$ such an interval cannot exceed the observed seed range, so every $\pm$ from three seeds is marked † and read as a range. The floor filters effect size, not false-positive rate, so the second is answered separately: a Holm–Bonferroni correction over every floor-clearing comparison in the paper ($m = 75$, on a paired t, since a percentile interval at $n = 3$ cannot widen to any corrected level) changes five verdicts, all at $n = 3$ (two Euclidean−sign-like cells we never claimed and three sites of the 210M depth profile, which then rests on layer 7 against 13–16) and no headline: every comparison involving the spectral arm survives, the smallest $|t|$ among the $n = 8$ claims is 11.4 against a Bonferroni critical value of 5.8, and the headline gaps sit at 28–190 standard errors. The floor is the stricter filter in the other direction, removing five comparisons Holm would retain, all with $|t|$ between 7 and 58 and effects of 0.0003 to 0.0092: seeds collapsing, p vanishing, nothing of consequence found (Appendix C). Every retraction in this work came from the analysis code that turns numbers into verdicts, so that layer carries *guards* (assertions that refuse to report a comparison the data cannot support) each tested against the retraction that motivated it (Appendix D).

3 At the final residual stream, the spectral arm reconstructs worse

Table 1 is the paper's central measurement. At the final residual stream, our primary site, the spectral arm reconstructs worse in every run, under both readouts, on both corpora: $+0.0300$, $+0.0461$, $+0.0573$, $+0.0544$ at matched loss across the four scale runs, and more at matched compute. An effect this size might wash out as models grow; it does not (the two largest runs agree within the floor), and since four points are no scaling law we claim only that the gap does not shrink with scale. Of the 24 pairwise comparisons in the table, 18 clear the floor and every one involving the spectral arm does; Euclidean and sign-like are separated in few cells and we build nothing on their order (Appendix H). The matched-compute gap exceeds the matched-loss gap in all four scale runs (matched compute handicaps no arm and gives a larger one) while our pre-set parity gate (that the arms' losses agree within 1.5%) is missed by all four, so the readout reduces the arms' loss disagreement rather than removing it (Appendix E).

It is not residual loss mismatch. Matched loss leaves a residual, and one might suspect the deficit is the spectral arm read slightly off its optimum. Three corrections say otherwise: projecting each arm's EV onto the common loss along its own EV-versus-loss curve shrinks the gap by 3–11%; at

equal loss the spectral arm's curve sits 0.038–0.052 below the Euclidean one, so loss explains 11–24% of the difference; and reading each arm at exactly the target loss by interpolation between its two bracketing cells, which fits nothing, gives $+0.0283$ $[+0.0273, +0.0294]$ against the selected-cell $+0.0300$. That last comparison is the strongest of the three: 40 new spectral cells were trained so that the two arms' loss frontiers overlap, and inside the overlap the gap is flat. Across a 0.055-nat sweep of matched loss it moves by 0.0007, every one of 25 targets clears the floor, and the same loss reached from either side of spectral's optimum, $95\times$ apart in learning rate, changes it by at most 0.0031. Put the arms at identical losses, seed by seed, and the gap does not move (Appendices E–G).

It is the norm's. The isolated instantiation of §2, in which the three arms share one code path and differ only in the sharp operator applied to the update, reproduces the deficit at full strength: $+0.0365$ $[+0.0353, +0.0376]$ at matched loss and $+0.0532$ at matched compute, with the spectral arm worse in 18 of 18 rate pairs, and the arms selecting the same rate so that no learning-rate difference remains to carry it. Replacing Muon's quintic with exact $UV^\top$ from the SVD gives $+0.0323$ and $+0.0432$ (Appendices J, K). The deficit belongs to the class of orthogonalizing updates, on the two instantiations we have, at the floor's resolution and not finer. It is also graded in how much of the network the norm acts on. Handing a fraction $f \in \{0, 0.25, 0.5, 0.75, 1\}$ of the 36 matrices, by parameter count, to the spectral norm lowers EV at every step. Which matrices are handed over is determined per seed, the rest are under Adam, and each dose is re-swept and read at the published matched-loss target. Every step's interval is off zero, $+0.0291$ $[+0.0277, +0.0304]$ end to end and close to linear at about 0.008 per quarter of the parameters ($R^2 = 0.79$). No single step clears the floor. The effect is the accumulation. Dosing the Euclidean geometry the same way does nothing (0.8204, 0.8225, 0.8196 at $f = 0$, 0.5, 1; Appendix S).

Training length moves it in both directions and removes it at neither. Tokens per parameter fall (4.71, 1.46, 1.08) exactly as the effect rises across the first three runs, so "grows with scale" and "grows with undertraining" are indistinguishable there. The two budget rows separate them. At 13.9M, raising the budget to 20.04 tokens per parameter (near compute-optimal (Hoffmann et al., 2022), still one epoch) enlarges the matched-loss gap, $+0.0300 \rightarrow +0.0422$, with matched compute flat at $+0.0414$. At 60.5M, raising it $4.4\times$ to 4.71 halves it, $+0.0573 \rightarrow +0.0286$ $[+0.0231, +0.0335]$, matched compute $+0.0663 \rightarrow +0.0475$, both still clearing the floor on every seed. The published 60.5M figure is therefore inflated about twofold by undertraining, the 13.9M figure is not, and undertraining explains the effect at neither scale (Appendix L).

Six more things it is not. Each was measured rather than argued, at 13.9M and matched loss unless stated. Weight decay switched on at a strength that halves the weight norm, at matched compute and three seeds†: $+0.0424$ $[+0.0414, +0.0434]$ against $+0.0414$ undecayed. Batch size $4\times$: $+0.0373$. A $100\times$ activation pool read in a single pass at 0.79 epochs, at matched compute: $+0.0405$ against $+0.0430$ at $1\times$. SAE hyperparameters: the arms' activation scales differ by $11.65\times$, yet tuned per arm all three select the same SAE rate, and across 8 configurations of expansion, k, rate and per-arm standardization the gap clears the floor in 8 of 8, the most under-tuned SAE reading $+0.0174$. SAE architecture: BatchTopK, Matryoshka and Gated dictionaries reproduce it at 0.96–$1.11\times$ the TopK gap under both readouts, and JumpReLU at $0.90\times$ once its sparsity penalty is on one scale across arms (§5.2). The auxiliary parameters: giving them each arm's own geometry enlarges the effect (Appendices M, N, O, P, Q, R).

What carries it, and what does not. The activation spread across the arms tracks the effect closely, and we had proposed it as the mechanism. It is not. A Gaussian surrogate carrying each arm's exact covariance reproduces the effective-rank (Roy & Vetterli, 2007) difference but only 18% of the final-site gap, inside the floor; PCA at the SAE's own reconstruction level sees 20%. The deficit at this site is therefore sparse-specific and not a second-order property of the activations, a negative result against our own account (Appendix T). What we can say is where it sits: stratified by training-set token frequency, the gap runs from $+0.0310$ on the rarest decile to $+0.0115$ on the commonest ($\rho = -0.90$; supported at both readouts), and extra dictionary capacity closes it $6.4\times$ faster on rare tokens than on common ones (Appendix U), consistent with, and no evidence for, the reports that Muon learns tail classes better than Adam (Wang et al., 2026) and Adam better than gradient descent (Kunstner et al., 2024).

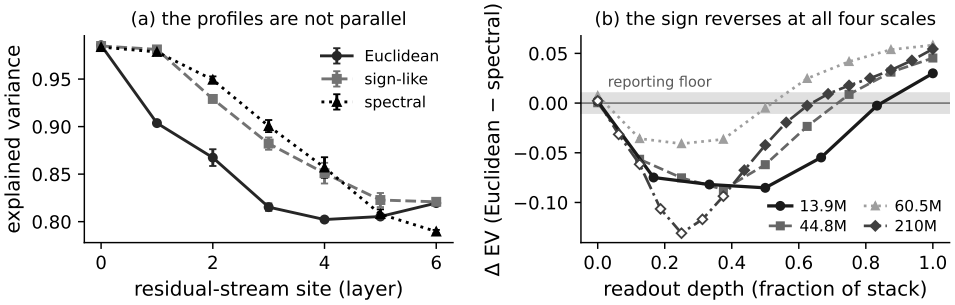

Figure 1: **(a)** At 13.9M, the same SAE refitted at every residual site (8 seeds; bars are seed-paired intervals). The Euclidean arm's explained variance falls steeply at the first block and then flattens; the spectral arm's descends gradually and keeps descending, so their difference changes sign once. **(b)** The Euclidean−spectral difference against readout depth at all four scales, floor-clearing negative in the middle of every stack and positive at its end. Open markers are sites excluded by the dead-latent rule (§2); the shaded band is the reporting floor. Values, intervals and dead fractions per site are in Appendices X, Y and Z.

It reaches what the model computes, and it is not the SAE's fault. Substituting each SAE's reconstruction into the forward pass raises the model's cross-entropy by $+0.0180$ $[+0.0167, +0.0193]$ nats more for the spectral arm than for the Euclidean one at 13.9M and by $+0.0240$ $[+0.0223, +0.0256]$ at 44.8M, KL agreeing; the conventional *loss-recovered* score, normalized by a mean-ablation of 12–13 nats, compresses every arm to ≈ 0.988 and does not resolve it. ΔCE tracks EV at $r = -0.905$, so this extends the deficit to behavior rather than confirming it independently (Appendix V). Linear probes find the spectral deficit larger in raw activations than in SAE features on both tasks we tried, so a better dictionary cannot remove it: the dependence belongs to the representation (Appendix W).

4 The sign depends on where the SAE is attached

Everything in §3 was read at one site. Refitting the identical SAE at every residual-stream site tests whether that choice was load-bearing, and it was (Figure 1). At 13.9M the Euclidean−spectral gap is floor-clearing and negative at layers 1–4 (-0.0749, -0.0819, -0.0853, -0.0547), inside the floor at layer 5, and positive only at layer 6, where it reproduces Table 1's $+0.0300$ exactly. Layer 0, the embedding, is the pre-registered control: it receives an identical rule in every arm, and the arms are indistinguishable there ($+0.0019$). Repeating the sweep at 44.8M and 60.5M (8 seeds, 9 sites, 432 fits) reverses the sign in both, each reproducing its published final-site gap under a pre-set gate, and the largest effect on an admitted site in the paper is mid-stack: $\mathbf{-0.0870}$ at layer 3 of the 44.8M run. At 210M, with 17 sites, it reverses a fourth time (negative at layers 7–9, positive from 12) with the six earliest sites negative but excluded by our dead-latent rule (Appendices Y, Z). Dead latents run against this result rather than producing it: the spectral arm carries more of them at exactly the sites where it reconstructs better, so its advantage there is understated.

Geometry changes the depth profile of reconstructability, not its level. The two arms' profiles are not parallel (Figure 1a), which is why their difference changes sign rather than merely shrinking. Attach the SAE at the final stream and the spectral model is harder to reconstruct; attach it mid-stack and it is easier, by more. No choice of layer escapes the dependence, because the layer chooses the sign: reconstruction quality compared across models at one depth is a statement about that depth.

The reversal has a mechanism; the final-site deficit does not. The Euclidean arm's first block overwrites 43% of the embedding where the other arms overwrite 4–6%, *embedding dilution*. One shared quadratic in that fraction fits every arm's explained variance at layers 1–4 with sub-floor residuals, and the final-site deficit survives at matched dilution, two-thirds of it (Appendix X). The two sites are dissociated on every mechanism axis we have: at layer 3 the covariance-matched surrogate that carried 18% of the final-site gap carries the whole reversal, PCA sees 73% of it against 20%, and the token-frequency profile is flat where the final-site gap concentrates on rare tokens. The

reversal is second-order and largely linear; the final-site deficit is neither (Appendix AC). Consistently, the isolated instantiation that reproduces the deficit at full strength (§3) does not re-establish the floor-clearing reversal: its profile crosses in place with the negative half inside the floor, for a reason we cannot yet give (Appendix J).

Where the crossing sits is measured, and unexplained. The gap's interpolated zero sits at 0.85, 0.72, 0.52 and $\mathbf{0.64}$ of depth across the four scales. A log-linear fit to the first three, registered before the fourth run, predicted 0.35 at 210M and put the reversal's vanishing near 1B; the measured 0.64 misses it by more than a block. The crossing moves away from the final layer with scale but along no law we can state; extending training at 13.9M moves it $+0.02$ (Appendices Y, AB).

Both readouts agree on the signs; they disagree on what can be measured. Reading the depth sweep at each arm's own best rate keeps every site's sign and enlarges most gaps, at 60.5M layer 3 moves from -0.0365 to -0.0658, and each final site reproduces Table 1's matched-compute column. But the spectral arm's best rate is $10\times$ its matched one, and there its mid-stack activations stop being fittable: dead latents reach 0.84 where the Euclidean arm, at an identical rate under both readouts, stays near zero, and our 2% rule removes six of nine sites at both larger scales and every mid-stack site at 210M. The floor-clearing reversal is thus established at matched loss at all four scales and at matched compute at 13.9M at its published budget; elsewhere the mid-stack is not measurable at this SAE budget, which is not the same as absent (Appendix AA).

Training and architecture. At 13.9M with $4.3\times$ the training the reversal holds and flattens: layer 3 halves to -0.0450, the final site grows to $+0.0422$, the crossing moves from 0.85 to 0.87, and a Matryoshka dictionary reproduces TopK's profile site for site ($0.99\times$, $1.05\times$ on admitted sites; Appendix AB). At 60.5M with $4.4\times$ the training, sites 1–7 are excluded under both readouts: the sign-like and spectral arms' mid-stack streams acquire a dominant shared direction (expected cosine between tokens 0.94–0.99 in the spectral arm against ≤ 0.26 in the Euclidean one) while every arm is near-isotropic again at the final site, so the reversal cannot be re-read there for want of admitted sites, not for want of sign (Appendix L).

5 What this means for cross-model SAE comparisons

5.1 The model you would pick is the model you would reject

A gap of 0.03–0.07 explained variance invites the reply that the effect is real but inconsequential; three measurements, one per subsection below, say what it costs. At the final residual stream, the arm reaching the lowest validation loss produces the worst SAE reconstruction in all four scale runs: a loss advantage of 0.075–0.234 nats against an EV deficit of 0.041–0.072, every spectral comparison floor-clearing, both orderings read at matched compute on the same three checkpoints (Appendix I), and again in both budget runs. What replicates in every run is the spectral arm's own movement, first by loss and last by EV; Euclidean and sign-like are separated in 2 of 8 cells, so the full three-way ordering is resolved in one run only. The inversion belongs to the site: mid-stack the ordering runs the other way (§4). We do not claim that practitioners select models this way, only that two characterizations of the same models disagree systematically, unflagged.

5.2 A shared hyperparameter is not a shared setting

The effect reaches into the instrument. JumpReLU's swept knob is the target initial L_0, with the threshold θ calibrated to each arm's own pre-activation quantile, so the arms start matched in sparsity and do not stay so. With identical settings, achieved L_0 is 32.83, 12.35 and 8.05, a $4.08\times$ spread, holding at 3.91–$4.26\times$ across all seven settings. JumpReLU's Euclidean$-$spectral gap read at the sparsity window's midpoint is $2.4\times$ TopK's. The cause is units. The arms' activations differ in scale by $11.65\times$; the calibration removed that difference from the threshold but not from the objective, where the L_0 penalty weighs $136\times$ more in the spectral arm. Scaling the inputs before fitting collapses the spread to $\mathbf{1.07}\times$ (21.8, 21.6, 23.1) and the gap to $+0.0298$ $[+0.0294, +0.0302]$, $0.90\times$ TopK on the same activations, whether the threshold is then calibrated per arm or shared (Appendix Q.2). Fixing an SAE hyperparameter across models silently compares dictionaries at

different sparsity, the error the field adopted TopK to avoid, unless the objective's units are shared too. the number to report is the achieved sparsity (Chanin & Garriga-Alonso, 2026).

5.3 SEED MOVES THE BASIS; GEOMETRY MOVES HOW WELL

Any claim that two models "find different features" must be read against how much two SAEs disagree with nothing changed. Fitted to identical activations with different SAE seeds, two dictionaries match at a median maximum cosine of 0.4256, with 66% of features unmatched at cosine 0.5, the floor every comparison below is read against, and a reproduction at our scale of an instability established at larger ones (Paulo & Belrose, 2026; Gerasimov et al., 2026). Changing the base model's seed leaves dictionaries $+0.0090$ above the random-vector level of 0.1780, below our floor; changing the update geometry moves them detectably (0.31–0.36) but far less than seed does. Yet matched by which features fire on frozen tokens rather than by decoder direction, seed and geometry perturb the feature set equally (0.4187 and 0.4180 against the SAE-seed level of 0.4539), and the $27\times$ by which two SAE seeds exceed two base seeds above chance in decoder cosine becomes $1.09\times$: seed rotates coordinates and leaves function. On reconstruction the order reverses: across 24 base models per geometry the seed-to-seed spread in EV is at most 0.0023, and the geometry gap is 18–$36\times$ that. Our geometry claim is measured at matched seed and stands; our seed number may not travel: McCann (2026) report 0.91–0.93 for the cross-seed comparison on Pythia-70m; matched bijectively as they do, ours is 0.1790, and the one explanation we could test, undertraining, moves it by $+0.0025$ at $4.3\times$ the training (Appendix AE).

5.4 WHAT A CROSS-MODEL SAE COMPARISON SHOULD REPORT

Each finding above names a quantity that current comparisons leave implicit. A comparison of SAE reconstruction across base models should state **(i)** the optimizer that trained each model and its learning rate, and whether the rate was chosen per model; **(ii)** the attachment depth, and the profile across depths where a conclusion's sign could depend on it; **(iii)** the achieved sparsity and dead-latent fraction per model; **(iv)** the readout: matched loss or matched compute, with the loss parity, or both; and **(v)** an absolute floor and a stability baseline (the SAE-seed and base-seed levels) against which a difference is read. Of three published cross-model reconstruction comparisons (Borobia et al., 2026; Gupte et al., 2025; Li et al., 2026), all three report the readout depth, none the base optimizer or its rate, and all fix the SAE hyperparameters across models (Appendix AG); this is no criticism of those papers, since the variable had not been shown to matter.

6 LIMITATIONS

Scale, corpora and training regime. Our largest model is 210M parameters, all decoder-only transformers of one family, on two corpora, at 1–5 tokens per parameter for the scale runs. Extending training moved the final-site gap in opposite directions at the two ends (§3); 44.8M and 210M remain undertrained, and the scale-budget interaction is measured at two points and understood at none. The residual streams we read are, at the published budgets, easier dictionary targets than production ones: near-isotropic, high-rank, without a GPT-2 stream's massive activations (Appendix A), and at $4.4\times$ the 60.5M budget two arms' mid-stacks leave that regime and stop being fittable (§4). Whether the final-site deficit survives the anisotropic streams of models trained at scale is untested here; we claim nothing about frontier scale or converged models.

Hybrid arms, and two instantiations of a class. Every arm but the sign-like one gives 4–35% of its parameters to Adam. The effect grows monotonically with the fraction manipulated, with a Euclidean control flat (§3), and is larger at 100% (Appendix R): the concession is bounded, not removed. "Class" rests on two instantiations of each geometry, both at 13.9M, agreeing at our floor and not finer; two rules are not the space of them.

The crossing is unexplained, and the mid-stack is often unmeasurable. What sets the depth at which the sign changes is unknown (§4), and our registered extrapolation of it failed out of sample. At matched compute, and at 60.5M with $4.4\times$ the training under both readouts, the 2% dead-latent rule removes most mid-stack sites; the raw gaps there keep the reversal's sign, but we draw nothing from a degenerate dictionary, and a recipe that first subtracts a stream's dominant direction is untried.

No mechanism for the headline. The reversal is embedding dilution; for the final-site deficit we have only where it sits and a refuted account of our own (§3). Pre-registered predictions that missed are reported as misses where they arose (Appendices Z, AB, AC, L).

Two tests that did not pass, and the thinnest claim. We predicted a task-level ranking flip between raw and SAE feature spaces; it did not, so the consequence we demonstrate is metric-level. A pre-registered test of whether a toy model of superposition (Elhage et al., 2022) predicts the transformer failed, the one resolving readout carrying the opposite sign (Appendix AJ). The claim that the deficit is not SAE-specific rests on linear probes at one scale whose differences are not shown to clear a floor; we rely on their direction, not their size (Appendix W), and the functional readout, which tracks EV at $r = -0.905$, is a second metric family, not a second line of evidence.

Statistical scope. Table 1 and everything derived from it rest on $n = 3$ at the three larger runs, so those intervals are observed ranges; every interval is seed-only, and four measured components of variation exceed the printed one, the largest 0.0063, none threatening a headline gap (Appendix AH); the cross-seed dictionary conflict of §5 is unexplained.

7 Related work

Optimizer geometry and representations. Optimizers select different solutions at similar loss (Wilson et al., 2017; Gunasekar et al., 2018); Bernstein & Newhouse (2024) frame their accumulation-free forms as steepest descent under a choice of norm, the taxonomy our arms follow, Muon (Jordan et al., 2024) being the spectral instance. Implicit-bias analyses now reach Adam and Muon in restricted settings (Wang et al., 2021; Fan et al., 2025; Gronich & Vardi, 2026; Vasudeva et al., 2025), and Muon has a lineage in spectral descent and a theory of its own (Carlson et al., 2015b; Bernstein & Newhouse, 2025; Chen et al., 2026, Appendix J). The closest Muon-specific work, Ruan et al. (2026), reports that spectrally normalized updates yield features more robust to corruption and better at transfer, supported by higher effective rank; we differ only on what effective rank licenses (§3, Appendix T). Dragutinović & Ranganath (2026) show, in deep linear networks and linear attention, that optimizers differ in their simplicity bias, our premise.

Sparse autoencoders and their evaluation. SAE research has concentrated on architectures and sparsity mechanisms (Bricken et al., 2023; Huben et al., 2024; Gao et al., 2025; Rajamanoharan et al., 2025; 2024; Bussmann et al., 2024; 2025), on frontier-scale releases (Templeton et al., 2024; Lieberum et al., 2024) and on benchmarks for them (Karvonen et al., 2025; Chanin, 2026); cross-model comparisons of reconstruction are made (Borobia et al., 2026; Gupte et al., 2025; Li et al., 2026), and none of the three we audited controls for how the compared models were trained (Appendix AG); cross-model SAE work asks instead whether feature spaces are shared across models (Lan et al., 2025; Thasarathan et al., 2025; Nasiri-Sarvi et al., 2026; Chen et al., 2025; Ognev et al., 2026). Dictionary instability is not our finding (Paulo & Belrose, 2026; Gerasimov et al., 2026; Fel et al., 2025; Brzozowski & Chung, 2026; Leask et al., 2025); our same-activation floor reproduces it. Representations have long been compared across trained networks by matched units (Li et al., 2015), CCA variants (Raghu et al., 2017; Morcos et al., 2018), CKA (Kornblith et al., 2019) and stitching (Lenc & Vedaldi, 2019; Bansal et al., 2021; Csiszárik et al., 2021); the measures encode different invariances and disagree on whether two initializations learn the same representation (Mehrer et al., 2020; Ding et al., 2021), hence our two criteria (Appendix AE). On the base model's seed, McCann (2026) find one function in rotated bases; our reading, our numbers disagreeing (Appendix AE). None varies the optimizer.

8 Conclusion

The optimizer that trained a language model is a variable in how well a sparse autoencoder reconstructs it, and the sign of that variable depends on where the autoencoder is attached. None of this says which optimizer to use; it says that a cross-model comparison of SAE reconstruction is a comparison of representations only once the base optimizer and its rate, the attachment depth and the achieved sparsity are on the table.

AI USE STATEMENT

Generative AI was used in this work, in four ways. We used it to aid and polish writing: a large language model, directed by the authors, drafted and edited the prose of this paper, and the authors revised it. We used it for retrieval and discovery: to search for related work, summarize what it found, and assemble the bibliography, every entry of which was then checked against its primary record before it was added. We used it to draft sections of the paper: the main text and the appendices were written from the stored results under the authors' direction, section by section, against an evidence map the authors fixed in advance. And we used it for purposes: to translate methods developed and specified by the authors into code, and to correct errors in code; the resulting code and corrections are verified rather than asserted. We have reviewed all AI-assisted work. We have not used generative AI to generate or fabricate data. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work involves no human subjects, annotators or personally collected data, and required no ethical review. It uses two public corpora: TinyStories, a synthetic corpus of machine-generated children's stories released under the CDLA-Sharing-1.0 licence, and WikiText-103, drawn from Wikipedia's Good and Featured articles under a Creative Commons Attribution-ShareAlike licence. The supplementary material includes the preprocessing scripts that rebuild our tokenized copies from those public sources and records the tokenizer, the vocabulary cap and the measured unknown-token rate for each; the tokenized copies themselves are released after decision notifications, under those licences and with the source, licence and processing record for each corpus. WikiText-103 carries whatever incidental information about people its source articles contain; we add none and extract none. The models we train are small (13.9M to 210M parameters) and are trained on those corpora only; their weights are too large to accompany the submission and are released after decision notifications, so that every measurement here can be recomputed from the models themselves rather than only from our records of them. They have no capabilities that raise dual-use or deployment concerns, and the compute they consumed is reported in full in Appendix AK. One consideration deserves a direct statement. The paper reports that a widely used interpretability metric depends on a training choice that current practice does not control for, a claim about the reliability of tools used in safety-adjacent research. Because an overstated reliability concern is as unhelpful as an unstated one, we have been explicit about what the evidence does not support: that the effect is specific to sparse autoencoders, or that any optimizer is preferable (§6); we report our own retracted claims rather than omitting them; and the audit of published comparisons (§5, Appendix AG) records only what each paper's text states, names the candidates it excluded, and claims no rate. The role of generative AI in this work is detailed in the AI use statement above.

REPRODUCIBILITY STATEMENT

Every number, table and figure in this paper is computed from released result files by released scripts, and every experiment is fixed by a configuration and a random seed. The main text states what is needed to re-run the study: the control surface, which parameters receive the manipulated update rule and which are held identical across arms (§2.1); the models, corpora, vocabulary rule and token budgets (§2); the sparse-autoencoder recipe, identical in every arm (§2.3); the per-geometry learning-rate protocol and the two readouts, matched loss and matched compute (§2.5); and the absolute-effect floors and seed-paired intervals under which a difference is reported (§2.6). Appendix A records every configuration, learning-rate grid, selected rate and seed count; Appendix B the readout in full; Appendix C the protocol as a checklist, with the consequence of omitting each control; Appendix D the assertions the analysis code enforces; Appendices AH and AI the statistical conventions; and Appendix AK the compute by stage and hardware. The supplementary material contains the source code (the three update rules, the model, the autoencoders, and the sweep and analysis library), every experiment script with its stages, the preprocessing scripts that rebuild the corpora from their public sources, the pinned environments, the test suite (which verifies bit-level agreement of the Euclidean and sign-like arms with reference optimizer implementations and of the spectral arm's orthogonalization with an explicit SVD, and checks each study's verdict against the rule registered for it) and, in an enclosed archive, every result file: the pre-registered rules and the verdicts computed under them, the per-cell records, the per-step training record and configuration of

every trained cell, and the run logs. From those records alone the released scripts reproduce every figure in this paper and re-derive the verdict files. What the submission cannot carry for reasons of size: the trained model and sparse-autoencoder weights, the tokenized corpora, and the columnar copies of the training records, is released after decision notifications; the configuration, seed and training record of every cell are included here, and the released scripts retrain it.

REFERENCES

Essential AI, :, Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of muon for pretraining, 2025. URL https://arxiv.org/abs/2505.02222.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. URL https://openreview.net/forum?id=HJ4-rAVtl.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Lukas Balles and Philipp Hennig. Dissecting adam: The sign, magnitude and variance of stochastic gradients. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 404–413. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/balles18a.html.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 225–236. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/01ded4259d101feb739b06c399e9cd9c-Paper.pdf.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=ux18f5nOpD.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=hErdffTsLu.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signSGD: Compressed optimisation for non-convex problems. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 560–569. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/bernstein18a.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Å. Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971. ISSN 00361429. URL http://www.jstor.org/stable/2949484.

Hector Borobia, Elies Seguí-Mas, and Guillermina Tormo-Carbó. How pruning reshapes features: Sparse autoencoder analysis of weight-pruned language models, 2026. URL https://arxiv.org/abs/2603.25325.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Michał Brzozowski and Neo Christopher Chung. Ablating archetypes: The stability of archetypal saes is an artifact of initialization and metric design, 2026. URL https://arxiv.org/abs/2606.02061.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. In *NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning*, 2024. URL https://openreview.net/forum?id=d4dpOCqybL.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=m25T5rAy43.

David Carlson, Volkan Cevher, and Lawrence Carin. Stochastic Spectral Descent for Restricted Boltzmann Machines. In Guy Lebanon and S. V. N. Vishwanathan (eds.), *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pp. 111–119, San Diego, California, USA, 09–12 May 2015a. PMLR. URL https://proceedings.mlr.press/v38/carlson15.html.

David Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015b. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/f50a6c02a3fc5a3a5d4d9391f05f3efc-Paper.pdf.

David Chanin. Are sparse autoencoder benchmarks reliable?, 2026. URL https://arxiv.org/abs/2605.18229.

David Chanin and Adrià Garriga-Alonso. Sparse but wrong: Incorrect l0 leads to incorrect features in sparse autoencoders, 2026. URL https://openreview.net/forum?id=wUzBBsrdB1.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring linear features across language models with model stitching. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48531–48563. Curran Associates, Inc., 2025. doi: 10.52202/085713-1621. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4569a868e7aa891248832ec08445d071-Paper-Conference.pdf.

Lizhang Chen, Jonathan Li, and qiang liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Blz4hjxLwU.

Dami Choi, Christopher J. Shallue, Zachary Nado, Jaehoon Lee, Chris J. Maddison, and George E. Dahl. On empirical comparisons of optimizers for deep learning, 2020. URL https://openreview.net/forum?id=HygrAR4tPS.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/2cb274e6ce940f47beb8011d8ecb1462-Paper.pdf.

Ashok Cutkosky and Harsh Mehta. Momentum improves normalized SGD. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2260–2268. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/cutkosky20b.html.

Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. Grounding representation similarity through statistical testing. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1556–1568. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0c0bf917c7942b5a08df71f9da626f97-Paper.pdf.

Sara Dragutinović and Rajesh Ranganath. To use or not to use muon: How simplicity bias in optimizers matters. In *Workshop on Scientific Methods for Understanding Deep Learning*, 2026. URL https://openreview.net/forum?id=GsZtgQf3IM.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sep 1936. ISSN 1860-0980. doi: 10.1007/BF02288367. URL https://doi.org/10.1007/BF02288367.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2958830.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 39622–39669. Curran Associates, Inc., 2025. doi: 10.52202/085713-1322. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/386432c7534eec9a1cd7cbeea90d7e9f-Paper-Conference.pdf.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba E. Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16543–16572. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/fel25a.html.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable features, reproducible subspaces: Understanding seed dependence in sparse autoencoders, 2026. URL https://arxiv.org/abs/2606.12138.

Eitan Gronich and Gal Vardi. The implicit bias of adam and muon on smooth homogeneous neural networks. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=DpIc1cpNKG.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1832–1841. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gunasekar18a.html.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1842–1850. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/gupta18a.html.

Suchit Gupte, Vishnu Kabir Chhabra, and Mohammad Mahdi Khalili. On the transferability of sparse autoencoders for interpreting compressed models, 2025. URL https://arxiv.org/abs/2507.15977.

Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, Yu-Gang Jiang, and Xipeng Qiu. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders, 2024. URL https://arxiv.org/abs/2410.20526.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Nicholas J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079. URL https://doi.org/10.1137/0907079.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. doi: 10.52202/068431-2176. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. ISSN 03036898, 14679469. URL http://www.jstor.org/stable/4615733.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A comprehensive benchmark for sparse autoencoders in language model interpretability. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=qrU3yNfX0d.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Anton Korznikov, Andrey V. Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do SAEs beat random baselines? In *Mechanistic Interpretability Workshop at ICML 2026*, 2026. URL https://openreview.net/forum?id=bEYHoD7fCj.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization, 2026. URL https://openreview.net/forum?id=hM44aMOjnq.

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970. ISSN 00361429. URL http://www.jstor.org/stable/2949656.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Frederik Kunstner, Jacques Chen, Jonathan Wilder Lavington, and Mark Schmidt. Noise is not the main factor behind the gap between sgd and adam on transformers, but sign descent might be. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=a65YK0cqH8g.

Frederik Kunstner, Robin Yadav, Alan Milligan, Mark Schmidt, and Alberto Bietti. Heavy-tailed class imbalance and why adam outperforms gradient descent on language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=T56j6aV8Oc.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8:355–362, 05 2017. doi: 10.1177/1948550617697177.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Quantifying feature space universality across large language models via sparse autoencoders, 2025. URL https://arxiv.org/abs/2410.06981.

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 73501–73548. Curran Associates, Inc., 2024. doi: 10.52202/079017-2338. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/8629b0fff229b8a27efb1422e990605f-Paper-Conference.pdf.

Patrick Leask, Bart Bussmann, Michael T Pearce, Joseph Isaac Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=9ca9eHNrdH.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. *International Journal of Computer Vision*, 127(5):456–476, May 2019. ISSN 1573-1405. doi: 10.1007/s11263-018-1098-y. URL https://doi.org/10.1007/s11263-018-1098-y.

Jiaming Li, Haoran Ye, Yukun Chen, Xinyue Li, Lei Zhang, Hamid Alinejad-Rokny, Jimmy Chih-Hsien Peng, and Min Yang. Breaking the block: Preserving data continuity to train superior SAEs for instruct models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SjNCl7f2e3.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon, 2025. URL https://arxiv.org/abs/2502.02900.

Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. Convergent learning: Do different neural networks learn the same representations? In Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar (eds.), *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*, volume 44 of *Proceedings*

of Machine Learning Research, pp. 196–212, Montreal, Canada, 11 Dec 2015. PMLR. URL https://proceedings.mlr.press/v44/li15convergent.html.

Yuanzhi Li and Ronen Eldan. Tinystories: How small can language models be and still speak coherent english, 2024. URL https://openreview.net/forum?id=yiPtWSrBrN.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, Miami, Florida, US, Nov 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.19. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2014. URL https://arxiv.org/abs/1312.5663.

Jordan F. McCann. Polymorphism is rotation: Operational mechanistic interpretability from a two-layer transformer to pythia-70m, 2026. URL https://arxiv.org/abs/2605.24577.

Johannes Mehrer, Courtney J. Spoerer, Nikolaus Kriegeskorte, and Tim C. Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11(1):5725, Nov 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19632-w. URL https://doi.org/10.1038/s41467-020-19632-w.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a7a3d70c6d17a73140918996d03c014f-Paper.pdf.

Ali Nasiri-Sarvi, Hassan Rivaz, and Mahdi S. Hosseini. SPARC: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=IJfvoc2GbZ.

Daniil Ognev, Célian Vasson, Lijie Hu, Kentaro Inui, and Benjamin Heinzerling. Sharedsae: One feature dictionary across language models, 2026. URL https://arxiv.org/abs/2609.04344.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1310–1318, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/pascanu13.html.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EjInprGpk9.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

Ofir Press and Lior Wolf. Using the output embedding to improve language models. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 157–163, Valencia, Spain, Apr 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2025/.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving dictionary learning with gated sparse autoencoders, 2024. URL https://arxiv.org/abs/2404.16014.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, Janos Kramar, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumpreLU sparse autoencoders, 2025. URL https://openreview.net/forum?id=mMPaQzgzAN.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, 2007.

Tianyu Ruan, Fengzhuo Zhang, Shuche Wang, and Shihua Zhang. Muon learns more robust and transferable features than adam, 2026. URL https://arxiv.org/abs/2606.09658.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451.

Yuval Sharon and Yehuda Dar. How do the architecture and optimizer affect representation learning? on the training dynamics of representations in deep neural networks, 2025. URL https://arxiv.org/abs/2405.17377.

Egor Shulgin, Sultan AlRashed, Francesco Orabona, and Peter Richtárik. Beyond the ideal: Analyzing the inexact muon update. In *The Third Conference on Parsimony and Learning (Recent Spotlight Track)*, 2026. URL https://openreview.net/forum?id=veVNVUSieC.

Devender Singh. The loss does not see the basis, but adam does, 2026. URL https://arxiv.org/abs/2608.05136.

Prabhu Teja Sivaprasad, Florian Mai, Thijs Vogels, Martin Jaggi, and François Fleuret. Optimizer benchmarking needs to account for hyperparameter tuning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9036–9045. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sivaprasad20a.html.

Teodor-Mihai Stupariu and Andrei Manolache. How the optimizer shapes learned solutions in equivariant neural networks. In *ICML 2026 Workshop on Weight-Space Symmetries: from Foundations to Practical Applications*, 2026. URL https://openreview.net/forum?id=ymiSLNQlkK.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G. Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=UoaxRN88oR.

Mark Tuddenham, Adam Prugel-Bennett, and Jonathon Hare. Orthogonalising gradients to speedup neural network optimisation, 2022. URL https://openreview.net/forum?id=-cII-Vju5C.

Bhavya Vasudeva, Jung Lee, Vatsal Sharan, and Mahdi Soltanolkotabi. The rich and the simple: On the implicit bias of adam and sgd. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 161725–161761. Curran Associates, Inc., 2025. doi: 10.52202/085713-5401. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/ec9b2a6ad5444caeff75efaa6176b3e4-Paper-Conference.pdf.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YzjS4jcfmS.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Bohan Wang, Qi Meng, Wei Chen, and Tie-Yan Liu. The implicit bias for adaptive optimization algorithms on homogeneous neural networks. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10849–10858. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21q.html.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent Y. F. Tan. Muon outperforms adam in tail-end associative memory learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=twbMFL0DMp.

Ashia Wilson, Rebecca Roelofs, Mitchell Stern, Nati Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/81b3833e2504647f9d794f7d7b9bf341-Paper.pdf.

Shuo Xie and Zhiyuan Li. Implicit bias of adamw: ℓ_∞-norm constrained optimization. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CmXkdlO6JJ.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10524–10533. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23988–24021. Curran Associates, Inc., 2024a. doi: 10.52202/079017-0756. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2ac79356a03fe5e9250e5e77ebc76e6e-Paper-Conference.pdf.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 131786–131823. Curran Associates, Inc., 2024b. doi: 10.52202/079017-4190. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/ee0e45ff4de76cbfdf07015a7839f339-Paper-Conference.pdf.

Jim Zhao, Tin Sum Cheng, Wojciech Masarczyk, and Aurelien Lucchi. Optimizer choice matters for the emergence of neural collapse. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=9EPYWJrib1.

Rosie Zhao, Depen Morwani, David Brandfonbrener, Nikhil Vyas, and Sham M. Kakade. Deconstructing what makes a good optimizer for language models. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=hxTKJNkyVz.

OVERVIEW OF THE APPENDICES

The appendices follow the main text and each opens by naming the claim it supports. Part I (Appendices A–D) gives the configurations, the sparse-autoencoder recipe as a specification, the protocol as a checklist with what each of its controls cost when it was absent, and the guards on the analysis code. Part II (Appendices E–W) supports §3: the loss readouts and their corrections, the pairwise table, the ranking inversion, the two further instantiations of each geometry, the 60.5M run at the 13.9M budget, the robustness controls, the hybrid surface, the mechanism controls, and the functional and probe readouts. Part III (Appendices X–AC) supports §4. Part IV (Appendices AD–AG) supports §5. Part V (Appendices AH–AK) gives the statistical audit, the derivation of the floors, the toy-model test that failed, and the compute accounting. Throughout, "the gap" is the Euclidean−spectral difference in SAE explained variance unless another pair is named, and every interval from three seeds is marked † and read as an observed range.

A EXPERIMENTAL CONFIGURATIONS

This appendix supports §2: the base models, the vocabulary rule that sets the second corpus's scale, the update rules as implemented, the learning-rate protocol and its bracketing check, and the residual streams the dictionaries were fitted to.

Base models. Decoder-only pre-LayerNorm transformers (Vaswani et al., 2017; Xiong et al., 2020) in the GPT-2 style (Radford et al., 2019): learned absolute position embeddings added to token embeddings, LayerNorm (Ba et al., 2016) (not RMSNorm) before attention and before the MLP, a $4d$ MLP with GELU (Hendrycks & Gimpel, 2023), residual connections around both sublayers, a final LayerNorm before the tied output projection (Press & Wolf, 2017), dropout 0 in every arm and run. The q, k and v projections are separate matrices, not a fused $[3d, d]$ block: orthogonalizing a fused block is a different operator from orthogonalizing three $[d, d]$ matrices, and the spectral arm's definition depends on which one is applied. Initialization is $\mathcal{N}(0, 0.02^2)$ for every embedding and linear weight, with residual projections scaled by $1/\sqrt{2n_{\text{layer}}}$. Context length 256, batch size

32, bfloat16 throughout; the Euclidean and sign-like arms are bit-identical to `torch.optim.SGD` and `torch.optim.Adam` in unit tests, both casting to fp32 internally, so the tests cover the code path that runs under bf16. Table 2 lists the six runs.

Table 2: The six base-model runs. Each is trained under all three geometries; "tok/par" is training tokens per parameter. The seed count is per arm at the selected rates; the two larger scale runs were re-trained at eight seeds for the depth sweeps of Appendix Y.

Params	**Corpus**	**Layers**	**Heads**	d_{model}	**Vocab**	**Steps**	**tok/par**	**Seeds**
13.9M	TinyStories	6	6	384	8,192	8,000	4.71	8
44.8M	TinyStories	8	10	640	8,192	8,000	1.46	3 (8 at depth)
60.5M	WikiText-103	8	10	640	32,768	8,000	1.08	3 (8 at depth)
210M	TinyStories	16	16	1,024	8,192	37,000	1.44	3
13.9M	TinyStories	6	6	384	8,192	34,000	20.04	3
60.5M	WikiText-103	8	10	640	32,768	34,800	4.71	3

The vocabulary rule, and why the second corpus is read at 60.5M. The manipulated update rule reaches only the matrix parameters, so the fraction of a model under the manipulation depends on how much of it is embedding. The rule, fixed before any measurement, caps the vocabulary by training-split frequency at the smallest size whose UNK rate is at most 2%, and requires at least 60% of parameters under the manipulation; a candidate that satisfies neither is reported as a failure, not rescued. On TinyStories (Li & Eldan, 2024) the cap of $8{,}192$ gives an UNK rate of 0.118% and puts 76.5% of the 13.9M model under the manipulation, where GPT-2's $50{,}257$-token vocabulary would leave 35.4%. On WikiText-103 (Merity et al., 2017) the rule fails at 13.9M (Table 3): no cap reaches the UNK bound while leaving 60% of parameters manipulated, because at width 384 a $32{,}768$-token embedding costs more parameters than every matrix put together. Rather than relax the rule for the candidate that failed it, the model was enlarged until the unchanged rule passed (eight layers and width 640, 65.0% manipulated at 0.391% UNK) which is why the second corpus is studied at 60.5M and not 13.9M.

Table 3: The vocabulary rule applied to WikiText-103. The cap must reach an UNK rate of at most 2% and leave at least 60% of parameters under the manipulated rule; only the $32{,}768$-token cap at the larger shape does both.

		% under mnp. (and resulting size)	
Cap	**UNK (WikiText-103)**	**6L,** d=384	**8L,** d=640
8,192	10.751%	76.5% (13.9M)	87.9% (44.8M)
16,384	4.312%	62.4% (17.0M)	78.7% (50.0M)
32,768	**0.391%**	45.6% (23.3M, fails)	**65.0%** (60.5**M**)

The update rules as implemented. The Euclidean arm is SGD with heavy-ball momentum (Polyak, 1964) 0.9; the sign-like arm is Adam (Kingma & Ba, 2015) with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$, sign-like in the sense that its update direction is set by the sign of the gradient and its magnitude by an estimate of relative variance (Balles & Hennig, 2018; Kunstner et al., 2023); the spectral arm applies the quintic Newton–Schulz iteration $X \leftarrow aX + b\,(XX^\top)X + c\,(XX^\top)^2X$ with $(a, b, c) = (3.4445, -4.7750, 2.0315)$ for five steps to the momentum-accumulated gradient normalized by its Frobenius norm, the short side taken as rows so the Gram matrix is the smaller one, with momentum 0.9 and Nesterov (Sutskever et al., 2013). We run the iteration in float32 where the reference implementation uses bfloat16: at these sizes there is no throughput argument, and the extra precision keeps the update's singular-value spectrum measurable rather than quantization-limited, which matters because Appendix K reads exactly that spectrum. The iteration's polynomial has no fixed point at 1 ($p(1) = 0.7010$), so five steps do not drive the singular values to one: measured on 720 real gradients from these runs, the applied update's singular values span 0.0026 to 1.20 with a median max/min ratio of 468, against 1.0001 for the exact-SVD arm. The five-step iteration is therefore a decidedly approximate orthogonalization (the inexact update is analysed by Shulgin et al., 2026), and whether that approximation matters is the question Appendix K answers. Auxiliary

parameters: embeddings, LayerNorm gains, and biases, take Adam at 3×10^{-4} in every arm and are never swept (Appendix R tests the alternative). The rate warms up linearly over the first 5% of steps and then follows a cosine decay (Loshchilov & Hutter, 2017) to zero; gradients are clipped at global norm 1 (Pascanu et al., 2013); weight decay is 0 in every headline run (Appendix M).

Learning-rate grids and selection. Rates are swept independently per geometry, since an optimizer comparison is only as fair as its tuning protocol (Choi et al., 2020; Sivaprasad et al., 2020): six log-spaced points per arm at 13.9M and seven, half a decade apart, at the three larger scales, each grid spanning two to three decades (the 60.5M budget run of Appendix L uses five half-decade points centered on the published optima), at half the run length and one seed, with the full sweep recorded rather than only its winner and an audit that fails any arm whose optimum lands on a grid edge. Every sweep optimum was interior in every run. The final stage retrains three rates per arm at full length and the run's seed count: the arm's own best, and the two rates whose converged loss lies nearest the common target, which is the worst arm's best loss. Selected rates differ across arms by $398\times$ to $1778\times$, which is why no comparison in this paper is made at a shared rate.

Are the full-duration operating points bracketed? The construction above leaves a gap a careful reader will see. Because the common target is the worst arm's best loss, the two companion rates of a better arm both lie below its best, so at full length that arm's optimum is bracketed from below only; its upper side was seen at half length and one seed. The stored rows say how much that matters. At every scale and arm the sweep's best is also the best of the three rates run at full length (12 of 12); the full-length best sits at the top edge of its grid for the sign-like arm at 13.9M, 44.8M and 60.5M and for all three arms at 210M; and the sweep's best is separated from its nearest neighbour by 2–3 seed standard deviations at 13.9M and 210M (14–90 at the two middle scales), so one seed alone cannot rank those pairs. We therefore ran the missing upper bracket at full length, with each run's own configuration, seed count and inline final-site readout, for the three cases a single card could reach (Table 4), under a rule fixed in advance: BRACKETED when the upper neighbour is worse by a seed-paired interval excluding zero.

Table 4: The upper bracket at full length ($8{,}000$ steps): validation loss at the sweep's next rate above the selected one, minus the selected rate's, with a seed-paired interval (the two three-seed rows† carry observed ranges); and the same margin at half length and one seed. All three are BRACKETED. At 13.9M and 60.5M the upper rate trains badly (2.82 and 5.32 nats); at 44.8M it closes most of its half-length deficit and still loses.

scale	arm	selected	upper	loss, upper − selected	half length	seeds
13.9M	sign-like	1.58×10^{-3}	3.98×10^{-3}	$+1.285$ $[+1.249, +1.326]$	$+1.178$	8
44.8M	sign-like	5.62×10^{-4}	1.47×10^{-3}	$+0.051$ $[+0.035, +0.064]$	$+0.486$	3
60.5M	sign-like	5.62×10^{-4}	1.47×10^{-3}	$+1.598$ $[+1.577, +1.631]$	$+0.855$	3

All three held, so the matched-compute operating points at these scales are interior at full length as well as at half. Two residues are stated rather than closed. At 210M the final grid also skipped the lower neighbours of the sign-like and spectral bests (2.15×10^{-4} and 3.16×10^{-3}), which lost at half length by 0.0025 and 0.0021 nats, inside one seed's noise, so those two operating points are the sweep's choice rather than verified optima; the matched-loss readout there sits below the best in any case and is unaffected. And the 44.8M margin shrank from 0.49 to 0.05 nats between half and full length, which is the objection's own mechanism in miniature: a longer run tolerates a higher rate, and a half-length sweep can only bracket what a full-length run would show by a margin larger than that drift. Fourteen runs, 1.7 h on one RTX 3090.

Are the residual streams unusual? A capped vocabulary on simple text makes small models unlike the ones dictionaries are usually fitted to, and the question that matters is whether their residual streams differ in the ways a sparse fit is sensitive to. Table 5 puts the statistics of our four scales beside the same statistics of two models sparse autoencoders are routinely fitted to, GPT-2 small (Radford et al., 2019) and Pythia-70m (Biderman et al., 2023), at the same two sites (the middle block and the final pre-LayerNorm stream) on the same text and the same frozen validation windows; GPT-2 shares our tokenizer, so it sees the identical token windows uncapped, and the uncapped stream was checked to reproduce the capped one token for token through the cap map.

Four expectations were fixed before the reference models were read: that every one of our streams is more isotropic than both references, has fewer massive activations than both, has a higher effective rank at the final site, and that dropping the UNK positions from the readout moves the gap by less than the floor.

Table 5: Residual-stream statistics over 163,840 frozen validation tokens: the effective rank of the covariance (exponential of the eigenvalue entropy) as a fraction of the width; the leading principal component's share of variance; anisotropy, the expected cosine between two tokens; the median excess kurtosis of a coordinate; the largest coordinate over the median coordinate; and the position-0 norm over the rest. Ours are the matched-loss arms averaged (seeds 0–7 at 13.9M, seed 0 elsewhere; the arms agree to the digits shown at the final site). Token-norm spread, not shown, is 0.11–0.23 for ours and 0.12–0.18 at the references' final sites.

model	site	rank / d	top PC	aniso.	kurt.	max / med.	pos. 0
13.9M, ours (8 seeds)	final	**0.44**	0.05	**0.04**	0.05	10	1.3
	mid	0.43	0.06	0.09	0.11	12	1.8
44.8M, ours	final	0.33	0.06	0.04	0.12	11	1.4
	mid	0.36	0.04	0.08	0.10	13	1.8
60.5M, ours (WikiText)	final	0.33	0.05	0.05	0.06	13	1.2
	mid	0.33	0.05	0.09	0.15	16	2.1
210M, ours	final	0.26	0.09	0.09	0.23	70	1.4
	mid	0.36	0.06	0.09	0.11	28	2.1
GPT-2 small, TinyStories	final	0.07	0.29	0.81	0.18	111	0.8
	mid	0.003	0.92	0.50	0.38	2,425	38
GPT-2 small, WikiText	final	0.12	0.19	0.83	0.22	144	0.8
	mid	0.003	0.91	0.46	0.33	2,298	38
Pythia-70m, TinyStories	final	0.09	0.39	0.97	0.45	4	1.0
	mid	0.01	0.72	0.41	0.54	503	9
Pythia-70m, WikiText	final	0.06	0.52	0.96	1.72	11	1.0
	mid	0.03	0.61	0.35	0.26	493	9

Three of the four held and one missed. Our streams are far more isotropic than either reference at every site, expected cosine 0.04–0.09 against 0.35–0.97, the references' final sites at 0.81–0.97, and of far higher effective rank (0.26–0.44 of the width at the final site against 0.06–0.12), with a leading component carrying 5–9% of the variance against 19–52%. On massive activations (Sun et al., 2024) the expectation missed: GPT-2's streams hold coordinates 111–2,400× the median and Pythia's middle block 500×, but Pythia's final site is cleaner than ours (4–11× against 10–70×), so ours are not below both references. The first-token effect behind those coordinates (38× the norm at GPT-2's middle block, 9× at Pythia's) is 1.2–2.1× in our models. Of the (run, site, statistic) cells, 76% lie inside the range the two references span, and every cell outside it is one of the same two statistics in the same direction. At the middle site the arms differ in the way Appendix X predicts: at 13.9M the spectral arm's stream is the most anisotropic (0.17 against 0.05 and 0.07) and of the lowest rank (0.33 against 0.54 and 0.43), the embedding it leaves standing being a shared direction, and agree at the final site to the second decimal. (A report that Muon-trained weights and representations carry higher stable and effective ranks, made for equivariant networks on point clouds (Stupariu & Manolache, 2026), is therefore not what these language-model streams show.) The UNK token is not a factor: it occupies 286 of the 163,840 validation positions (0.175%), its residual norm is 0.97–1.02× the median token's, and dropping every UNK position from the published readout moves the gap from +0.0299 to +0.0300.

So the premise of the worry is right and its direction is the opposite: at the published budgets these models' residual streams are easier targets for a sparse dictionary than the ones the field fits, being near-isotropic, high-rank, near-Gaussian, without the massive activations and first-token outliers that dominate a GPT-2 stream, and the gap is measured in that regime. Whether it survives the anisotropic, low-rank, outlier-carrying streams of models trained at scale is untested here and stated as such in §6; one step towards that regime is Appendix L, where 4.4× the 60.5M budget moves two arms' mid-stack streams into it, the final site stays out of it, and the mid-stack stops being fittable.

B THE SAE READOUT, SPECIFIED

This appendix supports §2 and §3: it states the readout precisely enough to be reproduced, lists every configuration the paper calls "the gap", and reports the controls on the readout itself: activation scale, the evaluation slice, outlier tokens and dead latents.

Every number the paper calls "the gap". The Euclidean−spectral gap at 13.9M appears as $+0.0300$, $+0.0320$, $+0.0414$ and $+0.0430$ in different places, and the Euclidean arm's absolute EV as 0.8196, 0.8259 and 0.8140. These are different configurations, not disagreeing measurements: they differ in the readout (matched loss or matched compute), the SAE learning rate (3×10^{-4} throughout the headline readout; 10^{-3} is what the per-arm tuning stage of Appendix P selected), and the activation source (validation-sourced everywhere except the pool experiment of Appendix O, which must draw from the training split to reach $100\times$). Table 1 is the reference for every headline claim.

gap	Eucl. EV	readout	SAE lr	activations / slice	where
$+0.0300$	0.8196	matched loss	3×10^{-4}	val, single held-out slice	Table 1
$+0.0414$	0.8196	matched compute	3×10^{-4}	val, single held-out slice	Table 1
$+0.0320$	0.8259	matched loss	10^{-3}	val, select + report slices	App. P
$+0.0430$	0.8140	matched compute	3×10^{-4}	*train*-sourced pool, $1\times$	App. O

One coincidence is worth naming so it is not read as a copy: Table 1's matched-compute column reads $+0.0414$ at both 4.71 and 20.04 tokens per parameter. The underlying values are 0.041366 and 0.041414, independent measurements from different runs that agree to four decimals; §3 calls this flat, which is what it is.

Where the activations come from. Dictionaries are fitted to the final pre-LayerNorm residual stream (layer 6 of the 13.9M model, layer 8 of the 44.8M and 60.5M models, layer 16 of the 210M model) at every token position. Pre-, not post-: the final LayerNorm and the tied output projection run after the site the dictionary sees, and the hook applies no LayerNorm. LayerNorm would remove most of the $11.76\times$ scale spread discussed next, so reading after it would answer the scale question by construction rather than by measurement; the pre-LayerNorm site is also what the functional metric substitutes into, which is what lets the LayerNorm and the head carry a reconstruction error into the model's own loss (Appendix V). We collect $163{,}840$ activation vectors per model from the frozen validation batches, the same token sequences for every arm, rate and seed, and use them raw: no normalization, no whitening, no centering beyond what the EV definition performs.

Activation scale does not carry the gap. Using activations raw is a conspicuous choice when the arms differ $11.76\times$ in mean ℓ_2 norm (41.87, 10.12, 3.56) and per-arm normalization is common practice. The whitening condition of Appendix T removes scale but destroys the covariance spectrum with it and moves EV levels by about 0.17, so it cannot isolate scale; dividing each arm's activations by their own mean norm does, one scalar per arm with directions, covariance shape and every higher moment preserved. Refitting the identical SAE under that transform at eight seeds gives $+0.0316$ $[+0.0304, +0.0329]$ against $+0.0320$ $[+0.0308, +0.0332]$ for the untransformed activations in the same run: 98.9% of the gap survives. EV is scale-invariant by construction, so this is a statement about the SAE's optimization under a fixed learning rate and initialization rather than about the metric, which is the form the objection takes; it does not survive.

The evaluation slice, and what its overlap with the training pool is worth. The SAE trains on all but a held-out slice of $8{,}192$ activations, and every reported number is computed on that slice. The split is at sequence granularity, $8{,}192 = 32 \times 256$ is exactly the first validation batch, so the held-out slice is 32 complete context windows and no window contributes vectors to both sides. One caveat needs a precise statement: validation windows are drawn at independent random offsets into the token file, so windows in different batches can overlap in the underlying token stream even though no window is split. At our seed, 7 of the $19{,}456$ holdout×train window pairs overlap and 9.0% of held-out token positions recur somewhere in the training pool; no activation vector is shared, since an overlapping token sits at a different offset within its window and carries a different context, and the pattern is byte-identical across arms. Rather than argue that this cannot bias the contrast,

we measured it: the 24 published dictionaries at 13.9M were held fixed and scored on 20 slices of 32×256 drawn the published way (overlapping the pool by 8.1% on average, range 1.0–15.1%), drawn so that no window touches any training-pool window or any other slice (0% by construction), and drawn from the training split, which the base model saw for one epoch and the SAE never did.

Scored on	Eucl.	sign	spectral	**Eucl. − spectral**
the published slice (9.0% recur)	0.8196	0.8209	0.7896	+0.0300
published-style slices (8.1%)	0.8215	0.8225	0.7914	+0.0300
disjoint slices (0%)	0.8198	0.8208	0.7896	**+0.0303**
training-split slices	0.8179	0.8190	0.7876	+0.0303

The overlap is worth about $+0.002$ of EV to every arm alike ($+0.0017$, $+0.0017$, $+0.0019$; slice standard error ≈ 0.0011, and the three are one measurement rather than three, since every arm is scored on the same slices): a fifth of the floor, and shared. The gap moves by $+0.0002$ when every recurring position is removed and by $+0.0003$ on the other split of the corpus, both an order of magnitude inside the ± 0.0014 the paper prints. What this does not settle is whether selecting learning rates on validation loss and then scoring SAEs on validation activations is circular in a way a third split would expose; it is the same procedure for every arm, there is no third split, and the training-split row is the nearest available check rather than a substitute for one.

Explained variance. For activations x and reconstruction $\hat{x}$,

$$\mathrm{EV} \;=\; 1 - \frac{\mathbb{E}\big[\|\hat{x} - x\|_2^2\big]}{\mathbb{E}\big[\|x - \bar{x}\|_2^2\big]},$$

with $\bar{x}$ the per-dimension mean over the evaluation slice. Both expectations are over tokens and the squared norms are summed over dimensions first, so EV is one aggregate ratio rather than a per-token or per-dimension average; $\mathrm{EV} = 0$ means no better than predicting the mean activation. Because it is an aggregate ratio, a few very high-norm tokens (position 0 above all) could in principle dominate numerator and denominator together, and the arms differ enough in activation scale to differ plausibly in outlier structure. Re-evaluating the 24 published SAEs on the same held-out slice under three alternative readouts settles it: the gap is $+0.0300$ as published, $+0.0297$ excluding position 0 (99.1% of it), $+0.0299$ excluding positions 0–3, and $+0.0251$ under a per-token mean EV (83.6%), all four clearing the floor at eight seeds. Outlier tokens account for under 1% of the effect; the convention matters for magnitude by about a fifth, which is a reason to state it, not a reason to prefer one.

Objectives. Expansion factor 8 ($d_{\text{sae}} = 8d_{\text{model}}$), Adam at 3×10^{-4}, 3,000 steps, batch 4,096, decoder columns renormalized to unit norm after every step. TopK (Gao et al., 2025), in the lineage of the k-sparse autoencoder (Makhzani & Frey, 2014), takes the encoder pre-activations $W_e(x - b_d) + b_e$, keeps the $k = 32$ largest, applies ReLU to those and zeroes the rest. Thus, $L_0 \leq 32$ always and $= 32$ whenever at least 32 pre-activations are positive. This holds at every site we report except layer 0, the embedding output, where the activation distribution is low enough in effective dimension that fewer than 32 survive the rectifier and achieved L_0 falls to 22.4–29.3. Layer 0 is a control, and no headline number is read there. JumpReLU gates on a learned per-feature threshold with a rectangle-kernel straight-through estimator, so its L_0 is an outcome rather than a setting. It calibrates the threshold from each arm's own pre-activation quantile so the arms begin at matched sparsity rather than a matched absolute threshold, with the straight-through bandwidth set in units of the positive pre-activation standard deviation. Appendix Q.2 shows what that calibration left uncorrected. The other architectures are specified in Appendix Q.

Dead features, effective features and interference. A differential dead-latent rate across arms would make an EV deficit a property of SAE optimization rather than of the representation, and would pull the median maximum cosine of Appendix AD toward chance mechanically, since dead features carry near-random decoder directions. At the final residual stream every one of the 3,072 dictionary features is alive in every arm and seed (dead fraction exactly 0, where dead means firing on no token of the evaluation slice), and across the hyperparameter grid of Appendix P the largest dead fraction in any cell is 0.0010; so the headline gap and the dictionary comparisons are never computed over a partly degenerate dictionary. Dead features do appear at other depths, up

to 0.14 at layer 1. which is why the depth comparison carries an exclusion rule of its own (§2). Usage is not uniform even where nothing is dead: the effective number of features by usage entropy, $\exp(-\sum_i p_i \log p_i)$ with p_i feature i's share of firings, is $2{,}508.8$, $2{,}547.7$ and $2{,}700.6$ for the three arms (means over base seeds 0–3), and this quantity is the one behind the 192-feature figure of Appendix W (the difference of those means; 219.6 at seed 0). For a dictionary with unit-norm columns w_i, interference is the total off-diagonal squared overlap $\Psi = \sum_i \sum_{j \neq i} \langle w_i, w_j \rangle^2$, normalized so an orthogonal dictionary gives 0; it is threshold-free, which is why the same quantity is read on the toy model and the transformer without choosing a cutoff.

C THE PROTOCOL AS A CHECKLIST, AND WHAT EACH CONTROL COST WHEN ABSENT

§5 ends with what a cross-model comparison should report. This appendix states the protocol behind it as something a reader can apply, and records what each of its controls cost this study when it was absent. Every consequence in Table 6 is an event in this work, not a hypothetical; most were paid for in the toy model, and the transformer study inherited the controls rather than rediscovering them. In particular the auxiliary-rule reversal is a toy-model event: the transformer arms have used an identical auxiliary rule and rate from the first run by design, so no transformer result turns on it.

Table 6: Controls, and what each cost us when absent. The study column records where the lesson was learned.

Control	Consequence when absent	Study
Per-geometry LR sweep	Arm pinned at grid ceiling, loss still falling 28%	toy
Converged endpoints only	97.8% of loss-matched selections were mid-training	toy
Identical aux rule *and* rate	Effect direction reversed	toy
Effect-band grid coverage	A false negative (1 in-band level)	toy
Absolute-effect floor	Two retracted claims	toy
Matched loss *and* compute	4 sign, 3 verdict disagreements of 12	transformer
SAE stability baseline	A null at the instrument's own floor	transformer
Guards on the analysis code	Four retractions, all in the analysis code	both

When comparing SAE reconstruction across base models:

1. **Report the base optimizer and its swept rate.** Optimal rates differed across arms by $398\times$ to $1778\times$ here; a shared rate measures at most one arm at its operating point.
2. **Read at matched loss and at matched compute.** Neither is correct alone; ours disagree on 4 signs and 3 verdicts of 12 transformer comparisons (Appendix E).
3. **State the readout depth, and do not generalize across it.** The sign of our effect reverses between layers (§4).
4. **Report achieved L_0 and the dead fraction, never the setting.** An identical target initial L_0 produced a $4.08\times$ spread in achieved sparsity (Appendix Q.2).
5. **Report a stability floor for any dictionary comparison.** Two SAEs on byte-identical activations reach only 0.4256 median maximum cosine here; without that number a dictionary difference has no scale (Appendix AD).
6. **Say whether "the dictionary changed" means basis or content.** Two base seeds are $+0.0090$ above chance geometrically and $+0.4360$ above chance functionally (Appendix AE).
7. **Require an absolute effect floor, not only an interval excluding zero.** At our seed variance, 11 of 18 comparisons excluded zero and 7 cleared the floor (Appendix AI).
8. **State the resampling unit of every interval.** Seeds and evaluation activations answer different questions (Appendix AH).
9. **Guard the analysis code, not only the sweep.** Every retraction in this project came from the analysis layer (Appendix D).

What we do not deliver. We have not re-analyzed a published cross-model comparison under this checklist. That would be the strongest demonstration, and doing it responsibly needs the original authors' checkpoints and swept rates, which are not generally released. What can be done without checkpoints is to audit what each paper reports, which is Appendix AG; the two are not the same and we do not present the weaker as the stronger.

Multiple comparisons, corrected rather than deferred. The absolute floor is a filter on effect size, not on false-positive rate, and the two are different questions, so the second is answered here rather than left to the reader. Every comparison in the paper whose verdict is "floor-clearing with an interval excluding zero" was pooled into one family: 24 pairwise comparisons, 35 depth sites (25 plus the ten measurable sites of the 210M run), 8 SAE-grid cells, 3 pool sizes, weight decay, the JumpReLU frontier and 3 dictionary contrasts, $m = 75$, and a Holm–Bonferroni step-down (Holm, 1979) at $\alpha = 0.05$ applied to a paired t on the per-seed differences. The t rather than the bootstrap, because at $n = 3$ a percentile interval saturates at the observed range and cannot widen to any corrected level, whereas a t-interval at two degrees of freedom widens enormously, which is the right answer for three seeds.

Five claimed verdicts change, every one at $n = 3$, and two of them are a pair we already decline to claim. The correction removes Euclidean−sign at 60.5M matched loss ($+0.0144$, $t = 7.3$, Holm $p = 0.18$) and at 210M matched compute ($+0.0106$, $t = 4.6$, $p = 0.32$). This is the middle pair, which Appendix AI lists as not asserted and Table 14 leaves unresolved. It also removes Euclidean−sign at three sites of the 210M depth profile: layers 8 and 9 on the negative side (-0.0423, -0.0195; $t = -15.9$, -14.5; $p = 0.055$, 0.061) and layer 12 on the positive ($+0.0177$, $t = 12.2$, $p = 0.080$). Under correction the 210M reversal rests on layer 7 (-0.0675, $p = 0.035$) against layers 13–16, which is still a sign reversal and is what three seeds can carry. Appendix H's "18 of 24" therefore reads 16 of 24 under correction, and every one of the 16 comparisons involving the spectral arm survives, as does every depth site at $n = 8$, every grid cell, every pool size, weight decay, the frontier and all three dictionary contrasts; the smallest $|t|$ among claimed $n = 8$ comparisons is 11.4 against a plain Bonferroni critical value of 5.8, and the headline gaps sit at 28–190 standard errors. The floor's claim to conservatism is true in the other direction: it removes five comparisons that a Holm-corrected test would retain (three layer-0 sites, one mid-stack site at 44.8M, and Euclidean−sign at 13.9M matched compute) all with $|t|$ between 7 and 58 and effects of 0.0003 to 0.0092. That is the plateau pathology §2 describes: seeds collapse, p vanishes, and nothing of consequence has been found. The floor and the correction guard against different errors; this paper needed the first more, and applies both.

D GUARDS ON THE ANALYSIS LAYER

Sweep-level guards: grid coverage, seed minima, loss parity, and learning-rate edge detection are standard, and would have caught none of this project's retractions. Every one came from the analysis code that turns numbers into verdicts: a floor specified in prose and never applied; a criterion evaluated on a grid too coarse to resolve it; a matched-sparsity readout that clamped to a boundary and reported the clamp as a match. The assertions below now sit inside that code, and each is tested against the retraction that motivated it.

- **Non-empty comparison interval.** When per-unit ranges are disjoint, $\max(\text{mins}) > \min(\text{maxes})$ silently yields an inverted interval whose endpoint is then used as a comparison point. An empty interval means no comparison exists. No result, not a negative one.
- **Interior readout.** A matched-value readout must be bracketed by real measurements in every unit's frontier, not merely within a pooled range; linear interpolation routines clamp outside their domain and return a boundary value that is then reported as an interpolated match.
- **Grid resolves the criterion.** A criterion of the form "differs by at least r" cannot be evaluated on a grid whose adjacent points differ by r or more, since one grid step satisfies it by construction.
- **Floor applied.** No difference may be reported unless it both excludes zero and clears its absolute floor.

The same discipline extends to the text: every number printed in this paper is re-derived from its result file by a check run against the released results, and every claim a stronger measurement superseded is either absent from the text or labeled as the reading it replaced. The retracted claims are named in the appendices where they arose, beside the measurement that replaced them, rather than collected here.

E VALIDATION LOSSES AT THE SELECTED CELLS, AND THE LOSS-ADJUSTED GAP

This appendix supports §3's statement that residual loss mismatch does not explain the deficit. It reports the converged loss of every selected cell, the parity gate and why it fails, the two loss projections and why they move in opposite directions, the within-arm baseline, and the training-budget confound in full.

The losses behind the residual. Table 7 gives the mean converged validation loss per arm at the matched-loss selected rate, and the resulting parity (largest over smallest), so a reader can see which arm carries the residual and in which direction rather than only the ratio. In every run the spectral arm sits below Euclidean even after matching, which is the comparison the headline gap is read on; spectral is not the lowest-loss arm after matching, since sign-like is lower at 44.8M (1.4484 against 1.4727) and at 60.5M (3.7510 against 3.7676). The matched-loss handicap therefore falls on each arm differently, and only the Euclidean$-$spectral residual is load-bearing.

Table 7: Mean converged validation loss per arm at the matched-loss selected rate, and parity (largest over smallest), for the three headline scale runs.

Run	Euclidean	sign-like	spectral	parity
13.9M	1.5877	1.5830	1.5614	1.0169
44.8M	1.4854	1.4484	1.4727	1.0256
60.5M	3.8052	3.7510	3.7676	1.0145

The parity gate, and why it fails. Our pre-set gate is that parity falls within 1.5%. All three runs miss it. Across each arm's best achievable loss the spread is 1.066–1.093; at the matched-loss cells actually compared it is 1.0169, 1.0256 and 1.0145, so the readout reduces a 6–9% disagreement to 1.5–2.6% rather than removing it, and two of the three still miss. That residual is what the projection below prices, and it is why we report the gate as failed rather than drop it. The direction of the residual bias is easy to get wrong. It is tempting to say that a spectral arm sitting at lower loss works against the reported gap "under the alternative that better models reconstruct better". That invokes the wrong alternative: $\mathrm{dEV}/\mathrm{d}L$ is positive within each arm, so in the regime we observe better models reconstruct worse, and a spectral arm at lower loss would if anything inflate the reported gap. That is exactly why the correction exists and is applied, and applying it leaves the gap floor-clearing with its sign unchanged.

Two projections, and why they move in opposite directions. Table 8 gives the gap at every cell raw and after projection onto each run's own matched-loss reference along each arm's fitted EV-versus-loss curve (projection i); this is the quantity the "3–11% shrink" of §3 summarizes, its matched-loss column spanning -2.6% to -8.1% (to -10.6% under a low-fit variant). The matched-compute column shrinks far more, 14–37%, which is the expected direction. That readout does not equalize loss at all, so the projection has more to remove. And we report it because §3 leans on that readout for giving a larger raw gap, most of whose excess is loss mismatch. No cell flips sign and none falls below the floor. Projection (i) also reconciles the two readouts: raw, the matched-loss and matched-compute gaps differ by 0.0114, 0.0231 and 0.0090 across the three runs; adjusted, by 0.0028, 0.0002 and 0.0012. A second projection (ii), onto a single reference $L = 1.7677$ common to all arms, enlarges the gap instead, to $+0.0326$. The two differ in three ways fixed before the runs: (ii)'s reference sits about 0.18 nats worse than any headline cell and outside every arm's operating region, because no two arms' measured frontiers overlapped; (ii) reads the grids extended downward by three points, which is what created an overlap to interpolate within; and the pre-registered R^2

rule routes each arm separately, putting Euclidean ($R^2 = 0.394$) and sign-like (0.222) on local interpolation and leaving spectral (0.949, slope $+0.0546$) on a global slope, so the arms are not projected by one functional form. We consider (ii) the stricter test and (i) the more interpretable one; a reader who trusts only a projection that stays inside the measured region should take (i), which shrinks the gap and leaves it floor-clearing, and one who wants all three arms at one loss should take (ii), which enlarges it. Neither changes a verdict.

Table 8: The gap at each headline cell, raw and after projection (i) onto the run's own matched-loss reference along the arm's fitted EV-versus-loss curve.

run	readout	raw gap	adjusted	change
13.9M	matched loss	$+0.0300$	$+0.0276$	-8.1%
13.9M	matched compute	$+0.0414$	$+0.0304$	-26.6%
44.8M	matched loss	$+0.0461$	$+0.0435$	-5.6%
44.8M	matched compute	$+0.0692$	$+0.0433$	-37.4%
60.5M	matched loss	$+0.0573$	$+0.0557$	-2.6%
60.5M	matched compute	$+0.0663$	$+0.0569$	-14.2%

The three slopes, including the one that runs the other way. Fitted $\mathrm{dEV}/\mathrm{d}L$ at 13.9M is $+0.0267$ for Euclidean ($R^2 = 0.010$), $+0.0916$ for spectral ($R^2 = 0.648$) and -0.1079 for sign-like ($R^2 = 0.415$). The main text quotes the first two because the Euclidean$-$spectral pair carries the headline; the sign-like slope is negative, so the correction does not displace all three arms in the same direction, and a reader reconciling the adjusted numbers should know that before trying. On the extended grids of projection (ii) the linear fit is adequate for one arm and fails for two (the R^2 of 0.394 and 0.222 above), which are therefore read by local interpolation between the two bracketing converged cells rather than from the fitted slope (a fallback fixed before the runs, not chosen after seeing the fits).

The within-arm baseline. Whether "update geometry" is the right unit of the claim depends on how far EV moves when only the base learning rate changes inside one geometry. Across the three converged rates in each arm's final sweep, the widest within-arm EV range is 0.0175 (Euclidean, 13.9M), 0.0231 (spectral, 44.8M) and 0.0356 (Euclidean, 60.5M): half to two-thirds of the cross-geometry gap in the same run. We report this because it is not small. It does not make the unit wrong, for a reason visible in the losses: the within-arm excursion is bought with loss (Euclidean at 13.9M spans 0.0674 nats across those three rates, $2.6\times$ the 0.0263-nat spread between arms at the matched-loss cells) and moving an arm off its operating point is exactly the manipulation the projection prices. The cross-geometry gap survives that projection; the within-arm spread is what it corrects for. Neither readout compares an arm at its operating point against an arm displaced from its own.

The training-budget confound, in full. All three headline runs at 13.9M, 44.8M and 60.5M share a 65.5M-token budget, so tokens per parameter fall (4.71, 1.46, 1.08) exactly as the effect rises, a rank correlation of -1 on which "grows with scale" and "grows with undertraining" cannot be told apart. Holding the model at 13.9M and raising the budget to 20.04 tokens per parameter, about the compute-optimal ratio of Hoffmann et al. (2022) and still one epoch, the confound predicts the gap shrinks; it does not, matched loss growing from $+0.0300$ to $+0.0422$ and matched compute flat at $+0.0414$ (two independent cells that coincide), and parity improves with budget ($1.0817 \to 1.0405$), which is why we call the parity failure real but budget-dependent rather than a property of the optimizers. The same test at 60.5M (Appendix L) halves the gap and leaves it floor-clearing; the two together are the basis for §3's statement that undertraining moves the gap in both directions and explains it at neither.

F THE INTERPOLATIVE COMPARISON, WITH THE OVERLAP BUILT

This appendix supports the third correction of §3. The strongest objection to the load-bearing pair is that the Euclidean and spectral loss frontiers do not overlap, so every matched-loss number extrapolates; the linear fit behind the correction above fails for two of three arms; and $\mathrm{dEV}/\mathrm{d}L$ is estimated

by moving the learning rate, so it conflates loss with distance from the optimum by construction. The experiment was pre-registered before any cell ran.

The premise is correct, and here it is measured. Table 9 gives Euclidean's lowest observed converged loss against spectral's highest, per scale: disjoint at every published scale. The last row was already on disk. The isolated family of §2 was built to separate the norm from the recipe, and its arms happen to overlap where the published ones do not, so it is an interpolative comparison we had run months earlier for a different reason: it gives $+0.0369$ $[+0.0357, +0.0381]$ at full seed coverage and $+0.0344$ to $+0.0384$ across its whole band, every point clearing the floor.

Table 9: Euclidean's lowest converged loss against spectral's highest, per scale. The published frontiers are disjoint; the isolated family's overlap.

scale	Euclidean min	spectral max	overlap
13.9M	1.5832	1.5702	none, disjoint by 0.0130
44.8M	1.4820	1.4771	none, disjoint by 0.0049
60.5M	3.8023	3.7745	none, disjoint by 0.0278
13.9M isolated family	1.5444	1.5749	$+0.0306$

Building an overlap at the published scale. Euclidean cannot be pushed below its own optimum, so the overlap has to come from the spectral side: 40 new cells at rates far enough from spectral's optimum for its converged loss to land in Euclidean's band. Which band matters. A comparison at loss L needs both arms interpolable at L for the same seed, and Euclidean's three rates give each seed two adjacent segments: the range all eight seeds can bracket is $[1.5919, 1.6472]$, narrower than the observed $[1.5832, 1.6668]$. Targeting the wider range would have spent eight seeds at 1.72, where no Euclidean seed can be compared at all. The estimator is deliberately unwilling: within one arm and one seed it interpolates EV linearly in loss between two adjacent rates and returns nothing rather than extrapolate past the observed range; it will not bracket across the loss optimum, where the same loss occurs twice; it skips segments with no loss difference; and a target supported by fewer than three seeds yields no verdict. Those refusals are the difference between a measurement and the fit it replaces, and each is a regression test.

The result. At loss matched seed by seed, with no fit, no slope and no extrapolation anywhere, the gap at the best-supported target ($n = 8$) is $+0.0283$ $[+0.0273, +0.0294]$; across the whole band it ranges from $+0.0281$ to $+0.0288$; 23 of 25 targets have full eight-seed coverage and every target clears the floor (Table 10, Figure 2). The headline target is chosen by seed coverage, a criterion that never inspects EV. The flatness is the strongest part: across a 0.055-nat sweep of matched loss the gap moves by 0.0007. The objection was that EV is a monotone function of model quality and the arms sit at different losses; put them at identical losses, seed by seed, and the gap does not move. The interpolative value differs from the published $+0.0300$ by 0.0017, inside the floor, so the extrapolating correction was not misleading us; reported because it is what the data say.

Table 10: The interpolative comparison at 13.9M.

gap at the best-supported target ($n = 8$)	$\mathbf{+0.0283}$ $[+0.0273, +0.0294]$
range across the whole band	$+0.0281$ to $+0.0288$
targets at full 8-seed coverage	23 of 25
every target clears the floor	yes
published matched-loss gap	$+0.0300$

The two-sided control. Spectral's loss is U-shaped in the rate, so each target loss exists twice, under- and over-shooting the optimum. Twenty-four further cells put spectral at Euclidean's loss from the far side, at rates about $95\times$ apart. Interpolating each branch separately, never across the optimum, EV at matched loss differs by at most 0.0031, side-agnostic at all 25 targets. The intervals do exclude zero, so a small systematic side effect is real; it is a tenth of the gap and sub-floor, which is the situation the absolute floor exists to adjudicate, and we report it as detectable rather than absent. Coverage here is $n = 3$ to 5, not 8, because both branches must bracket the same target for

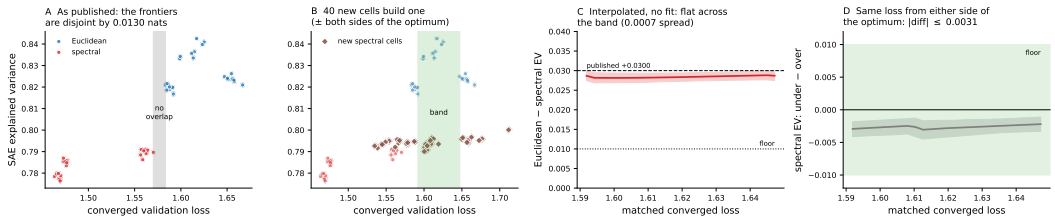

Figure 2: **A**: as published, the two arms' loss frontiers miss each other, so a matched-loss number must extrapolate. **B**: 40 cells build an overlap from the spectral side, at rates below and above its optimum. **C**: the comparison inside that band (interpolated within each arm, never extrapolated, no slope and no shared curve) is flat: across a 0.055-nat sweep of matched loss the gap moves by 0.0007. **D**: the same loss reached from either side of spectral's optimum, 95× apart in learning rate, changes EV by less than a third of the floor.

the same seed and their rate grids do not align per seed. So $\mathrm{dEV}/\mathrm{d}L$ does conflate loss with rate by construction, as objected, and at matched loss the distance from the optimum contributes at most 0.0031 EV against a gap of 0.030: what the matched-loss readout reads is loss.

What this does not do. It is one scale. The other two remain disjoint and their matched-loss numbers rest on the correction of Appendix E; the new cells were not run there because each larger run costs several times a 13.9M one and the objection is answerable at the scale with eight seeds. Nothing in Table 1 changes.

G Explained variance against converged loss: offset curves or one curve?

This appendix supports the second correction of §3. Within-arm learning-rate sensitivity is 50–65% of the cross-geometry effect (Appendix E), so a reader is entitled to ask whether the arms lie on a single curve of EV against converged loss, in which case the effect would be about operating point rather than geometry. The question is sharper than it looks: $\mathrm{dEV}/\mathrm{d}L$ is positive within an arm, so lower loss goes with lower EV, and spectral has both the lowest loss and the lowest EV, a shared curve would predict our ordering with no reference to the update rule.

Test. Within each scale, ordinary least squares with and without an arm term, over every cell carrying both a converged loss and an SAE explained variance (198 cells: three rates per arm, eight seeds at 13.9M and in the isolated family, three at the larger scales). Offsets are bootstrapped by resampling whole seed blocks (Appendix AH), since the three rates are a designed grid rather than exchangeable draws.

Table 11: The spectral arm's offset at equal loss, the fit with and without an arm term, and the sign-like offset. The 44.8M and 60.5M rows rest on three seeds†, so their intervals are observed ranges.

Scale	spectral offset at equal loss	R^2 shared	R^2 offset	sign offset
13.9M	$\mathbf{-0.0376}$ $[-0.0386, -0.0364]$	0.510	0.923	-0.0001
44.8M	$\mathbf{-0.0513}$ $[-0.0528, -0.0495]$	0.577	0.959	-0.0032
60.5M	$\mathbf{-0.0524}$ $[-0.0606, -0.0459]$	0.656	0.889	-0.0165
isolated family	$\mathbf{-0.0390}$ $[-0.0407, -0.0374]$	0.234	0.913	-0.0008

Every spectral offset clears the 0.0100 floor with its interval excluding zero, and one arm term takes R^2 from about 0.5 to about 0.9 (Table 11). The sign-like offset is sub-floor at 13.9M, at 44.8M and in the isolated family: the arms do not simply fan out; spectral separates (Figure 3).

How much loss explains. The arm-mean gap decomposes exactly into a loss part and an arm part, and the loss part must use the within-arm slope (Table 12): loss explains 11–24% of the gap and the

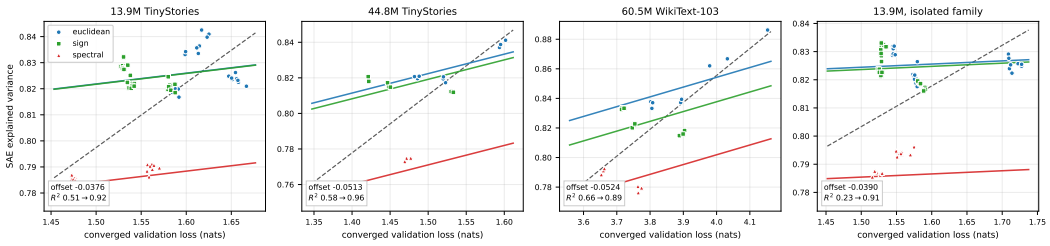

Figure 3: EV against converged validation loss for every cell of every arm. Dashed black is the single shared curve; the colored lines are the offset model, one slope common to the three arms with a per-arm intercept. If the operating point were the whole story the points would lie on the dashed line. Spectral sits below it at every scale, including in the isolated family of Appendix J, where the arms' losses are closest together.

arm offset carries the rest, the two summing to the observed gap to four decimals in every row. The last column is what a shared-curve fit would give, and we report it because it is why the objection looks compelling, and because we first computed the decomposition that way ourselves. It is wrong: the shared slope is fitted through all three arms at once, so it absorbs the between-arm effect it is supposed to separate; using it, the decomposition does not close ($0.0291 + 0.0376 = 0.0667$ against an observed 0.0424 at 13.9M), and that failure to close is what identifies the error. The isolated family is the cleanest case at 3%, because making the arms commensurable in step size also brought their converged losses closer together (0.097 nats apart against 0.117–0.252 for the published arms): the same offset, with less to confound it. Projecting each run onto its own matched-loss reference (Appendix E) shrinks the gap by 3–11%; conditioning on loss across all rates here removes 11–24%. These are different operations: the first corrects a residual mismatch at one operating point, the second removes the whole loss dependence, and they agree in direction and magnitude, which is the check that matters.

Table 12: The gap decomposed into a loss part (within-arm slope times the loss difference) and an arm part; the last column is the naive shared-slope figure that does not close.

Scale	**gap**	within-slope $\times\Delta L$	arm offset	**loss explains**	naive shared
13.9M	$+0.0424$	$+0.0048$	$+0.0376$	**11**$\%$	69%
44.8M	$+0.0647$	$+0.0134$	$+0.0513$	**21**$\%$	62%
60.5M	$+0.0691$	$+0.0167$	$+0.0524$	**24**$\%$	66%
isolated family	$+0.0401$	$+0.0011$	$+0.0390$	**3**$\%$	35%

H THE FULL PAIRWISE COMPARISON

This appendix supports Table 1: all three pairs, both readouts, all four scale runs, with eight seeds at 13.9M and three elsewhere (Table 13). Entries are ΔEV with the half-width of a paired seed-block bootstrap (the unpaired column, and what pairing changed, is in Appendix AH); bold clears the 0.0100 floor. Rows marked † rest on $n = 3$, where a percentile interval cannot exceed the observed seed range: those $\pm$ are observed ranges, not calibrated 95% coverage, and the ± 0.0007 in the last row is the clearest case: three seeds happening to land close together, not a precise measurement.

Eighteen of 24 clear the floor, and 16 of 24 also survive a Holm correction over all 75 comparisons in the paper (Appendix C); the two that do not are the 60.5M and 210M Euclidean$-$sign cells, which we do not claim. Every comparison involving the spectral arm clears in every run under both readouts (16 of 16); the six that do not are all Euclidean$-$sign-like, which is the paper's position: those two arms are not distinguishable at this resolution and the claim concerns spectral. One consequence for Table 14: at 13.9M the Euclidean$-$sign-like interval includes zero (-0.0014 ± 0.0016), so that run's EV ordering is not the exact reverse of its loss ordering, and the middle pair is unresolved in three of four runs. The inversion we claim is spectral's own movement, first by loss and last by EV, not a full three-way reversal.

Table 13: Every pairwise difference at the final residual stream. Bold clears the floor; † marks $n = 3$.

Run	Readout	Eucl. − sign	Eucl. − spectral	sign − spectral
13.9M	matched loss	-0.0014 ±0.0016	$\mathbf{+0.0300}$ ±0.0013	$\mathbf{+0.0314}$ ±0.0015
	matched compute	-0.0092 ±0.0018	$\mathbf{+0.0414}$ ±0.0014	$\mathbf{+0.0506}$ ±0.0010
44.8M†	matched loss	$+0.0047$ ±0.0013	$\mathbf{+0.0461}$ ±0.0016	$\mathbf{+0.0414}$ ±0.0011
	matched compute	$+0.0015$ ±0.0014	$\mathbf{+0.0692}$ ±0.0007	$\mathbf{+0.0677}$ ±0.0021
60.5M†	matched loss	$\mathbf{+0.0144}$ ±0.0032	$\mathbf{+0.0573}$ ±0.0035	$\mathbf{+0.0429}$ ±0.0012
	matched compute	$+0.0029$ ±0.0023	$\mathbf{+0.0663}$ ±0.0024	$\mathbf{+0.0635}$ ±0.0007
210M†	matched loss	$+0.0046$ ±0.0014	$\mathbf{+0.0544}$ ±0.0011	$\mathbf{+0.0499}$ ±0.0024
	matched compute	$\mathbf{+0.0106}$ ±0.0039	$\mathbf{+0.0722}$ ±0.0007	$\mathbf{+0.0616}$ ±0.0033

I THE SCALE SERIES AND THE RANKING INVERSION

This appendix supports §3's scale statement and §5's inversion. Figure 4 shows the four scale runs' gaps under both readouts; Table 14 ranks the same models by loss and by EV.

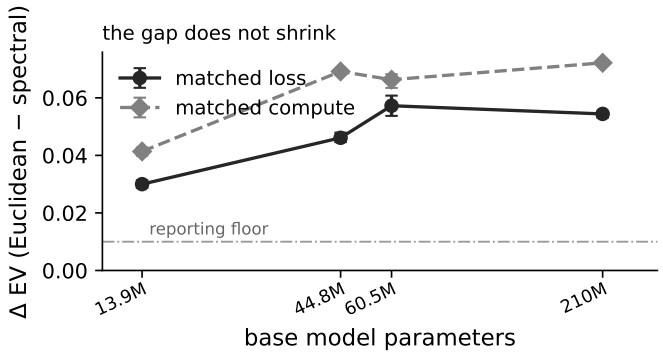

Figure 4: The four scale runs at the final residual stream (the budget runs vary a different axis and are in Table 1). SAE explained variance, Euclidean minus spectral, against base-model size under both readouts; error bars are the paired seed-block bootstrap on the difference, the same intervals as Table 13. Adding the two arms' seed spreads in quadrature would be the formula for two independent samples applied to arms that share a seed, a bar that would be wrong rather than merely different (Appendix AH). The gap exceeds the floor everywhere and does not shrink; only $13.9\text{M} \to 44.8\text{M}$ and $44.8\text{M} \to 210\text{M}$ isolate scale, the second at matched tokens per parameter, and both also change depth, width and head count, which is why no trend is fitted.

The effect does not track manipulation strength. The fraction of parameters under the manipulated rule is non-monotonic across the four runs (76.5, 87.9, 65.0, 95.9%) while the gap runs $+0.0300$, $+0.0461$, $+0.0573$, $+0.0544$, so the gap is not a monotone function of how much of the model the manipulation reaches. This is a control against the obvious alternative reading of the scale series and it survives; it does not identify what the gap does track.

In every run spectral goes from first by loss to third by EV. The Euclidean−spectral gaps at this readout are $+0.0414$, $+0.0692$, $+0.0663$ and $+0.0722$, all clearing the floor; the smallest is the $+0.0414$ that Appendix B lists among the gap's configurations. The three larger runs are $n = 3$ here, as in Table 1.

Why four of four may be one observation rather than four. This is the strongest caveat we hold against our own inversion result. The four runs are not independent replications if the inversion follows from whatever single property makes the spectral arm both lowest-loss and worst-reconstructing at the final site: one cause observed four times would look exactly like this. Four points cannot separate the two, and we do not claim to.

Table 14: The same four models ranked by the objective they were trained on and by the metric used to evaluate their features. Both orderings are read at matched compute, each arm at its own best rate, so both columns rank the same three checkpoints; the loss and the EV in each row come from one model. Taking the loss ordering from each arm's best-achievable loss and the EV ordering at matched loss would rank two different sets of models; the orderings are the same either way, and the EV differences are larger at this readout. The inversion is inherently a matched-compute statement: at matched loss the arms sit at equal loss by construction and there is no loss ordering to invert. The middle pair that does not clear our floor is resolved cell by cell in Table 13.

Run	Best → worst by loss	loss	EV	Best → worst by SAE EV
13.9M TinyStories	spectral	1.4678	0.7782	sign, Euclidean, **spectral**
	sign	1.5309	0.8288	
	Euclidean	1.5877	0.8196	
44.8M TinyStories	spectral	1.3588	0.7512	Euclidean, sign, **spectral**
	sign	1.4217	0.8189	
	Euclidean	1.4854	0.8204	
60.5M WikiText-103	spectral	3.5710	0.7697	Euclidean, sign, **spectral**
	sign	3.7175	0.8332	
	Euclidean	3.8052	0.8360	
210M TinyStories	spectral	1.0972	0.7198	Euclidean, sign, **spectral**
	sign	1.1153	0.7814	
	Euclidean	1.1721	0.7919	

J ISOLATING THE NORM FROM THE REST OF THE RECIPE

This appendix supports §3's statement that the deficit is the norm's, and §4's statement that this family does not re-establish the reversal. Our three arms are heavy-ball SGD, Adam and Muon, and they differ in more than the norm they are steepest under, so a gap between them is not by itself attributable to geometry. Table 15 reads the differences off the optimizers' definitions: Euclidean and spectral differ in three respects, and either differs from sign-like in five. The exact-SVD arm of Appendix K does not close this, since it controls the orthogonalizer and keeps the rest of the Muon pipeline.

Table 15: What the three named optimizers differ in, besides the norm.

Component	Euclidean	sign-like (Adam)	spectral (Muon)
accumulator	heavy-ball	debiased EMA	heavy-ball
Nesterov	no	n/a	**yes**
second moment	none	**yes**	none
update map	identity	elementwise	matrix
shape factor	none	none	$\sqrt{\text{fan}_{\text{out}}/\text{fan}_{\text{in}}}$

What is known about these geometries, and what this paper takes from it. The three arms sit in a literature the main text can only point at. That the optimizer selects the solution, not only the speed of reaching it, is the implicit-bias result: adaptive methods reach solutions unlike gradient descent's (Wilson et al., 2017), and for linear models the selected solution is characterized by the norm the method is steepest under (Gunasekar et al., 2018). For Adam the picture is scoped and not uniform: on homogeneous networks its limiting direction can coincide with gradient descent's (Wang et al., 2021), on separable data it maximizes the ℓ_∞ margin (Zhang et al., 2024a), AdamW solves an ℓ_∞-constrained problem (Xie & Li, 2024), and Adam resists the simplicity bias that SGD shows (Vasudeva et al., 2025); representation dynamics differ by optimizer (Sharon & Dar, 2025), as does the geometry of neural collapse (Zhao et al., 2026), and a recent preprint argues that Adam, unlike gradient descent, sees the coordinate basis of a factored model (Singh, 2026). For spectral descent and Muon the implicit bias is characterized on multiclass separable data (Fan et al., 2025) and on smooth homogeneous networks, where Adam and Muon follow approximate steepest-descent trajectories toward different margin problems (Gronich & Vardi, 2026). Muon's

own lineage runs from spectral descent for Boltzmann machines and deep networks (Carlson et al., 2015a;b) and per-dimension preconditioning (Gupta et al., 2018) through gradient orthogonalization (Tuddenham et al., 2022) to the modular norm and its duality map (Large et al., 2024; Bernstein & Newhouse, 2025); its orthogonalizer is the polar factor $UV^\top$ (Higham, 1986), approximated by Newton–Schulz-type iterations whose numerical study predates deep learning (Kovarik, 1970; Björck & Bowie, 1971) and whose inexactness in Muon is analysed by Shulgin et al. (2026); its convergence (Li & Hong, 2025), its reading as a spectral-norm trust region (Kovalev, 2026) or a spectral-norm constraint (Chen et al., 2026), its equal-rate learning of principal components on imbalanced data (Vasudeva et al., 2026) and its behavior at pretraining scale (Liu et al., 2025) are each studied. On language models specifically, the Adam-over-SGD gap has been traced to sign descent rather than noise (Kunstner et al., 2023), to heavy-tailed class imbalance (Kunstner et al., 2024) and to block-heterogeneous curvature (Zhang et al., 2024b), and a deconstruction of Adam finds signed momentum to recover its performance and stability, with adaptivity on the last layer and the LayerNorm parameters in particular necessary to keep them (Zhao et al., 2024); close to the hybrid our arms are, whose auxiliary group holds the tied embedding and the LayerNorm gains under Adam. None of these results is used as a premise here: the paper measures what the update norm leaves in the representation and reads that with a fixed instrument, and the theory above is the reason to expect the arms to differ at all, not a prediction of the sign or size of what we report.

The isolated family. Three arms in one code path, differing only in the sharp operator of a norm:

$$\text{buf} \leftarrow \mu\,\text{buf} + g, \qquad D \leftarrow g + \mu\,\text{buf}, \qquad U \leftarrow \frac{\Phi(D)}{\|\Phi(D)\|_F}\sqrt{\max(1, m/n)}, \qquad p \leftarrow p - \eta\, U,$$

with $\Phi \in \{\mathrm{id}, \mathrm{sign}, \mathrm{NS5}\}$ the sharp operators of the Frobenius, entrywise-max and spectral norms respectively (Bernstein & Newhouse, 2024); the sign arm is sign descent with momentum (Bernstein et al., 2018), the identity arm normalized SGD with momentum (Cutkosky & Mehta, 2020). Momentum, Nesterov, the vector reduction, the shape factor, weight decay, gradient clipping, the schedule and the auxiliary rule are shared by construction; with Φ neutralized to the identity the three arms are bit-identical over momentum-carrying steps, which is asserted as a test rather than by inspection. The family uses the five-step iteration rather than exact $UV^\top$ because Appendix K already separates those two. Note what the construction costs: under this unit-norm sharp-direction convention $\Phi = \mathrm{id}$ is normalized SGD, not the published Euclidean arm's raw heavy-ball step. Raw gradient descent is Euclidean steepest descent only in the quadratic-penalty form of the problem, where the step carries the dual-norm factor, $\|D\|_F$ here, that the named sign and spectral optimizers drop; the isolated family drops it in every arm. These are therefore new arms, and we report how they sit against the published ones rather than substituting one family for the other.

The unit-Frobenius rescaling is load-bearing. The three maps have natural norms $\|D\|_F$, $\sqrt{mn}$ and about $\sqrt{\min(m, n)}$. Those are constants per shape, and a per-geometry learning-rate sweep absorbs a constant; but only one. A transformer holds several shapes at once, so without the rescaling the arms differ in step size relative across parameters as well as in direction, and no single swept rate can match them. This is a confound in our own published design, and the evidence that the rescaling removes it is that all three isolated arms then select the same rate, $\eta = 0.1$, none at a grid edge, against 0.63, 0.0016 and 0.005 for the published arms, a spread of three orders of magnitude.

Result. 13.9M on TinyStories, eight seeds, the published readout, 72 cells at full budget. Every final-site cell reached L_0 exactly 32.0 with zero dead latents; seed standard deviations are 0.0006–0.0016. The published matched-loss gap is $+0.0300$; isolated, it is $+0.0365$, or 122% of it (Table 16). Removing Nesterov's asymmetry, the second moment, the differing accumulators and the per-shape scaling did not shrink the effect.

Table 16: The isolated family at the final site, eight seeds, paired seed-block intervals.

	ℓ_2- spectral		sign $-$ spectral		ℓ_2- sign	
Readout	Δ	95% CI	Δ	95% CI	Δ	95% CI
matched loss	$+\mathbf{0.0365}$	$[+0.0353, +0.0376]$	$+0.0372$	$[+0.0359, +0.0384]$	-0.0007	$[-0.0021, +0.0007]$
matched compute	$+0.0532$	$[+0.0522, +0.0541]$	$+0.0457$	$[+0.0447, +0.0469]$	$+0.0074$	$[+0.0060, +0.0088]$

The selection rule did not decide it. All 72 cells were trained, so the comparison is available at every rate pair rather than only the selected one, and all 18 pairs are floor-clearing and positive, spanning $+0.0240$ to $+0.0539$ (Table 17). This is not decoration: the matched-loss rule chose $\eta = 1.0$ for the spectral arm over $\eta = 0.032$ on a 0.0009 loss margin, and those sit on opposite sides of its optimum.

Table 17: The gap at every rate pair of the isolated family, eight-seed means; every entry clears the floor. The matched-compute selection is the $(0.1, 0.1)$ cell of each block and the matched-loss selection the $(0.1, 1.0)$ and $(0.316, 1.0)$ cells.

	ℓ_2- spectral				sign $-$ spectral		
$\eta_{\ell_2}\downarrow$	η_{sp}=.032	.1	1.0	$\eta_{sign}\downarrow$	.032	.1	1.0
0.01	+.0402	+.0482	+.0314	0.032	+.0327	+.0407	+.0240
0.032	+.0357	+.0437	+.0269	0.1	+.0378	+.0457	+.0290
0.1	+.0453	+.0532	+.0365	0.316	+.0460	+.0539	+.0372

A cleaner statement than the published family supports. Isolated, the two non-spectral arms are equivalent at the final site: -0.0007 with the interval wholly inside ± 0.0100, our own equivalence criterion, where the published family separates Euclidean and sign-like in few cells (Appendix H). Isolation makes spectral the sole outlier rather than one of three unresolved orderings. The isolated arms land $+0.0110$, $+0.0104$ and $+0.0045$ from their published counterparts, despite sharing no momentum semantics, adaptivity or scaling convention with them.

The depth reversal does not replicate in this family. The first depth pass returned sign-reversing, and it did not survive the dead-latent check; we record the sequence because the first reading was wrong. At matched loss the selection put the sign-like arm at 0.316 and the spectral arm at 1.0, both above their optima of 0.1: an artifact of having made the arms commensurable, which flattens the loss curves and moves the nearest-to-target rate off the optimum. SAEs on those over-driven models came back with 84–97% of the dictionary dead at layers 1–4, against at most 13.6% anywhere in Appendix X, and the site with the largest dead-fraction gap (layer 1, 16% against 97%) was also the site with the largest apparent effect ($+0.3097$): a training artifact in the shape of a result. The profile is therefore read at both operating points with the 2% exclusion applied. At matched compute every arm sits at 0.1 and the dead fractions collapse (spectral, layer 4: $0.8407 \rightarrow 0.0003$); Table 18 gives that profile. The gap is monotone in depth and crosses zero between layers 3 and 4, which is the published shape, but the negative half clears the floor at no usable site: layer 3 reaches only -0.0085 against a 0.0100 floor, and the two sites that do clear negatively are exactly the two the SAE could not fit. So the isolated family attributes the final-site deficit to the norm and reproduces the depth trend, but does not independently re-establish the floor-clearing sign reversal. What the ablation buys is the attribution of the headline deficit, not a second proof of depth-conditionality.

Table 18: The isolated family's depth profile at matched compute, eight seeds; * marks floor-clearing, and sites where any arm's dictionary is more than 2% dead are excluded.

Layer	ℓ_2- spectral	sign $-$ spectral	**max dead**	
0	+0.0009	+0.0005	0.002	
1	−0.0148*	−0.0001	0.353	EXCLUDED
2	−0.0225*	−0.0062	0.190	EXCLUDED
3	−0.0085	+0.0007	0.013	
4	+0.0081	+0.0122*	0.000	
5	+0.0325*	+0.0347*	0.000	
6	+0.0532*	+0.0457*	0.000	

Two further observations. At layer 1 the isolated sign-like and spectral arms are indistinguishable (-0.0001) and the Euclidean arm is the outlier (0.9666 against 0.9813 and 0.9814), matching the published profile's 0.9037 against 0.9812 and 0.9786; on this reading the early "reversal" is substantially Euclidean being poor early rather than spectral being good early, which is what the embedding-dilution mechanism of Appendix X says. The comparison is sub-floor at the usable layer 3, so it is reported as consistent with the published profile, not as a result. And this family's

intermediate activations are harder to fit than the published arms': even at matched compute, layers 1–2 carry 19–35% dead latents where the published run had under 1%. Whatever makes these arms commensurable in step size also makes their mid-stack representations less amenable to a fixed-k SAE at the published budget; we do not know why, and it bounds what the family can measure away from the final site.

K THE SECOND SPECTRAL INSTANTIATION: EXACT $UV^\top$

This appendix supports §3's statement that an exact orthogonalizer reproduces the deficit. Every other spectral number in this paper comes from one orthogonalizer, so the question is blunt: is "spectral" a geometry class, or is it one quintic? The experiment was pre-registered with four gates and a decision rule for each outcome.

Design. A second instantiation orthogonalizes exactly, as $UV^\top$ from the SVD. Both spectral arms run one shared code path, so momentum, Nesterov, the vector reduction and the fan-in/fan-out factor are identical by construction; with the orthogonalizer neutralized the two arms are bit-identical over momentum-carrying steps. Nothing is retrained: the new arm is compared against the stored Euclidean and spectral rows, and the code change was verified to leave all three existing arms bit-identical first. The matched-loss target is held at the value the original three arms set (1.587722). The learning rate is swept per arm as everywhere, which is load-bearing here specifically: the quintic has no fixed point at 1 ($p(1) = 0.701$, and iterating it oscillates), so it returns $0.71\times$ an already-orthogonal matrix where the exact rule returns it unchanged, and a shared rate would have handed the exact arm a roughly $1.4\times$ larger effective step and let us call the difference geometry.

Table 19: The four pre-registered gates. G2 is stated as a proper equivalence test, the interval must lie entirely inside the floor, because "fails to clear the floor" is absence of evidence and would not support a claim that the instantiations agree. G3 exists because a manipulation that never engaged produces a clean null meaning nothing, and is measured on real gradients mid-training, not on synthetic matrices.

Gate	Result	Value
G1 reproduction	**PASS**	Eucl. − svd = $+0.0323$ $[+0.0310, +0.0337]$ matched loss; $+0.0432$ $[+0.0419, +0.0445]$ matched compute
G2 equivalence	**EQUIVALENT**	spectral − svd = $+0.0023$ $[+0.0008, +0.0038]$ and $+0.0018$ $[+0.0006, +0.0030]$; both CIs wholly inside ± 0.0100
G3 manipulation	**PASS**	median update singular-value ratio 468.4 (NS5) vs 1.0001 (exact), 720 samples on real gradients (singular values, *not* $O^\top O = I$; see below)
G4 sweep validity	**PASS**	optimum interior at 5.01×10^{-3}

Result. All four gates pass (Table 19). Against the stored Euclidean−spectral gaps of $+0.0300$ and $+0.0414$, the exact arm gives $+0.0323$ $[+0.0310, +0.0337]$ and $+0.0432$ $[+0.0419, +0.0445]$; the two spectral arms differ by $+0.0023$ $[+0.0008, +0.0038]$ and $+0.0018$ $[+0.0006, +0.0030]$, both intervals wholly inside ± 0.0100. The pre-registered prediction, that the arm flattening the update spectrum completely would show the deficit at least as strongly as the one that only approximates it, held under both readouts, by a small margin.

Which orthogonality invariant G3 uses. A natural worry is whether the audit is well posed for wide matrices: $O^\top O = I$ characterizes a tall partial isometry while a wide one satisfies $OO^\top = I$ instead, and our MLP carries both orientations ($[4d, d]$ and $[d, 4d]$). The worry is correct about that invariant and does not apply to ours, because G3 does not use it. G3 reads the singular values of the applied update, and for $O = UV^\top$ every nonzero singular value is 1 whichever way the matrix is shaped, so the statistic is orientation-free by construction. Measured on our own shapes, $\max / \min$ singular value is 1.000001 for $[1536, 384]$, $[384, 1536]$ and $[384, 384]$ alike, while $\|O^\top O - I\|_\infty$ is 0 tall and 0.80 wide and $\|OO^\top - I\|_\infty$ is 0.80 tall and 0 wide; either one-sided identity would have been wrong for half our matrices, which is why the audit was not written on one. A unit test pins this at all three orientations.

The loss correction, and which comparison it is worth most for. The arms do not converge to identical losses and $\mathrm{dEV}/\mathrm{d}L$ is positive within each, so the correction of Appendix E is applied with each arm's own slope (Euclidean $+0.0267$, spectral $+0.0916$, exact $+0.0999$). Which pairs admit an interpolative correction is measured, not assumed. The spectral and exact frontiers overlap, $[1.4634, 1.5551]$, so the equivalence comparison carrying G2 is interpolative: sweeping the reference across the whole overlap rather than choosing a point, the difference runs $+0.0013$ to $+0.0006$ and clears the floor at 0 of 5 references; the arms are equivalent everywhere the comparison is defined. The Euclidean frontier is disjoint from both spectral arms', so those corrections necessarily extrapolate; routed through the empty-interval guard of Appendix D and labeled as extrapolated, they give Euclidean−exact $+0.0354$ against Euclidean−spectral $+0.0344$, the exact arm carrying 103% of the quintic's deficit. The $+0.0344$ here is not the $+0.0326$ of Appendix E, and Table 20 says why: they are adjustments of different quantities at different references, this appendix corrects the matched-compute gap, because that is the readout its four-arm grid was run at, at a reference interior to its own frontiers. Both are extrapolative for this pair, both clear the floor, and neither is the paper's headline number. The inherited caveat applies: within-arm loss variation is generated by varying the learning rate, so these slopes absorb both what loss does to EV and what the rate does at fixed geometry.

Table 20: Two loss-adjusted Euclidean−spectral figures, and why they differ.

	readout adjusted	**raw**	**reference**	**adjusted**
§3	matched loss	$+0.0300$	$L = 1.7677$	$+0.0326$
here	matched compute	$+0.0414$	$L = 1.5767$	$+0.0344$

Numerical provenance. Orthogonalization runs in float32 via cuSOLVER's `gesvdj`, PyTorch's CUDA default. Measured on 108 real update matrices from this configuration, its departure from orthogonality is $\max|O^\top O - I| = 1.19 \times 10^{-4}$ (median; worst 1.33×10^{-4}), four orders of magnitude smaller than the $O(1)$ deviation of the arm it is contrasted against. An experiment whose whole claim is exactness should not leave its decomposition to a library default: the nominally more accurate `gesvda` driver was tested and rejected, being more orthogonal on synthetic input but throwing on 0.9% of real gradient matrices, which have many repeated singular values.

A correction, recorded because the paper's numbers changed. This experiment was first run with the new arm trained for $4{,}000$ steps against stored arms trained for $8{,}000$; every comparison drawn from it was confounded and is withdrawn. The cause was a runtime argument defaulting to half the budget the stored runs used, in a script whose other constants were imported precisely so they could not drift. The budget is now derived from the stored training records and a guard refuses any mismatch. The correction reversed every qualitative conclusion: the arms had appeared to differ, and to contradict the pre-registered prediction, purely because one was trained half as long.

L THE 60.5M RUN AT THE 13.9M BUDGET

This appendix supports §3's statement that training length halves the 60.5M gap without removing it, and §4's statement that the mid-stack becomes unfittable there. The budget sweep of Appendix AB raises training at the smallest scale. The 60.5M run is the one that most invited the undertraining objection: 1.08 tokens per parameter, validation loss 3.57–3.81, the largest matched-loss gap in Table 1 and the second corpus. So it was repeated at the 13.9M headline's own budget, 4.71 tokens per parameter ($34{,}800$ steps, 285M tokens), with six predictions and five gates fixed before the run.

Design. Everything is the published protocol: the same shape (eight blocks, width 640, vocabulary $32{,}768$), per-arm grids of five half-decade points centered on the published optima with the grid-edge guard, the sweep at half length, three rates per arm and three seeds in the final stage, both readouts, the depth readout at all nine sites. One registered exception: the published run used the first of WikiText-103's two training shards (59.0M tokens) for 1.1 passes; this one uses both shards, each tokenized on the published run's frozen vocabulary map (the first shard reproduces the published training file byte for byte and the validation set is the published file) for 117.9M tokens and 2.42

passes' worth of random windows, with a gate that no selected cell's final validation loss sits more than 0.01 above its own trajectory minimum.

Gates. Every arm's best loss improves on its published best by far more than the 0.15 nats required: $3.805 \to 3.2715$ (Euclidean), $3.718 \to 3.2371$ (sign-like), $3.571 \to 3.1971$ (spectral); perplexity 36–45 becomes 24–26. Every sweep optimum is interior; Euclidean's and spectral's are the published ones $(1.0, 10^{-2})$, the sign-like optimum is 3.2×10^{-4} on this grid against the published 5.6×10^{-4}, with 10^{-3} taking the best full-length loss by 0.001 nats. No cell's final loss exceeds its trajectory minimum (0.000 excess in all 27: the 2.42 passes did not overfit). Dead latents at the final site are at most 0.02%. The matched-loss selection lands Euclidean and spectral 0.0002 nats apart (3.2715 against 3.2713); the sign-like grid has no cell near the target, so its nearest (3.2382, residual 0.033 nats) is used, and the interpolated reading is given beside it in Table 21.

Table 21: The 60.5M model at 4.71 tokens per parameter against its published self at 1.08. Final residual stream, three seeds† (paired seed bootstrap; the intervals are observed ranges), floor 0.0100. The sign-like matched-loss EV interpolated in loss between its two bracketing cells is 0.7692 (selected cell 0.7707), which moves the sign-like pairs by at most 0.0015 and the Euclidean−spectral pair not at all.

		EV			**ΔEV**, Eucl. − spectral	
readout	**tok/par**	Eucl.	sign	spectral	this run [95% CI]	published
matched loss	4.71	0.7809	0.7707	0.7523	$+\mathbf{0.0286}$ $[+0.0231, +0.0335]$	$+0.0573$
matched compute	4.71	0.7809	0.7808	0.7334	$+\mathbf{0.0475}$ $[+0.0433, +0.0509]$	$+0.0663$

The headline halves and survives (a registered prediction missed). We predicted the matched-loss gap would grow, as it did at 13.9M under $4.3\times$ the training. It halved: $+0.0573 \to +0.0286$ $[+0.0231, +0.0335]$, still $2.9\times$ the floor and clearing it on every seed; at matched compute $+0.0663 \to +0.0475$ $[+0.0433, +0.0509]$. All three matched-loss pairs clear the floor (sign−spectral $+0.0184$, Euclidean−sign $+0.0102$); at matched compute Euclidean and sign-like tie ($+0.0001$) and both sit $+0.047$ above spectral. The inversion holds: the arm with the best loss, spectral at 3.1971, has the worst reconstruction, 0.7334. So at the two ends of the scale series, extending training moves the final-site gap in opposite directions: up by $1.4\times$ at 13.9M on TinyStories, down by half at 60.5M on WikiText-103, and removes it at neither. The published 60.5M figure is, by this measurement, inflated about twofold by undertraining; Table 1 carries both budgets, and the registered outcome map applies: the headline is stated for the budgets tested, not extrapolated.

The mid-stack is not fittable at either readout (two predictions missed, one held). At the published budget the 2% dead-latent rule excluded sites 1–2 at matched loss and 1–6 at matched compute (Appendix AA). At 4.71 tokens per parameter it excludes sites 1–7 at both, the same set under either published form of the rule (worst seed, or seed mean). The spectral arm at its matched-loss rate (0.0316, the grid point above its best) leaves 90–99% of the dictionary dead at sites 1–6 and 7.5% at site 7; at its best rate, 97% at site 1 falling to 10% at site 7; the sign-like arm 37%, 29%, 8% at sites 1–3 at matched loss and $95\% \to 17\%$ across sites 1–7 at its best rate; the Euclidean arm never exceeds 1.4% at any site under either readout. What the rule admits is the embedding ($+0.0034$ at matched loss, inside the floor; $+0.0168$ at matched compute) and the final site. There is no admitted negative site, so the reversal cannot be re-read here and no crossing can be located among admitted sites; the registered interpolation, which uses every site, returns 0.94 against the published 0.52 because the only sign change left is between the excluded site 7 (-0.0296) and the final site, and we report that as the miss it is. The prediction that the most negative admitted gap would be smaller in magnitude than -0.0365 holds only vacuously. The prediction that at least half of the mid-stack sites would be excluded at matched compute held in full (7 of 7), as at every larger scale and at 20.04 tokens per parameter. The excluded raw gaps at matched loss are large and positive ($+0.15$ to $+0.44$ at sites 1–5), which is what a dictionary with 50 live latents against $5{,}000$ reads, and we draw nothing from them.

What the streams look like there (post hoc). Written after the verdict, because the exclusions asked for it: Appendix A's statistics at all nine sites of every selected cell, three seeds. At the

published budget the 60.5M streams were near-isotropic (expected cosine 0.06–0.11 at the middle block, 0.05–0.06 at the final site) and high-rank (0.29–0.37 of the width). At 4.71 tokens per parameter the sign-like and spectral arms' streams at sites 1–7 have moved into the regime of the reference models and past it: the spectral arm at its matched-loss rate has expected cosine 0.94–0.99 at sites 1–5, a mean direction carrying 99% of the energy at site 1, effective rank 0.04–0.13 of the width at sites 1–4, and token norms of 5,100–5,600 against the Euclidean arm's 114–190; the sign-like arm 0.57, 0.55, 0.50 at sites 1–3 with effective rank 0.09; at their own best rates $0.92 \to 0.64$ and $0.95 \to 0.59$ across sites 1–7 (GPT-2 small: 0.46–0.50 at its middle block, 0.81–0.83 at its final site, effective rank 0.003–0.12). The Euclidean arm's streams stay comparatively benign: expected cosine 0.26 at sites 1–2 and ≤ 0.15 from site 3, effective rank 0.24–0.52; though its largest coordinate is now 64–128× the median against 11× at the published budget. Every arm is near-isotropic again at the final site (0.12–0.18; effective rank 0.38–0.46), which is where the headline is read and where no latent is dead. The dead fraction rises with the shared direction's share of the energy: below a quarter (every Euclidean site) the worst site loses 1.4% of the dictionary; near half (the sign-like arm's sites 1–3 at matched loss) 8–37%; above nine tenths, 90–99%. So what the rule excludes is a stream dominated by one shared direction, which the fixed recipe (no input normalization, a decoder bias that moves little in 3,000 steps at 3×10^{-4}) spends its live latents on. Whether a recipe that first subtracts that direction would read these sites is a test we have not run. Two things follow for the rest of the paper: the benign-stream characterization of Appendix A is a statement about the published budgets, and at 4.4× one of them it stops holding in the middle of the stack for two of the three arms, optimizer-conditionally, which we note as an observation and did not predict; and the depth reversal reported for this model at its published budget cannot be re-read at this one, for want of admitted sites rather than for want of sign.

Cost. Fifteen sweep and 27 final-grid runs with the inline readouts and 135 depth fits: 19.9 h of one RTX 4090, three processes sharing the card; the stream statistics, 135 sites, on the local machine's GPU.

M WEIGHT DECAY, AND WHY IT CANNOT BE MATCHED BY ITS COEFFICIENT

This appendix supports the weight-decay control of §3. Decay is applied as $p \leftarrow p(1 - \mathrm{lr} \cdot \mathrm{wd})$, so the per-step shrinkage is the product. With selected rates differing by 398× across arms, a shared coefficient is not a shared manipulation: at $\mathrm{wd} = 0.1$ the cumulative weight factor over 8,000 steps would be 3.7×10^{-227} for Euclidean against 0.28 for sign-like. Matching the pressure $\lambda = \mathrm{lr} \cdot \mathrm{wd}$ instead is better but still not enough, because achieved weight norm is an equilibrium between decay pulling down and gradients pushing up, and the arms' gradient scales differ too: at matched λ the achieved reductions come out 0.63, 0.34 and 0.49, a spread of 1.84.

What works. Calibrate per arm on the achieved reduction, the same principle the paper applies to L_0: sweep λ per arm, select the value landing nearest a common target of 0.50, and run the finals at that value, each arm at its own best rate (the matched-compute readout), decayed and undecayed alike at three seeds† (Table 22). Euclidean−spectral is $+0.0424$ $[+0.0414, +0.0434]$ decayed against $+0.0414$ $[+0.0408, +0.0421]$ undecayed, ratio 1.02 with overlapping intervals: unchanged, not enlarged; the undecayed figure is the published matched-compute gap re-read at those three seeds. Validation losses move by at most 1.6% and the arm ordering is preserved, so training was not disrupted. Two gates fixed in advance both pass: the selected λ is interior to the swept grid (4.5 and 6.0×10^{-4} inside $[3, 10] \times 10^{-4}$), and the achieved reductions agree across arms within 1.15 (spread 1.07).

Table 22: Weight decay calibrated per arm on the achieved weight-norm reduction; each arm at its own best rate, three seeds†.

Arm	selected λ	achieved	EV decayed	EV undecayed
Euclidean	6.0×10^{-4}	0.5065	0.8237	0.8191
sign-like	6.0×10^{-4}	0.4756	0.8255	0.8286
spectral	4.5×10^{-4}	0.5099	0.7814	0.7776

What fails, and why it is the useful part. Coarser calibrations do not work, and the reason is instructive for anyone attempting this comparison. Calibrating at half the run length selects a λ that overshoots by the end, because decay compounds: ratios measured as 0.51, 0.53, 0.47 at $4{,}000$ steps became 0.42, 0.37, 0.35 by $8{,}000$, failing the equality gate at 1.19. And a grid coarser than the arms' actual differences cannot separate them, the three arms respond to λ within 2–12% of each other, so on a $3\times$ grid all three select the same value and the per-arm mechanism never engages. The calibration above uses a $1.33\times$ grid at full length, which is what makes spectral separate from the other two. Runs that failed either gate are excluded from the claim rather than averaged into it.

N BATCH SIZE

This appendix supports the batch-size control of §3. Every headline run trains at batch 32×256, and these update rules are batch-size sensitive, so a batch-size disadvantage for one arm is a live alternative to every claim. Muon's published case is better large-batch scaling than Adam; if that holds here, our small batch handicaps the spectral arm and the deficit is partly a batch artifact, which predicts that it shrinks at a larger batch. That is the direction tested, pre-registered with four branches named before the data (robust, attenuated, amplified, killed).

Design. Batch 128×256 for $2{,}000$ steps: the same 65.5M training tokens in a different grouping, four times fewer optimizer steps, the cosine schedule compressed with them. Learning rates were re-swept per arm on six-point grids centered above the published optima (18 runs, one seed; every best rate interior), then the published final-stage construction applied: each arm's best rate plus the neighbour bracketing the common matched-loss target, eight seeds, the SAE fitted inline (48 cells). One control had to be built first: the training loop drew its validation batches at the training batch size, so a batch sweep would have silently changed the loss target and the SAE's activation pool; a separate validation batch size, defaulting to the training batch size so the published path is bit-identical, holds the validation set at the published $20 \times 32 \times 256$.

The sweep answers half the question before any dictionary is fitted. At matched tokens and a quarter of the steps every arm reaches a worse loss, and the Euclidean arm pays three times what the spectral arm pays: $+0.219$, $+0.141$ and $+0.078$ nats over the published optima for Euclidean, sign-like and spectral. Muon's large-batch tolerance is real at 13.9M and shows up in the loss. The matched-loss target is therefore Euclidean's 1.8265, and reaching it reads the spectral arm $16\times$ below its own best rate, a heavier handicap than the published 1.07 parity.

Table 23: The gap at batch 128×256 against the published batch 32×256, eight seeds, both readouts.

readout	**pair**	batch 128	95% CI	batch 32	
matched loss	Eucl. − spectral	$\mathbf{+0.0373}$	$[+0.0343, +0.0402]$	$+0.0300$	ROBUST
	sign − spectral	$+0.0244$	$[+0.0234, +0.0255]$	$+0.0314$	
	Eucl. − sign	$+0.0129$	$[+0.0102, +0.0157]$	-0.0014	
matched compute	Eucl. − spectral	$\mathbf{+0.0432}$	$[+0.0403, +0.0461]$	$+0.0414$	
	sign − spectral	$+0.0445$	$[+0.0434, +0.0457]$	$+0.0506$	
	Eucl. − sign	-0.0014	$[-0.0039, +0.0013]$	-0.0092	

Robust. The headline gap at $4\times$ the batch is $+0.0373$ against $+0.0300$, within the floor of the published value and larger, not smaller; matched compute gives $+0.0432$ against $+0.0414$ (Table 23). Every spectral comparison clears the floor under both readouts with no diverged cell and no dead latents. The prediction that a larger batch would shrink the gap by relieving a spectral handicap was tested in the direction that would hurt and did not happen: the spectral arm is less handicapped by the batch in loss, and reconstructs worse by the same margin regardless. Absolute EV is higher in every arm (0.82–0.87 against 0.79–0.82), because every arm is less trained and $\mathrm{d}\mathrm{EV}/\mathrm{d}L > 0$ within arms; the gap between them is what survives. The middle pair does what it does everywhere else in this paper: clears at one readout and not the other.

Scope. One alternative batch, $4\times$ in one direction, at 13.9M, at the final site only; the depth profile was not re-run. A smaller batch is predicted by the same logic to enlarge the gap and was not tested,

because a confirmation in the easy direction would establish nothing. The loss-side finding (that the spectral arm tolerates the larger batch best) is consistent with Ruan et al. (2026)'s framing and with the report that Muon retains data efficiency at large batch sizes beyond the critical batch size (AI et al., 2025), and is reported as an observation, not a claim: it rests on one seed per rate at the sweep stage and eight at two rates per arm.

O THE ACTIVATION POOL

This appendix supports the activation-budget control of §3. Our SAEs see $163{,}840$ activations redrawn for about 79 effective epochs, where production SAEs see far more, single-pass; if the effect were an artifact of a small pool seen many times, the warning would not transfer. Pre-registered with four gates.

The manipulation is pool size, not SAE compute. SAE steps stay at $3{,}000$ with batch $4{,}096$, so the trainer always draws 12.29M samples; growing the pool therefore lowers the epoch count, which is the axis production differs on. The evaluation slice is byte-identical in every condition (the paper's own $8{,}192$ held-out validation activations, asserted equal in code rather than assumed) so a change in EV cannot come from a change in what EV was measured on. Only the training pool grows, and the pools are nested prefixes (Table 24).

Table 24: The gap against SAE training-pool size, eight seeds per cell, 72 fits.

pool	**activations**	**epochs**	Eucl.	sign	spectral	**Eucl. − spectral**
$1\times$	155,648	78.95	0.8140	0.8229	0.7710	$+0.0430$ $[+0.0416, +0.0445]$
$10\times$	1,556,480	7.89	0.8331	0.8409	0.7924	$+0.0408$ $[+0.0397, +0.0419]$
$\mathbf{100\times}$	**15,564,800**	**0.79**	0.8352	0.8427	0.7946	$\mathbf{+0.0405}$ $[+0.0394, +0.0416]$

Gates and result. G1, the source control: the $1\times$ train-sourced gap is $+0.0430$ against the published validation-sourced $+0.0414$, so swapping the activation source moves it by 0.0016, a sixth of the floor, without which the size sweep would be confounded by source. G2: $+0.0405$ at $100\times$, floor-clearing. G3: epochs reach 0.79 and the SAE is not degenerate there; dead fraction 0.0000, achieved L_0 exactly 32.0. G4: each model's held-out slice is constant across its three pool sizes. The pool matters, and the gap does not move with it: absolute EV rises by about 0.021 for every arm between $1\times$ and $100\times$, so a hundredfold pool really does train a better SAE, while the gap changes by $+0.0025$ $[+0.0005, +0.0045]$, which does not clear the floor. Reporting only the differences would have concealed that the budget matters a great deal for everything except the quantity we are measuring.

Why the pool is drawn from the training corpus. The batch sampler draws overlapping windows, so more batches is not more unique tokens: a $100\times$ pool taken from the 2M-token validation split would be $8.19\times$ duplicated and would measure re-reading, the thing this experiment exists to escape. Against the 60M-token training split it is $1.14\times$ duplicated. The base model saw about 25.2M of those tokens, but the batcher's seed depends only on the run seed and never on geometry, so for a given seed all three arms are exposed to byte-identical token sequences: the exposure is shared and cannot differentiate the arms. The SAE's evaluation slice is drawn from the validation split throughout, so nothing is ever scored on data the SAE trained on. What remains untested is site count, all of this is one site, and the step from 10^7 activations to the 10^8–10^9 of the largest production runs.

P THE SAE HYPERPARAMETER GRID

This appendix supports the SAE-hyperparameter control of §3. We call a shared base-model rate indefensible in §2 and then fit SAEs at a shared rate, which is defensible only because it was checked.

Two SAE learning rates appear in this paper. The published readout (Table 1, the depth sweeps and every headline number) uses 3×10^{-4}. The grid below sits at 10^{-3}, because 10^{-3} is what the

per-arm selection stage chose, in all three arms; it is the selected setting, not a fixed published one, and the distinction matters because without it a reader cannot tell which rate produced Table 1. It also explains why the grid quotes $+0.0320$ where Table 1 quotes $+0.0300$: it would be natural to attribute the whole 0.0020 to the change of evaluation slice, and Table 25 shows that is wrong, about 86% is the SAE learning rate and 14% the slice. Neither value changes a verdict, and the 3×10^{-4} cell agrees with Table 1 to 0.0003.

Table 25: Why the grid's reference cell reads $+0.0320$ where Table 1 reads $+0.0300$.

	configuration	**gap**	**attributable to**
Table 1	lr 3×10^{-4}, single held-out slice	$+0.0300$	—
this grid	lr 3×10^{-4}, select-and-report slices	$+0.0303$	slice $+0.0003$
this grid	lr 10^{-3}, select-and-report slices	$+0.0320$	**rate** $+0.0017$

The SAE learning-rate sweep, which selects the rate and bounds under-tuning. Five rates, three arms, eight seeds (Table 26). The arms differ by $11.65\times$ in mean activation standard deviation, and yet, tuned per arm on one held-out slice and reported on a second, all three select the same SAE rate. The gap is floor-clearing at all five rates and rises monotonically with the rate, which is what makes the under-tuning direction unambiguous: an SAE trained too weakly understates the effect ($+0.0174$ at the lowest rate against $+0.0320$ at the selected one). The 3.16×10^{-4} row ($+0.0304$) brackets the published readout's 3×10^{-4} and agrees with Table 1's $+0.0300$ to 0.0004. The readout is also converged in SAE steps: $|\mathrm{EV}(3000) - \mathrm{EV}(12000)| \leq 0.0011$, so the 3,000-step budget used throughout is not truncating one arm more than another.

Table 26: The SAE learning-rate sweep at 13.9M, three arms, eight seeds.

SAE lr	Eucl.	sign	spectral	**eucl−spec**	**95% CI**	
3.16×10^{-5}	0.7145	0.7204	0.6971	$\mathbf{+0.0174}$	$[+0.0160, +0.0189]$	*lowest*
10^{-4}	0.7849	0.7888	0.7599	$\mathbf{+0.0250}$	$[+0.0235, +0.0266]$	
3.16×10^{-4}	0.8170	0.8184	0.7866	$\mathbf{+0.0304}$	$[+0.0292, +0.0318]$	
10^{-3}	0.8259	0.8267	0.7938	$\mathbf{+0.0320}$	$[+0.0308, +0.0332]$	**selected**
3.16×10^{-3}	0.8248	0.8260	0.7921	$\mathbf{+0.0326}$	$[+0.0315, +0.0339]$	

The grid. Every cell is three arms times eight seeds, selected on one held-out slice and reported on a second disjoint one, varying one factor at a time around the grid's reference point (expansion 8, $k = 32$, SAE rate 10^{-3}, no per-arm standardization). All eight configurations clear the floor and the whole range is $+0.0284$ to $+0.0348$ (Table 27). The structure is worth stating rather than hiding in a range: the gap shrinks as the SAE is given more capacity, monotonically from $k = 16$ upward, $+0.0348$, $+0.0320$, $+0.0284$ at $k = 16$, 32, 64 ($k = 8$ gives $+0.0342$), and $+0.0326$ at expansion 4 down to $+0.0309$ at expansion 16. A larger dictionary at looser sparsity closes some of the difference without eliminating it, which is consistent with the deficit being a property of the representation that a richer code can partly but not wholly absorb; per-arm standardization moves the gap by $+0.0013$. Appendix U gives an account of why this curve has the shape it does, tested against this table rather than fitted to it: the table predates the mechanism by four months.

Q Other SAE architectures

This appendix supports §3's architecture control and §5's statement that a shared hyperparameter is not a shared setting. The main text reads the effect through TopK; here it is read through four other sparse-autoencoder architectures: JumpReLU, and three each differing from TopK in exactly one respect: BatchTopK, Gated and Matryoshka. All are at 13.9M, the final residual stream, the published activation set; one scale, one site, one SAE seed.

Table 27: The SAE hyperparameter grid at 13.9M, matched loss, eight seeds; one factor varied at a time around the reference configuration.

exp.	k	**SAE lr**	**std.**	Eucl.	spectral	**gap**	**95% CI**
4	32	10^{-3}	no	0.8173	0.7847	$+0.0326$	$[+0.0314, +0.0339]$
8	32	10^{-3}	no	0.8259	0.7938	$+0.0320$	$[+0.0308, +0.0332]$
16	32	10^{-3}	no	0.8292	0.7982	$+0.0309$	$[+0.0297, +0.0323]$
8	8	10^{-3}	no	0.7373	0.7030	$+0.0342$	$[+0.0329, +0.0357]$
8	16	10^{-3}	no	0.7835	0.7488	$+0.0348$	$[+0.0335, +0.0361]$
8	64	10^{-3}	no	0.8704	0.8420	$+0.0284$	$[+0.0271, +0.0298]$
8	32	10^{-3}	**yes**	0.8244	0.7910	$+0.0333$	$[+0.0322, +0.0345]$
8	32	3×10^{-4}	no	0.8161	0.7858	$+0.0303$	$[+0.0291, +0.0316]$

Q.1 BatchTopK, Gated and Matryoshka

Design, fixed before any fit. The three architectures were added to the one SAE implementation so that each shares everything else with TopK, the tied initialization included. BatchTopK (Bussmann et al., 2024) changes the selection: the $k \cdot B$ largest activations of a batch rather than the k largest of each token, so L_0 is k on average and varies per token. Gated (Rajamanoharan et al., 2024) changes the activation and the objective: a Heaviside gate times a separately scaled magnitude, an L1 penalty on the gate's pre-activation, and an auxiliary reconstruction through a frozen copy of the decoder. Matryoshka (Bussmann et al., 2025) changes the objective alone: the reconstruction loss summed over nested dictionary prefixes of $1/16$, $1/8$, $1/4$, $1/2$ and all of the latents, with TopK selection over the whole. No auxiliary dead-latent loss anywhere, because the TopK arm has none. Everything else is the published recipe: the frozen validation batches, 200k tokens, expansion 8, 3,000 Adam steps at 3×10^{-4}, batch 4,096, one SAE seed, unit-norm decoder columns, the same 8,192-token held-out slice, eight base seeds, seed-pair intervals with the 0.0100 floor inside the routine. BatchTopK and Matryoshka fix $L_0 = 32$ by construction and are read at both readouts. Gated's L_0 is an outcome of its L1 weight, so it follows the JumpReLU protocol: one weight grid in every arm, EV interpolated per seed at matched achieved L_0 inside the common window. Two choices made it read cleanly. Gated inputs are scaled by a per-arm scalar so that $\mathbb{E}\|x\|^2 = d$, which is what makes one L1 weight mean the same thing across the $11.65\times$ activation-scale spread (EV, L_0 and the dead fraction are invariant to that scalar); and the five-point grid was cut from a calibration on seed 0 by a rule that reads achieved L_0 only, with EV never consulted. Four predictions were fixed before the first fit: the Euclidean−spectral sign replicates under all three at matched loss; BatchTopK lands within 25% of TopK at both readouts, since the selection differs only in aggregation; Gated is at least $1.5\times$ TopK, as JumpReLU was; sign−spectral replicates under all three.

Result. Every gate passed: $L_0 = 32.00$ in every fixed-L_0 cell, dead latents at most 0.85%, Gated's common window $[12.85, 65.59]$ at matched loss and $[12.88, 62.37]$ at matched compute, both containing 32, so the registered target is read directly rather than substituted; and the stored TopK SAEs re-read on the same slice reproduce the published $+0.0300$ and $+0.0414$ to four decimals, which is what makes the ratios in Table 28 comparable. All six blocks replicate. Three predictions held: BatchTopK is within 4% of TopK at both readouts, and its own threshold inference (achieved $L_0 = 32.2$–32.6) agrees with the batch reading on every class, the gaps within 0.0004; Matryoshka lowers every arm's absolute EV by 0.005–0.008, as a nested objective must, and leaves the gap intact; Gated holds its sign across the whole window ($+0.0288$ at $L_0 = 12.85$ to $+0.0225$ at 65.59, matched loss), and the three arms' frontiers coincide in achieved L_0 to within 3–8% at every weight, calibrating the input does what calibrating a threshold did not. One prediction missed: Gated is $1.11\times$ TopK, not $1.5\times$ or more, its difference from TopK $+0.0031$ $[+0.0012, +0.0052]$: real, and a tenth of what JumpReLU added under its published protocol. What that said at the time was that a gate-type objective with an L1 penalty does not double the gap, so the doubling was not the objective's; the refit below then found the candidate left standing.

Table 28: Explained-variance gaps at the 13.9M final residual stream under three further SAE architectures, eight base seeds, seed-pair intervals on the headline pair; "loss" and "compute" are the two readouts. Every Euclidean−spectral and sign−spectral entry is positive, clears the 0.0100 floor and has its interval off zero; every Euclidean−sign entry is sub-floor, as under TopK. BatchTopK and Matryoshka at $L_0 = 32$ exactly; Gated at matched achieved $L_0 = 32$ inside the common window. The last column is the ratio to the TopK gap read on the same slice and seeds.

architecture	readout	Eucl.−spectral	sign−spectral	Eucl.−sign	vs. TopK
BatchTopK	loss	$+0.0287$ $[+0.0273, +0.0299]$	$+0.0311$	-0.0024	$0.96\times$
BatchTopK	compute	$+0.0411$ $[+0.0392, +0.0429]$	$+0.0502$	-0.0091	$0.99\times$
Matryoshka	loss	$+0.0317$ $[+0.0305, +0.0327]$	$+0.0325$	-0.0008	$1.06\times$
Matryoshka	compute	$+0.0439$ $[+0.0422, +0.0453]$	$+0.0507$	-0.0069	$1.06\times$
Gated	loss	$+0.0331$ $[+0.0308, +0.0357]$	$+0.0365$	-0.0034	$1.11\times$
Gated	compute	$+0.0452$ $[+0.0422, +0.0481]$	$+0.0529$	-0.0077	$1.09\times$
TopK, same slice	loss	$+0.0300$	$+0.0313$	-0.0013	—
TopK, same slice	compute	$+0.0414$	$+0.0506$	-0.0092	—

Q.2 The JumpReLU frontier, and the units of the sparsity penalty

The published protocol. The 13.9M matched-loss checkpoints, three arms at four seeds (fewer than the eight used elsewhere) on the published activation set. JumpReLU replaces TopK; the swept knob is the target initial L_0, with the threshold θ calibrated to each arm's own pre-activation quantile so the arms begin at matched sparsity by construction. Because achieved L_0 diverges across arms, a matched-sparsity comparison exists only inside the overlap window $L_0 \in [13.00, 16.25]$; across it the gap runs $+0.0808$, $+0.0844$, $+0.0877$ at the low edge, midpoint and high edge, each an interpolation of the per-arm seed-mean frontiers and each far above the floor. The $2.4\times$ of §5 is read at the midpoint. (The released result file also carries $+0.0871$ with Cohen's $d = 64.1$; that is a superseded reading at $L_0 = 15.85$, at 88% of the window, which our own interior-readout guard exists to prevent; it is kept in the file and not quoted.) We hit the shared-hyperparameter error ourselves first: our first two sparsity grids gave achieved-L_0 ranges that did not overlap across arms at all, and comparing those EVs would have produced a large, confident, entirely artifactual effect; the window is what replaced them. And one thing about the published reading was not checked until the refit below and is now on the record: the Euclidean cell that brackets the midpoint from below (target initial $L_0 = 16$) has 2.1–2.2% dead latents on three of four seeds, so under the 2% rule applied everywhere else the $+0.0844$ would be shown and not read. The refit is read under that rule; the published number is left as it was quoted.

The $2.4\times$ was the protocol, not the architecture. $+0.0844$ is about $2.4\times$ the TopK gap at comparable sparsity ($+0.0348$ at $k = 16$, Appendix P), and the Gated result above showed the doubling is not the objective's. What the published protocol leaves unshared is the objective's units. The threshold is calibrated per arm, but the L_0 penalty is one coefficient against a reconstruction error in each arm's own units, and the arms' activation norms are 41.7, 10.2 and 3.6, so the same coefficient weighs $136\times$ more, relative to reconstruction, in the spectral arm than in the Euclidean one. That is why an identical grid lands the arms at mean achieved L_0 of 32.83, 12.35 and 8.05 (the $4.08\times$ of §5, between $3.91\times$ and $4.26\times$ at all seven settings) and why the only overlap pairs Euclidean's sparsest, most-penalized cells with spectral's densest. The test, fixed before any fit: the same activations, seeds, grid, steps and rate under three protocols on one card; the published one, refitted; the inputs scaled per arm to $\mathbb{E}\|x\|^2 = d$ (the Gated scaling above) with θ still per arm; and the same scaling with one θ and one straight-through bandwidth per seed and grid point, calibrated on the three arms' pooled slices, each read at matched achieved L_0 against a TopK reference ($k = 8$–64) fitted on the same slice under the same scaling, with the ratio classified as calibration at $\leq 1.25\times$ and architecture at $\geq 1.75\times$. Five predictions: both scaled protocols read calibration; the achieved-L_0 spread under the shared threshold is $\leq 1.25\times$; the sign holds throughout; the published protocol reproduces on the new card. All five held (Table 29).

Scaling the inputs collapses the spread from $4.08\times$ to $1.07\times$ (the arms now sit at 21.9, 21.6 and 23.1 with the threshold per arm, and at 21.8, 21.6 and 23.1 with one threshold) and the gap at the published midpoint from $+0.0844$ to $+0.0298$ $[+0.0294, +0.0302]$, $0.90\times$ the TopK gap on the

Table 29: JumpReLU under three protocols, 13.9M, matched loss, four seeds, the same activations and grid, one card. Spread: ratio of grid-mean achieved L_0 between the extreme arms (windows $[13.2, 15.8]$, $[12.3, 36.4]$ and $[12.1, 36.3]$); gap: Euclidean−spectral by per-seed interpolation at the published midpoint, seed-paired interval; TopK: the reference at the same L_0 on the same slice and scaling; ratio: gap over TopK's, with intervals by joint resampling $[2.48, 2.71]$, $[0.88, 0.93]$ and $[0.87, 0.94]$. ‡Fails the dead-latent rule (Euclidean's bracketing cell 2.2%). The last column is the ratio at TopK's own k.

protocol	**spread**	**gap at L_0=14.63**	**TopK**	**ratio**	**at 32**
published: unscaled, θ per arm	4.08×	+0.0844 [+0.0829, +0.0857]‡	+0.0326	2.59×	—
scaled, θ per arm	1.07×	+0.0298 [+0.0294, +0.0302]	+0.0330	**0.90**×	0.70×
scaled, one θ	1.07×	+0.0298 [+0.0294, +0.0303]	+0.0330	**0.91**×	0.70×

same slice; sharing the threshold on top of that changes nothing (+0.0298, 0.91×; once the inputs are scaled, per-arm and pooled thresholds differ by under 2% at initialization). The TopK reference itself moves at most 0.0016 under the scaling, so the collapse is JumpReLU's alone, and the published protocol refitted on this card gives +0.0844 against the earlier +0.0845. Read at TopK's own $k = 32$, which the scaled windows now reach, the JumpReLU gap is +0.0207, 0.70× TopK's, and there the JumpReLU frontiers have flattened, 0.05 EV below TopK at the same L_0, because a unit penalty on unit-scale inputs stops admitting features near $L_0 \approx 20$: an under-tuned SAE understates the gap, as Appendix P reports for TopK. The other two pairs say the same: sign-like−spectral goes from 1.4× TopK's to 0.92×, and Euclidean−sign, which the published protocol put at +0.0328 where TopK sees nothing (−0.0042, sub-floor), is sub-floor under both scaled protocols (−0.0038), the unshared units had manufactured a difference between two arms that agree. So the size transfers between architectures once the SAE's settings are shared settings, and the 2.4× was our own instance of the error §5 describes, committed in the one place we had calibrated a threshold and thought that sufficient (Figure 5).

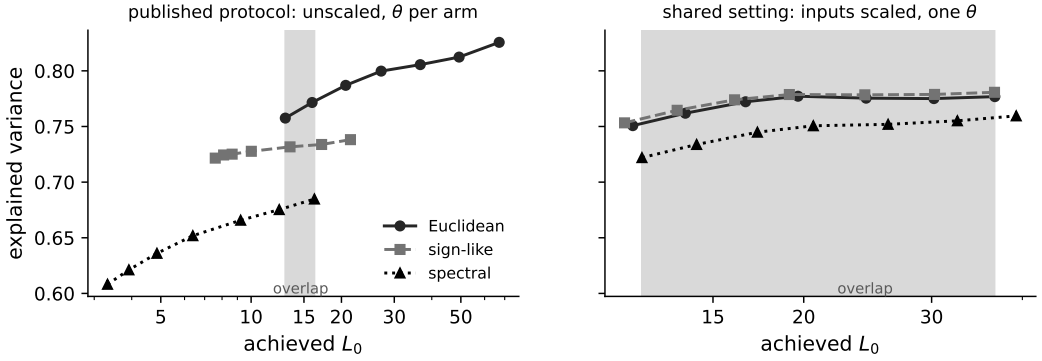

Figure 5: Explained variance against achieved L_0 for all three arms on the 13.9M run, sweeping the JumpReLU sparsity hyperparameter identically in every arm; the shaded band is the overlap window, the only region where a matched-sparsity comparison exists. **Left**, the published protocol: the ordering holds across the whole band, and the horizontal displacement between the curves is the 4.08× achieved-L_0 spread, the same setting places the three arms at visibly different sparsities. **Right**, the same grid with the inputs scaled per arm and one threshold: the arms land on top of each other in L_0 (1.07×), the window spans the grid, and the vertical gap is TopK's (0.9×).

What this does and does not license. Achieved sparsity is not only a reporting matter: Chanin & Garriga-Alonso (2026) find that an L_0 set too low mixes correlated features and one set too high finds degenerate solutions, so two arms read at different achieved L_0 are compared on different feature sets as well as different reconstruction levels. The direction and the fact of floor-clearing transfer across five architectures at both readouts; the magnitude is 0.90–1.11× TopK's for all four once each architecture's sparsity setting is a shared setting, and 2.4× for JumpReLU when it is not. The sign is architecture-independent, and the size is too, on the condition that the SAE's own hyperparameters mean the same thing in every arm. One scale, one site, one SAE seed.

Cost. 289 fits for the three architectures and 348 for the JumpReLU refit, each on one RTX 4090 with six processes sharing the card (1.8 h and 1.0 h of occupancy), and 18 calibration fits on a local machine's GPU (3.4 h); the scaled TopK reference was first fitted unscaled by mistake, noticed from the stored scale factors, and refitted before any number was read.

R THE HYBRID CONTROL SURFACE

This appendix supports §2's auxiliary rule and §6's hybrid concession. The geometry acts on matrix parameters; auxiliary parameters: token and position embeddings, LayerNorm gains, and biases, receive Adam at 3×10^{-4} in every arm, identical in rule and rate. Every arm is therefore a hybrid, and the question is how much of the effect that leaves outside the manipulation. The question has force because in the toy model, failing to hold the auxiliary rule identical reversed the effect's direction (Appendix C). The first half of this appendix is a bound, computed from artifacts already on disk before any decision rule was fixed and stated as exploratory; the second half is a pre-registered counterfactual.

What the manipulation reaches, and what the auxiliary path contributes. Table 30 recomputes the split from the same parameter partition the optimizer uses. The auxiliary weights do move apart: at 13.9M, comparing arms at matched seed, the auxiliary tensors differ by 0.4170 to 0.6040 in relative RMS against 0.8203 to 0.9195 for the matrix tensors, a factor of only 1.48 to 2.16, so the manipulation propagates into the auxiliary parameters, as it must (the rule and rate are identical but the gradients reaching them are not) and both sit below the seed-to-seed floor (0.9841 auxiliary, 1.4137 matrix). What they produce, however, is nearly identical. Layer 0 is the residual stream before any block, and every parameter feeding it is auxiliary under an identical rule at an identical rate; its arm gap is therefore the auxiliary path's own contribution, and it is sub-floor at all three scales, at 0.7%, 6.3% and 13.1% of each run's final-site gap. Ordered by how much auxiliary each run has (12.1, 23.5, 35.0%) the layer-0 gaps ascend with it ($+0.0003$, $+0.0019$, $+0.0075$); three points are not a trend, but rank agreement across all three is what one expects if layer 0 measures the auxiliary path rather than noise.

Table 30: The parameter split, and the layer-0 gap that bounds the auxiliary path's direct contribution.

run	matrix (manipulated)	aux (Adam in every arm)	layer-0 gap
44.8M	87.9%	12.1%	+0.0003
13.9M	76.5%	23.5%	+0.0019
60.5M	65.0%	35.0%	+0.0075

The counterfactual: tying the auxiliary rule to each arm's geometry. A bound is not a test, and the sharper question is what happens when the auxiliary rule is not held identical, since that is what reversed the toy model. This part was pre-registered and required 66 new runs. Tied means one optimizer, the arm's own geometry at the arm's own swept rate, over every parameter: 100% manipulated instead of 65–88%. The rate is re-swept per arm because tying moves each optimum, which is not a formality: routing a vector through an optimizer at the matrix rate is what destroyed a toy model in our own logs. Nothing diverged and no optimum landed at a grid edge. Tying is not free, and it does not hurt the arms equally: best achievable loss moves from 1.5877 to 1.6062 for Euclidean, 1.5309 to 1.4873 for sign-like and 1.4678 to 1.4318 for spectral, costing the Euclidean arm and helping the other two and widening parity from 1.082 to 1.122. The arms become less comparable, not more, which is one reason the published configuration is the one we report.

Table 31: The tied-auxiliary arms at the final site, eight seeds.

readout	Eucl.	spectral	gap	
matched loss	0.8689	0.8051	$+\mathbf{0.0638}$ $[+0.0615, +0.0661]$	vs published $+0.0300$
matched compute	0.8689	0.8774	-0.0085	*not measurable, below*

The direction does not reverse (Table 31). At matched loss the tied gap is $+0.0638$ $[+0.0615, +0.0661]$, floor-clearing with its interval excluding zero and larger than the published $+0.0300$: removing the hybrid entirely does not weaken the effect, and the toy-model reversal does not replicate at transformer scale. Matched compute is not measurable here and we do not read the number it gives: spectral's tied optimum sits at 0.0251, about $16\times$ its matched-loss rate, and at that rate dead latents reach 9.9% mean and 13.3% maximum, so our own 2% exclusion removes the cell; the raw value there is -0.0085, negative but sub-floor, which would not clear on its own terms either, and reporting it as a reversal would be the error this paper exists to criticise. Two caveats on the $+0.0638$: the matched-loss residual is 0.0319 nats, wider than the published parity, because tying pushes the arms' frontiers apart; and the tied arms are a different manipulation, the norm now acting on embeddings too, so this does not license the tied configuration as the better one. It answers the narrower question asked: the published result is not an artifact of leaving 12–35% of parameters outside the manipulation.

S Dose–response on the manipulated fraction

This appendix supports §6's statement that the effect grows monotonically with the fraction of parameters manipulated. It is the cleanest causal test available. Every published arm puts 76.5% of parameters, the 36 attention and MLP matrices, under its geometry; Appendix R put 100% under it and the effect grew; nothing had put less. If the deficit is caused by the norm acting on the weights it should grow with how many weights it acts on; if it belongs to some particular matrix, or to the recipe as a whole, it should not.

Design. A fraction f of the 36 matrices, by cumulative parameter count, receives the spectral geometry at a swept rate; the rest join the auxiliary group under Adam at 3×10^{-4}, the rule they carry in every published arm. Which matrices is drawn per seed, so eight seeds average over placement and the seed spread is the honest spread of "a fraction f" rather than of one arbitrary layout, random matrices rather than blocks, since Appendix X shows block 0 is special. $f \in \{0, 0.25, 0.5, 0.75\}$, with $f = 1$ the published arm. The zero point is Adam at 3×10^{-4} on every parameter, run fresh at eight seeds rather than borrowed from the published sign-like arm whose rate was swept, and it lands at EV 0.8204, indistinguishable from the published Euclidean (0.8196) and sign-like (0.8209) arms. Rates were re-swept per dose (five-point grids, every optimum interior) and each dose is read at the published Euclidean matched-loss target of 1.5877, so the whole curve sits at the loss the published arms were compared at. The Euclidean geometry at $f = 0.5$ is the control for "any dose curve". Pre-registered with four branches (monotone, threshold, non-monotone, flat).

One correction to the readout, made before any result was read. The registered selection took each dose's best rate plus the neighbour whose one-seed sweep loss was nearest the target. At $f = 0.5$ that neighbour was 0.0316, six times the optimum, where the eight seeds scatter four times more widely than any other cell (1.50–1.61) and average 0.039 nats below the target the single seed had hit; read there, the curve showed a spurious rise at $0.5 \to 0.75$. The design said "bracket the target"; that cell did not, so the bracketing cells were added (0.0008 at $f = 0.5$; 0.002 at 0.5 and 0.75) and every dose is read by Appendix F's rule: per seed, linear interpolation between adjacent rates on the low side of the optimum that bracket the target, never across it, never extrapolated. The $f = 1$ endpoint uses the published spectral cells and the low-side ladder (0.00063, 0.00089, 0.00126). The nearest-rate reading is kept beside it so the failure stays visible.

Monotone, and close to linear. EV falls at every step, every step's interval excludes zero, and the total clears the floor by three times (Table 32). A straight line in f fits at $R^2 = 0.79$ with slope -0.030 per unit dose: each quarter of the parameters handed to the spectral norm costs about 0.008 of explained variance, the first quarter slightly less ($+0.0032$ at the zero point's own loss, $+0.0067$ with it adjusted to target). No single step clears the floor on its own; the effect is the accumulation. The three readouts: interpolated, nearest-rate, and matched compute, return the same class. Matched compute falls faster ($+0.0422$ over the curve) because loss improves with dose as well ($1.5696 \to 1.4678$ at the best rates) and $\mathrm{dEV}/\mathrm{d}L > 0$ within arms; the matched-loss column holds loss fixed and is the one the claim rests on. Dosing the Euclidean geometry does nothing: at $f = 0$, 0.5 and 1 the Euclidean curve reads 0.8204, 0.8225, 0.8196, both steps sub-floor. The effect

Table 32: Explained variance at the published matched-loss target against the fraction of matrices under the spectral norm, eight seeds with placement drawn per seed; steps are paired seed-block intervals. *The zero point at its own loss; within-dose $\mathrm{dEV}/\mathrm{d}L$ puts its value at the target at 0.8238.

f	**EV at matched loss**	step	95% CI (paired)	matched compute
0	0.8204 *			0.8204
0.25	0.8172	$+0.0032$	$[+0.0012, +0.0055]$	0.8040
0.50	0.8091	$+0.0081$	$[+0.0037, +0.0125]$	0.7926
0.75	0.8011	$+0.0079$	$[+0.0026, +0.0131]$	0.7857
1	0.7913	$+0.0098$	$[+0.0035, +0.0159]$	0.7782
$0 \to 1$		$+\mathbf{0.0291}$	$[+0.0277, +0.0304]$	$+0.0422$

is not "any geometry replacing Adam"; it is the spectral norm specifically, and it is graded in how much of the network it acts on.

Scope. 13.9M, the final site, one direction of the Euclidean check. Placement is randomized, not resolved: whether some matrices carry more of the effect than others is answerable from the per-seed subsets on disk and was not asked here. Ninety-two cells plus the twenty-four bracketing ones: 5.5 hours of one GPU with two to four processes sharing it, about 12 hours sequential-equivalent.

T THE SPECTRUM-MATCHED NULL, AND A LINEAR BASELINE

This appendix supports §3's statement that activation spread does not carry the deficit and that the deficit is sparse-specific at the final site. It is the record of our own refuted account, and of our disagreement with Ruan et al. (2026), who read higher effective rank as evidence of less superposition; that work's claim is ordinal and ours takes no position on it, and its design already sweeps rates independently and reasons at matched loss, so the disagreement is only about what effective rank licenses. The inference fails here: the spectral arm has the highest effective rank and the worst reconstruction in every run, so occupancy and interference move in opposite directions in a floor-clearing case. Effective rank measures how many directions carry variance, not how cleanly a dictionary separates them.

Setup. The stored 13.9M matched-loss checkpoints, three geometries at eight seeds, layer 6, 163,840 activations from the frozen validation batches, the published readout's activation collection, with the two-holdout separation: trained on activations 16,384 onward, reported on the disjoint slice 8,192–16,383, so the levels here are not Table 1's slice. The SAE is identical in every condition (TopK, expansion 8, $k = 32$, 3,000 steps, one SAE seed).

The surrogates. For each arm's activations take the mean μ and covariance $\Sigma = V\Lambda V^\top$. The synth condition draws $z \sim \mathcal{N}(0, I)$ and returns $z(V\Lambda^{1/2}V^\top)^\top + \mu$, which matches μ and Σ exactly while destroying every higher-order moment; the white condition returns $(x-\mu)V\Lambda^{-1/2}V^\top$, whose covariance is the identity, preserving higher-order structure while removing the spectrum (verified numerically: the whitened covariance departs from I by at most 2×10^{-9}, and the synthetic covariance matches to sampling error). Each manipulation alone is weak (a surrogate that keeps the spectrum and loses the effect could fail for an unrelated reason, and whitening that keeps the effect could be creating it) but run together they point in opposite directions by construction, so agreement between them is informative in a way either alone is not. We pre-registered that a disagreement would be reported as such rather than resolved by preferring one.

Result. Effective rank is a function of the covariance spectrum, so the synth condition carries the same effective-rank difference as the real activations (at seed 0 the Euclidean$-$spectral difference is -6.03 real and -6.16 spectrum-matched) while the EV gap falls from $+0.0320$ to $+0.0057$: the same effective-rank difference, 18% of the effect (Table 33). Spectrum-matched Gaussians reconstruct far worse than real activations in every arm (about $0.83 \to 0.59$), so the structure SAEs exploit at this width and sparsity is overwhelmingly higher-order; whitened activations sit between (about 0.66) and keep the gap.

Table 33: The spectrum does not carry the effect. Identical SAEs on the same activations, three arms at eight seeds. The last column normalizes each gap by its own condition's headroom, $\Delta/(1-\overline{\mathrm{EV}})$, because the three conditions sit at very different EV levels and comparing absolute differences across them is the error Table 1's caption warns about; under normalization the spectrum reproduces 8% of the effect rather than 18%. One secondary reading does not survive it: whitening leaves the gap 61% larger in absolute EV but 13% smaller relative to headroom, so we claim no masking effect in either direction.

Condition	Eucl.	sign	spectral	**gap**	**95% CI**	**norm.**
real	0.8259	0.8268	0.7939	$+\mathbf{0.0320}$	$[+0.0308, +0.0332]$	17.33%
synth (spectrum kept)	0.5867	0.5927	0.5810	$+0.0057$	$[+0.0044, +0.0070]$	1.38%
white (spectrum destroyed)	0.6795	0.6730	0.6281	$+\mathbf{0.0515}$	$[+0.0501, +0.0528]$	15.15%

A linear baseline: is the sparse code the point? If a linear reconstruction at comparable capacity shows the same gap, the effect is generic compressibility of the activation distribution and the sparse autoencoder is incidental. PCA gives the optimal rank-r least-squares linear reconstruction after centering (Eckart & Young, 1936) and needs no training, so it was run on the published activations of all 24 matched-loss cells, components and mean fitted on the SAE's training pool and scored on its held-out slice. Two anchors, because a gap at EV 0.45 and a gap at EV 0.82 are not the same currency: rank 32, the SAE's per-token active budget; and the rank at which Euclidean's PCA reconstruction first reaches its own SAE EV, so that linear and sparse are compared at the same rank and the same reconstruction level. A dense nonlinear autoencoder ($384 \to 32 \to 384$, ReLU, trained with the SAE's settings) is the third row of Table 34.

Table 34: Linear and dense reconstructions against the sparse one, 13.9M, matched loss.

reconstruction	Eucl.	sign	spectral	**Eucl. − spec.**	of SAE
TopK SAE, $k = 32$ (Table 1)	0.8196	0.8209	0.7896	$+0.0300$	—
PCA, rank 32	0.4564	0.4628	0.4424	$+0.0140$	47%
dense AE, bottleneck 32	0.4515	0.4595	0.4375	$+0.0140$	47%
PCA, rank 135 (Eucl. at its SAE EV)	0.8206	0.8294	0.8145	$+\mathbf{0.0061}$	**20**%

At the same rank and the same reconstruction level as the SAE, a linear code sees $+0.0061$ of the $+0.0300$: sub-floor, and 20% of the effect. That is the same fraction the Gaussian surrogate reports (18%, $+0.0057$) by an entirely different route: PCA on the real activations and an SAE on a Gaussian with the real covariance are both measurements of what second-order structure carries, and they agree to 0.0004. The remaining four fifths is what the sparse code sees and the linear one cannot. At rank 32 the linear gap is larger in absolute terms ($+0.0140$) and smaller per headroom (0.025 against the SAE's 0.167); a nonlinear bottleneck of the same width reconstructs no better than PCA and shows the same gap, so nonlinearity without sparsity buys nothing here. The linear gap peaks at rank 60 ($+0.0187$) and is gone by rank 192; spectral needs 123 linear dimensions to reach its own SAE EV where Euclidean needs 135, so at matched target the spectral activations are if anything slightly more linearly compressible. The effect is a property of the sparse decomposition, at this site; at layer 3 the same two controls say the opposite (Appendix AC).

Effective rank does not predict interference. Across a depth axis the two quantities do not move together: the correlation is near zero with the wrong sign, and the largest layer transition is a floor-clearing counterexample, effective rank falls by 71.6 while interference rises by 5.4. The design excludes only $|\rho| > 0.50$ at 80% power, so this is a bound rather than a null. It is the reason we frame the effect as reconstructability and make no claim about superposition.

U WHERE THE DEFICIT LIVES: TOKEN FREQUENCY

This appendix supports §3's statement of where the deficit sits. It was pre-registered before any cell ran, on the natural mechanism: spectral learns tail content better, a fixed-capacity SAE drops tail content first, so the deficit should concentrate on rare tokens. The premise is not ours: Wang

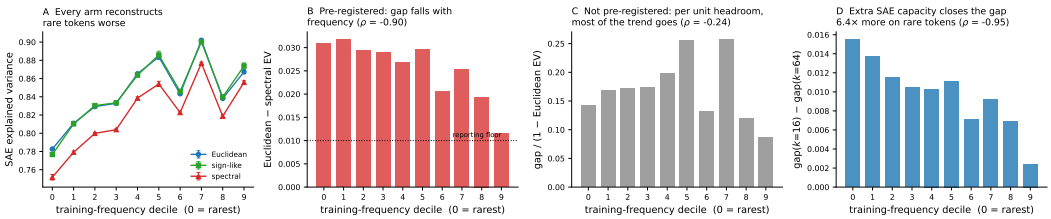

Figure 6: SAE reconstruction stratified by training-set token frequency, 13.9M, eight seeds. **B** is the pre-registered statistic and it is supported. **C** is a control added after seeing B, and it removes most of B: every arm reconstructs rare tokens worse (**A**), so the commonest deciles have less headroom for any gap to occupy. **D** is the difference-in-differences that survives C and is the mechanistic evidence.

et al. (2026) report Muon learning low-frequency classes better than Adam, which is our spectral-versus-sign-like contrast rather than the headline one, though the two behave alike here; on language models, Kunstner et al. (2024) trace Adam's advantage over gradient descent to heavy-tailed class imbalance, sign-based methods being less sensitive to it.

Design. Token frequency from the training split (60M tokens, vocabulary 8,192, counts 26 to 4.45M), because the claim is about training exposure. Ten deciles of equal evaluation mass, so every bucket carries the same n and none is a single rare word. Every evaluation activation keeps the id of the token that produced it. The EV denominator is the global activation mean restricted to the bucket (the paper's own baseline on a subset) not the bucket's own mean, which would give tight buckets an easier baseline and manufacture the trend under test. Numerator and denominator are reported separately for the same reason: across deciles the denominator moves only 558–725 and mean activation norm 18.0–19.6, so what follows is not a denominator effect. Position 0 is excluded throughout, and including it changes no verdict.

The pre-registered result, supported at both readouts. The Euclidean−spectral gap runs from $+0.0310$ on the rarest decile to $+0.0115$ on the commonest (Figure 6), a rare−common difference of $+0.0194$ $[+0.0169, +0.0220]$ with $\rho = -0.90$; at matched compute $+0.0163$ $[+0.0141, +0.0187]$, $\rho = -0.75$. Both clear the floor with intervals excluding zero, and unlike the depth result this one does not depend on the readout.

A control, not pre-registered, that removes most of it. Absolute EV rises with frequency ($0.783 \rightarrow 0.867$), so the gap has less room to occupy at the common end. Expressed per unit of headroom, ρ falls from -0.90 to -0.24 at matched loss and from -0.75 to -0.07 at matched compute; at the second readout, essentially no trend at all. We report this because it cuts against us, and we do not claim it settles the question in the other direction: bucket difficulty is plausibly the mechanism's own pathway, since rare tokens are hard to reconstruct because they are rare, and dividing it out then removes the effect being measured. On the level statistic alone, the two readings are not separable with what we have.

The capacity check, which is separable, and which passes. The level comparison is confounded by how reconstructable a bucket is at all; a difference-in-differences is not. Refitting at $k = 16$ and $k = 64$, the question is not whether the gap is larger on rare tokens but whether extra dictionary capacity closes it more there (Table 35). Extra capacity closes the gap $6.4\times$ more on the rarest decile than the commonest ($\rho = -0.95$), and this survives the headroom control that the level result did not: $3.9\times$ per unit of headroom, $\rho = -0.37$. The stratification stays supported at every capacity, with the frequency concentration itself shrinking as capacity grows: rare−common $+0.0261$, $+0.0194$, $+0.0130$ at $k = 16, 32, 64$.

What this does and does not buy. It explains the curve in Appendix P: the gap narrows with capacity because the capacity is spent disproportionately on the tail the spectral arm loads more heavily, and that grid was recorded before the mechanism was proposed, so the prediction was not fitted to it. It does not explain why spectral loads the tail more heavily, that is Wang et al. (2026)'s

Table 35: The gap by frequency decile at three dictionary capacities, and how much extra capacity closes it, absolute and per unit of headroom.

decile	**gap** k=16	**gap** k=32	**gap** k=64	**closed**	**per headroom**
0 rarest	+0.0378	+0.0310	+0.0223	**+0.0155**	0.0581
1	+0.0370	+0.0318	+0.0233	+0.0137	0.0604
2	+0.0338	+0.0294	+0.0223	+0.0115	0.0561
3	+0.0325	+0.0289	+0.0221	+0.0104	0.0516
4	+0.0309	+0.0267	+0.0206	+0.0103	0.0631
5	+0.0341	+0.0297	+0.0230	+0.0111	0.0791
6	+0.0231	+0.0206	+0.0160	+0.0071	0.0372
7	+0.0292	+0.0252	+0.0201	+0.0092	0.0769
8	+0.0217	+0.0194	+0.0147	+0.0069	0.0349
9 commonest	+0.0117	+0.0115	+0.0092	**+0.0024**	0.0148

question, not one we answer; and it does not identify the higher-order structure the covariance surrogate failed to capture. It is one scale, one corpus, and the level half of the evidence is largely a ceiling effect. We report it as where the deficit sits, with its weakest joint named, not as why.

V A FUNCTIONAL METRIC: DOES THE DEFICIT CHANGE WHAT THE MODEL COMPUTES?

This appendix supports §3's statement that the deficit reaches the model's own loss. Explained variance says how well a dictionary reproduces an activation and nothing about whether the model still computes the same thing, and gains on such proxy metrics do not reliably translate to practical performance (Karvonen et al., 2025). Here the reconstruction is substituted back into the residual stream, the rest of the model is run, and its own loss is measured.

Method. A resume-from-layer forward pass takes a supplied residual stream at a given site. It must be the exact inverse of the function that produced that stream, and this is asserted rather than argued: reconstructing the clean stream and resuming from it is required to be bit-identical to an ordinary forward pass at every layer, checked in the test suite and again before any cell runs, since a drift between the two paths would masquerade as a functional deficit. No training was needed; the eight-seed checkpoints already existed. Three quantities were fixed in advance because the conventions are not neutral: ΔCE = CE(substituted) − CE(clean) in nats, which needs no baseline and no denominator; KL(clean ‖ substituted) per token; and loss recovered, $(\mathrm{CE}_{\mathrm{abl}} - \mathrm{CE}_{\mathrm{sae}})/(\mathrm{CE}_{\mathrm{abl}} - \mathrm{CE}_{\mathrm{clean}})$ against a mean-ablation baseline. ΔCE and KL are unbounded, so the 1%-of-range rule yields no floor for them and intervals alone are reported; the 0.0100 floor applies to loss recovered, which is bounded. Two sites: the final pre-LayerNorm stream where Table 1 is read, and mid-stack (layer 3 of 6 and 4 of 8, the midpoint fixed by arithmetic), where real computation follows the substitution.

Table 36: Euclidean−spectral differences in the model's own loss after substituting each arm's reconstruction, matched loss, eight seeds; positive means spectral is worse. 192 measurements over both readouts, both sites and two scales.

Scale	**Site**	**ΔCE**	**KL**	**LR**	**EV**
13.9M	final	+0.0180 [+0.0167, +0.0193]	+0.0157	+0.0014	+0.0300
13.9M	mid	−0.0288 [−0.0322, −0.0252]	−0.0282	−0.0012	−0.0851
44.8M	final	+0.0240 [+0.0223, +0.0256]	+0.0205	+0.0018	+0.0453
44.8M	mid	−0.0109 [−0.0144, −0.0071]	−0.0141	+0.0041	−0.0620

The deficit is functional. At the final site spectral's reconstruction costs the model 0.018–0.024 more nats of its own cross-entropy, at both scales, under both readouts, every interval excluding zero (Table 36). The depth reversal transfers too: at mid-stack spectral costs less, and KL agrees at every cell. Both halves of the central claim survive a change of metric family.

The conventional metric fails to resolve it. Loss recovered gives differences of $+0.0014$ and $+0.0018$ at the final site, below our own 0.0100 floor, because mean-ablating the final residual stream costs 12.2–13.0 nats and every arm therefore scores about 0.988: real differences are compressed into the fourth decimal by a denominator two orders of magnitude larger than the effect. Reported alone, loss recovered would have said "no functional difference", which is why ΔCE was pre-registered as primary and why every denominator is printed; our own thesis arriving at our own instruments, a normalization chosen by convention deciding the answer. Worse, the denominator is not constant across arms where it matters most: at 44.8M mid-stack the mean-ablation cost is 9.16, 11.60 and 5.81 nats for the three arms, and there loss recovered reports spectral worse by $+0.0041$ while ΔCE and KL report it better by -0.0109 and -0.0141. Three functional metrics, two directions, one cause: the arm with the best reconstruction also has the least to lose from ablation, so dividing by that loss inverts the comparison.

How much independent evidence this is: less than it looks. Over all 192 cells the correlation between ΔCE and EV is $r = -0.905$ at the final site and $r = -0.724$ at mid-stack. The second site carries more independent information, but neither is a separate line of evidence: at the final pre-LayerNorm stream only the LayerNorm and the unembedding follow the substitution, so ΔCE there is close to a monotone restatement of reconstruction error. We therefore claim that the deficit is visible in the model's behavior and not only in a dictionary-fit statistic; not that it has been confirmed by an independent measurement, which would need a downstream task we do not have. 13.9M and 44.8M only: the 60.5M run costs as much as the other two together and would add a scale and corpus point rather than the metric-family point at issue.

W THE LINEAR-PROBE EXPERIMENT

This appendix supports the claim, made in §1 and scoped in §6, that the deficit is not specific to SAEs, and it fixes what kind of "uncontrolled variable" the title refers to.

Confounding variable, not instrument artifact. One may object that a measurement artifact of the SAE would not survive removing the SAE, so a difference that is larger in raw activations cannot be an artifact. That is correct, and it is the distinction. A measurement artifact would mean the metric reports a difference the representation does not have; that is not what we found, and the probes are how we know: strip the dictionary and the difference is still there, slightly larger. A confounding variable means that a difference which is really in the representation is attributable to a factor the comparison did not control. That is what the optimizer is for cross-model SAE comparisons, and the two have opposite remedies: an instrument artifact is fixed by a better instrument, a confounding variable is not. Because the difference survives into raw activations, no SAE improvement removes it; the only remedy is to report and control the optimizer, which is what Appendix C asks for. One thing this does not license is calling one arm's features worse: that needs an external criterion of feature quality we do not have, and "reconstructs worse" is the whole of what we assert.

Design. Linear probes (Alain & Bengio, 2017) are trained on the 13.9M run only, at eight base-model seeds, comparing raw residual-stream activations against SAE codes from the same models, on concepts derivable from the token stream alone. Features are standardized (fitted on the training split only) before fitting, because raw activations and SAE codes differ in scale by roughly $4.6\times$ on top of an $8\times$ dimensionality difference, and a shared regularization constant would otherwise penalize the two spaces unequally rather than compare them.

The result, and why it cuts against our framing. Spectral is worst in both spaces on both non-saturated tasks, 8 of 8 floor-clearing, and the deficit is larger in raw activations than in SAE features: $+0.0227$ against $+0.0177$ on the first task and $+0.0299$ against $+0.0184$ on the second. If the SAE were manufacturing the deficit the ordering would run the other way; instead the phenomenon belongs to the representation, which the SAE inherits and partly compensates for. The raw-versus-SAE differences themselves (0.0050 and 0.0115) are not shown to clear a floor, so the direction is what we rely on and not the size, and this is the thinnest claim in the paper. Costs compound rather than trade off: at matched sparsity the spectral arm spreads usage over 192 more effective features (the usage-entropy perplexity of Appendix B, with no dead features in any arm at this readout) while explaining less variance: a dictionary that is both more spread out and worse at reconstruction is not

trading one against the other. Two caveats bound what this shows: two of the four tasks saturate, so the load-bearing evidence is two comparisons; and because the concepts are token-derivable, the probe measures how linearly recoverable the input is from the residual stream, a quantity related to but not the same as feature quality, and one influenced by the embedding, which receives an identical update rule in every arm by design. We also pre-registered the stronger prediction that a downstream task ranking would flip between raw and SAE feature spaces. It did not: orderings agree on three of four tasks and the single disagreement is 0.002, far below its floor.

X THE DEPTH SWEEP AT 13.9M, AND THE MECHANISM OF THE REVERSAL

This appendix supports §4: the profile behind Figure 1a and Table 1's smallest run, the dead-latent check, and the embedding-dilution mechanism.

Design. Three arms at eight seeds at all seven residual-stream sites of the stored 13.9M matched-loss checkpoints, 168 SAE fits, at the published readout throughout (3,000 steps, 163,840 activations, SAE rate 3×10^{-4}, one SAE seed, the single held-out slice). Only the activations differ between cells. Two things were fixed before the fits: layer 0 is a control (embeddings receive an identical rule in every arm, so the arms should be indistinguishable there) and layer 6 must reproduce Table 1's $+0.0300$, which is the comparability gate.

Table 37: Explained variance by readout depth at 13.9M, eight seeds, with the Euclidean−spectral difference and a seed-level bootstrap. Floor-clearing negative at four sites, positive at one. Layer 0 is the control and the arms are indistinguishable there; layer 6 reproduces Table 1's $+0.0300$ exactly.

Layer	Eucl.	sign	spectral	**eucl−spec**	**95% CI**	**verdict**
0 (embed.)	0.9856	0.9848	0.9837	$+0.0019$	$[+0.0017, +0.0021]$	sub-floor
1	0.9037	0.9812	0.9786	$\mathbf{-0.0749}$	$[-0.0759, -0.0738]$	**negative**
2	0.8674	0.9291	0.9493	$\mathbf{-0.0819}$	$[-0.0885, -0.0765]$	**negative**
3	0.8153	0.8821	0.9006	$\mathbf{-0.0853}$	$[-0.0898, -0.0801]$	**negative**
4	0.8022	0.8510	0.8570	$\mathbf{-0.0547}$	$[-0.0611, -0.0470]$	**negative**
5	0.8054	0.8228	0.8080	-0.0026	$[-0.0061, +0.0005]$	sub-floor
6 (readout)	0.8196	0.8209	0.7896	$\mathbf{+0.0300}$	$[+0.0285, +0.0315]$	**positive**

Result. The two arms' profiles are not parallel (Figure 1a): Euclidean's explained variance falls steeply at the first block and then flattens, spectral's descends gradually and keeps descending, which is why their difference changes sign once rather than merely shrinking (Table 37). The middle pair is worth noting: at layer 6, Euclidean−sign-like is -0.0013 with interval $[-0.0028, +0.0001]$ at $n = 8$, sub-floor and including zero; an independent eight-seed confirmation of the limitation Table 14 records. Dead latents run against the early-layer result rather than producing it: the dead fraction is exactly 0 at layer 6 in all 168 fits, and off the readout layer it reaches 0.14 (sign-like, layer 1) and 0.09 (spectral, layer 2). Dead features remove usable dictionary capacity, so the spectral arm carrying more of them at exactly the sites where it reconstructs better means its early-layer advantage is understated rather than manufactured.

The reversal is embedding dilution; the final-site deficit is not. A mechanical account of Table 37 deserved a measurement rather than an argument. The anomalous early arm is Euclidean, not spectral (0.9037 at layer 1 against 0.9812 and 0.9786), and the early residual stream is mostly the embedding, a lookup table any dictionary reconstructs nearly perfectly; if Euclidean's blocks overwrite it faster, its early-layer EV is low for a reason that has nothing to do with feature geometry. We measured, at every site and every one of the 24 matched-loss cells, the fraction of held-out residual variance that is linearly predictable from the embedding output h_0, $R^2_{\mathrm{emb}}(L)$, with the 384×384 predictor fitted on the SAE's own training pool and scored on its own held-out slice, so the split is the one the EV used (Table 38).

The premise holds, dramatically: Euclidean's first block overwrites 43% of the embedding's linear structure where the other two arms' first blocks overwrite 4–6%, and R^2_{emb} orders the arms the same way EV does at every site ($r = +0.90$ across the 18 site×pair gaps). Pooling all 144 cells, a single quadratic in R^2_{emb} fits EV with $R^2 = 0.990$ (the arms lie on one curve) and the residual

Table 38: The fraction of each site's held-out residual variance linearly predictable from the embedding output, per arm; the EV gap; and the gap that remains at matched R^2_{emb} under the pooled quadratic fit.

	R^2_{emb}				Eucl.−spec. at matched R^2_{emb}	
Layer	Eucl.	sign	spectral	EV gap	residual	dilution explains
1	**0.566**	0.945	0.964	−0.0749	+0.0047	106%
2	0.475	0.649	0.735	−0.0819	−0.0048	94%
3	0.342	0.504	0.557	−0.0853	−0.0012	99%
4	0.306	0.431	0.455	−0.0547	+0.0094	117%
5	0.311	0.365	0.357	−0.0026	**+0.0184**	—
6	0.329	0.337	0.307	+0.0300	**+0.0197**	**34%**

from that curve at each site is what dilution does not account for. At layers 1–4 it is sub-floor: the reversal, the negative region behind "spectral is reconstructed better mid-stack", is embedding dilution. At layer 6 it is not: the three arms have overwritten the embedding to the same degree (0.31–0.34), and 0.0197 of the +0.0300 survives at matched R^2_{emb}, two-thirds of the headline gap. This is a better position than the one it replaces and a less comfortable one. The depth reversal has a mechanism, not feature geometry but how much of the embedding each optimizer's blocks leave standing, a property of the update rule (Euclidean's matched-loss rate is three orders of magnitude larger in nominal units) and not of what the dictionary finds. The phenomenon and the warning are unchanged; the interpretation of the mid-stack sign as a feature-level fact is withdrawn. What is left unexplained is exactly the headline: the final-site deficit persists at matched dilution and is not this. The pre-registered rule returned partial, because the residual clears the floor at two sites; the registered test forced one proportionality across sites whose R^2_{emb} runs from 0.3 to 1.0, and its layer-1 residual was the model's own linearity. The shared-curve reading above was added after seeing that, is labeled post hoc, and lands in the same class with the structure the registered test could not resolve. The transition from "explained" to "not explained" falls at layer 5, where the gap crosses zero; we note the coincidence and do not build on it.

Both halves replicate at the two larger scales, on the one seed per arm whose checkpoint is on disk there (R^2_{emb} has a seed spread of about 0.01 at 13.9M, so one seed orders the arms and puts no interval on anything). Euclidean's first block leaves 0.561 and 0.654 of the embedding standing at 44.8M and 60.5M against 0.92–0.95 for the other arms; at the final site all three arms sit at 0.31–0.34; and the Euclidean−spectral residual at matched R^2_{emb} there is +0.0284 of +0.0453 and +0.0527 of +0.0583, 63% and 90% of the gap. Dilution explains less of the final site as scale grows, which is the same fact Appendix Y sees from the other side: the reversal, which is dilution, shrinks with scale, and the final-site deficit, which is not, grows.

Y THE DEPTH SWEEP AT 44.8M AND 60.5M

This appendix supports §4's reversal at the two larger scales and its crossing. Pre-registered: the 13.9M sweep established that the sign depends on readout depth, and this repeats it at the two larger scales, the measurement we would most want repeated.

A retraining, not a re-analysis. The matched-rate checkpoints behind Table 1's larger rows no longer existed in full: of the 18 cells, 10 survived on disk; the numbers survive in the per-seed record, the weights do not. So all 48 cells (two runs, three geometries, eight seeds) were retrained on one machine rather than only the missing eight, so that a machine-level offset cannot masquerade as seed variance inside a bootstrap; the hardware matched the original GPU model and PyTorch build. Learning rates are the stored matched-loss rates, read from file and not re-selected, since re-choosing them on eight-seed losses would let the data pick its own comparison. Seeds are eight rather than the three behind Table 1, because a percentile bootstrap at $n = 3$ has 27 distinct resamples and returns the observed range, which would make a per-layer decision weaker than the 13.9M result it exists to corroborate. The SAE readout is the published one throughout, and achieved L_0 is exactly 32.0 at every site except layer 0. Two gates fixed before the run both passed: retrained final validation losses reproduce the published values in all six arms, every deviation smaller than that arm's own

seed standard deviation (largest $|\Delta| = 0.0021$ against SDs of 0.0013–0.0044); and the retrained final-site gaps return $+0.0453$ and $+0.0583$ against the published $+0.0461$ and $+0.0573$, inside the pre-registered ± 0.0075. Table 1 is not updated to these eight-seed values because the retraining used the frozen matched-loss rates only, so eight seeds exist for that readout and not for matched compute; replacing one column would leave a row reporting its two readouts at different n, a worse defect than reporting both at $n = 3$ with the replication stated here.

Table 39: Explained variance by readout depth at both larger scales, eight seeds, with the Euclidean−spectral difference and a seed-level bootstrap against the 0.0100 floor. Both runs reverse sign. Layer 0 is the control and is sub-floor in both, in a sweep whose mid-stack effects reach -0.087; the final layer is the site Table 1 reports. The 44.8M run differs from 13.9M in scale alone; the 60.5M run differs from it in corpus, vocabulary, size and manipulated fraction.

Layer	Eucl.	sign	spectral	**eucl−spec**	**95% CI**	**verdict**
44.8*M, TinyStories (isolates scale against* 13.9*M)*						
0 (embed.)	0.9882	0.9879	0.9879	+0.0003	[+0.0002, +0.0005]	sub-floor
1	0.9128	0.9740	0.9694	**−0.0565**	[−0.0578, −0.0555]	**negative**
2	0.8721	0.9319	0.9473	**−0.0752**	[−0.0772, −0.0734]	**negative**
3	0.8199	0.8915	0.9069	**−0.0870**	[−0.0896, −0.0844]	**negative**
4	0.8035	0.8635	0.8654	**−0.0620**	[−0.0663, −0.0570]	**negative**
5	0.8012	0.8436	0.8247	**−0.0235**	[−0.0273, −0.0194]	**negative**
6	0.8037	0.8184	0.7950	+0.0088	[+0.0072, +0.0103]	sub-floor
7	0.8100	0.8060	0.7788	**+0.0311**	[+0.0299, +0.0323]	**positive**
8 (readout)	0.8202	0.8152	0.7748	**+0.0453**	[+0.0441, +0.0465]	**positive**
60.5*M, WikiText-103 (differs from* 44.8*M in corpus, vocabulary, size* and *manipulated fraction)*						
0 (embed.)	0.9452	0.9407	0.9377	+0.0075	[+0.0072, +0.0077]	sub-floor
1	0.9008	0.9497	0.9366	**−0.0358**	[−0.0379, −0.0335]	**negative**
2	0.8718	0.9256	0.9127	**−0.0409**	[−0.0464, −0.0357]	**negative**
3	0.8266	0.8894	0.8630	**−0.0365**	[−0.0423, −0.0312]	**negative**
4	0.8184	0.8522	0.8234	−0.0051	[−0.0100, −0.0004]	sub-floor
5	0.8204	0.8319	0.7957	**+0.0247**	[+0.0206, +0.0286]	**positive**
6	0.8256	0.8162	0.7838	**+0.0418**	[+0.0396, +0.0438]	**positive**
7	0.8317	0.8148	0.7779	**+0.0538**	[+0.0520, +0.0554]	**positive**
8 (readout)	0.8364	0.8220	0.7781	**+0.0583**	[+0.0564, +0.0600]	**positive**

The largest effect on an admitted site in the study, and how to read its size. It is -0.0870, at layer 3 of the 44.8M run (Table 39; the 210M run's -0.1309 of Appendix Z sits on an excluded site). Against the range of final-site gaps across runs ($+0.0300$ to $+0.0583$) it is 1.5–2.9×, a ratio to a range and therefore ambiguous; against that run's own final-site gap of $+0.0453$ it is 1.9× and opposite in sign, the within-run comparison we quote. By the same rule the 13.9M sweep's -0.0853, at the same site, is 2.8× its own $+0.0300$. That the two runs land within 0.002 of each other at layer 3, having been trained separately on different hardware at different times, is a near-replication we did not design for and did not predict.

Dead latents do not carry the reversal, and the check is stronger than at 13.9M. Dead fraction is confined to layers 1–2, where it is strongly asymmetric (spectral 0.31 against Euclidean 0.003 at layer 1 of the 44.8M run), and across the 18 run–layer cells the dead-fraction difference correlates with the EV gap at $+0.54$. But the association runs the wrong way for an artifact: where spectral has more dead latents it reconstructs better. Excluding every layer at which either arm exceeds 2% dead, layers 0–2 of the 44.8M run and 1–2 of the 60.5M run, both runs remain sign-reversing, and the largest effect (-0.0870, dead fraction 0.017) survives the exclusion.

The crossing, and what extrapolating it would have cost. Taking the crossing as the linearly interpolated zero of the Euclidean−spectral gap as a fraction of depth, it sits at 0.85 at 13.9M, 0.72 at 44.8M and 0.52 at 60.5M (the midpoint of the bracketing pair of sites gives 0.92, 0.69 and 0.56; the ordering is the same under either rule, the values are not, and no rule was fixed in advance). Scale and corpus both change across those three points and are not separated at the last one, and no part of this pattern was predicted. The natural worry is that the crossing moves earlier with scale, so the final-site sign could reverse at production scale and invert the headline. Extrapolated,

the crossing moves away from the final layer, at -0.44 of depth per decade of parameters (-0.53 under the midpoint rule); for it to reach the final layer one extrapolates downward, to about 7–10M parameters, below our smallest model, while the final-site gap grows with scale. Both trends make the final-site deficit more secure at scale. What the same extrapolation threatens is the reversal: the fraction of the stack whose gap is floor-clearing negative falls from 0.67 to 0.62 to 0.38, the deepest reversal at the largest scale is -0.0409 against -0.0853 and -0.0870, and a log-linear fit puts the whole stack positive somewhere near 0.8–1.3B parameters (the two crossing definitions give 794M and 1,289M, a $1.6\times$ disagreement that is itself the measure of what an R^2 over three points is worth). We could not exclude that, and declined to assert it: the shrink is carried entirely by the 60.5M point, where corpus, vocabulary cap and manipulated fraction all change, and the reversal depth is not monotone, being slightly larger at 44.8M than at 13.9M. The fourth point was run to answer the question in the direction that would hurt (Appendix Z), and the fit's prediction of 0.35 at 210M was contradicted out of sample by a measured 0.64; the reasoning above is left as it stood before that run.

One difference between the two larger runs we cannot account for. The sign-like−spectral gap is floor-clearing positive at every site 1–8 of the 60.5M run but reverses at 44.8M, being negative at layers 2–3 and positive from 5. The Euclidean−spectral comparison, which carries every claim in this paper, reverses in both. We report the discrepancy rather than resolve it.

Z A FOURTH SCALE POINT: 210M

This appendix supports Table 1's largest run and §4's fourth reversal. Three scale points, one on a different corpus, with tokens per parameter falling exactly as the gap rises, is one observation confounded with undertraining, and Appendix Y ends by saying that the depth claim's scaling is the open question a fourth point would answer in the direction that would hurt. This is that point, pre-registered, amended before launch to a larger model once the compute made the wall clock tractable, with predictions fixed before any run.

Design. Sixteen layers, 16 heads, width 1,024, vocabulary 8,192 on the frozen 13.9M map: 210.0M parameters, 201.3M of them in the 96 manipulated matrices (95.9%, the highest fraction in the study), 37,000 steps $\times 32 \times 256 = 303$M tokens $= 1.44$ tokens per parameter, one pass of a 330M-token TinyStories pool. The 44.8M run is eight layers of width 640 on the same corpus and vocabulary at 1.46 tokens per parameter, so $44.8\text{M} \to 210\text{M}$ is the only pair in the study that isolates scale at matched tokens per parameter: $5.1\times$ the manipulated parameters, $8 \to 16$ blocks, and nothing else the undertraining confound can hide in. Not held fixed, and disclosed: 0.92 epochs of the larger pool against 1.09 of the 60M subset, same tokenization. Everything else is the published protocol, imported rather than copied: the 44.8M run's own seven-point per-arm grids, the sweep at half length and one seed, three rates per arm in the final grid, the inline SAE readout of Table 1, the depth readout at all 17 sites, the selection rule of the scale study, a seed-pair bootstrap against the 0.0100 floor, the 2% dead-latent exclusion. Three seeds, so every $\pm$ is an observed range. Selections were frozen to files that later stages read and never recompute, and the step budget is asserted against the stored checkpoints.

Gates. Every arm's sweep optimum is interior, and every optimum (Euclidean 1.0, sign-like 5.6×10^{-4}, spectral 10^{-2}) is identical to the 44.8M run's at $4.7\times$ the size, an unplanned check on the design. Projected wall time under 24 hours from a calibration probe: 6.2 h projected, 5.6 h taken. The matched-loss residual to the common target is reported and never waived: Euclidean 0.00%, sign-like 1.52%, spectral 3.09%, the last outside the 1.5% parity aim as at the other two large runs and for the same reason, a half-decade grid; unlike those runs, this grid's bracketing cells straddle the target on their own, so Appendix F's reading at exactly the target is available without new cells and is reported beside the registered nearest-cell readout. Sites where any arm exceeds 2% dead latents are excluded per readout before classification. The pool holds one full pass (330M $\geq$ 303M).

The headline survives at both readouts. At the final residual stream, Euclidean−spectral is $+0.0544$ $[+0.0531, +0.0552]$ at matched loss and $+0.0722$ $[+0.0714, +0.0727]$ at matched compute, both floor-clearing, against the 44.8M run's $+0.0461$ and $+0.0692$ on the same corpus at the same tokens per parameter. The four-point series at matched loss is $+0.0300$, $+0.0461$, $+0.0573$,

$+0.0544$: it does not shrink, and the two largest points agree within the floor. Levels are lower than at the smaller scales (0.79, 0.79, 0.74 against 0.82–0.84), as a fixed $k = 32$ against a 1,024-wide residual should give, which is why Table 1 prints levels beside gaps. Sign-like−spectral is $+0.0499$ and $+0.0616$; Euclidean−sign-like is sub-floor at matched loss ($+0.0046$) and $+0.0106$ at matched compute, which the Holm correction removes (Appendix C). Read at exactly the common target (each arm's cell within 0.005 nats of it where one exists, else interpolated seed by seed between the two low-side rates that bracket it) the gap is $+0.0491$ $[+0.0482, +0.0499]$: spectral's 3.1% residual was worth about 0.005 of EV, in the direction the within-arm slope predicts, and the verdict is unchanged. The inversion holds a fourth time: spectral reaches the lowest loss (1.097 against 1.115 and 1.172) and the lowest EV (Table 14).

Table 40: Explained variance by readout depth at 210M, three seeds† (every interval an observed range), matched-loss rates, with the Euclidean−spectral difference, a seed-pair bootstrap against the 0.0100 floor, and the largest dead-latent fraction among the three arms at that site. Sites at which any arm exceeds 2% dead are excluded from the classification; their gaps are printed because they were measured and because the artifact runs the wrong way for a confound. Layer 0 is the control.

Layer	Eucl.	sign	spectral	**eucl−spec**	**95% CI**	**dead**	**verdict**
0 (embed.)	0.9907	0.9891	0.9885	$+0.0023$	$[+0.0022, +0.0025]$	0.084	control, excluded
1	0.9263	0.9768	0.9578	-0.0315	$[-0.0328, -0.0301]$	0.757	excluded (dead)
2	0.8777	0.9492	0.9393	-0.0616	$[-0.0706, -0.0525]$	0.815	excluded (dead)
3	0.8170	0.9166	0.9235	-0.1065	$[-0.1171, -0.0971]$	0.610	excluded (dead)
4	0.7703	0.8870	0.9012	-0.1309	$[-0.1332, -0.1272]$	0.357	excluded (dead)
5	0.7583	0.8410	0.8753	-0.1170	$[-0.1202, -0.1140]$	0.131	excluded (dead)
6	0.7554	0.8182	0.8490	-0.0937	$[-0.0959, -0.0917]$	0.030	excluded (dead)
7	0.7544	0.8027	0.8219	$\mathbf{-0.0675}$	$[-0.0741, -0.0639]$	0.010	**negative**
8	0.7549	0.7942	0.7971	$\mathbf{-0.0423}$	$[-0.0468, -0.0376]$	0.006	**negative**
9	0.7555	0.7874	0.7750	$\mathbf{-0.0195}$	$[-0.0220, -0.0174]$	0.005	**negative**
10	0.7575	0.7781	0.7601	-0.0026	$[-0.0062, -0.0007]$	0.004	sub-floor
11	0.7599	0.7712	0.7506	$+0.0092$	$[+0.0071, +0.0111]$	0.003	sub-floor
12	0.7627	0.7658	0.7449	$\mathbf{+0.0177}$	$[+0.0155, +0.0205]$	0.002	**positive**
13	0.7656	0.7644	0.7407	$\mathbf{+0.0249}$	$[+0.0237, +0.0266]$	0.002	**positive**
14	0.7696	0.7635	0.7363	$\mathbf{+0.0332}$	$[+0.0328, +0.0336]$	0.002	**positive**
15	0.7774	0.7683	0.7345	$\mathbf{+0.0429}$	$[+0.0416, +0.0438]$	0.001	**positive**
16 (readout)	0.7919	0.7874	0.7375	$\mathbf{+0.0544}$	$[+0.0531, +0.0552]$	0.002	**positive**

Depth at matched loss: sign-reversing a fourth time. Every site from 1 to 9 is negative with an interval excluding zero and sites 12–16 are positive and floor-clearing (Table 40). The 2% rule removes sites 1–6, where the spectral arm at its matched-loss rate (and the sign-like arm at its) carries 13–81% dead latents, so the classification rests on sites 7–16: negative at 7–9, positive at 12–16. The exclusion is conservative in the direction that matters: dead latents depress the EV of the arm that has them, and here that arm is spectral, whose EV is nonetheless the higher one at every excluded site, so the artifact works against the reversal as it did at 13.9M. Counting every site whose interval excludes zero, the negative region spans 0.56 of the stack; counting measurable sites only, 0.19. The largest effect anywhere in the study is now -0.1309, at layer 4 here, $2.4\times$ this run's own final-site gap and opposite in sign, on an excluded site; the largest on a measurable site is -0.0675 at layer 7, $1.2\times$. Under the Holm correction over 75 comparisons, layers 8, 9 and 12 fall at $n = 3$ and the reversal rests on layer 7 against 13–16.

The crossing: both registered predictions miss, and the miss is the finding. Appendix Y's three-point fit predicted a crossing of 0.35 at 210M, and the two TinyStories points alone predicted 0.54; both were registered, with a band of one block (0.0625), and no side was taken. Measured: 0.64 (0.66 under the midpoint rule). Neither prediction is within its band. From 44.8M to 210M the crossing moved -0.08 of depth where the fit said -0.29: the migration decelerates, and the extrapolation is contradicted out of sample in the direction that helps the reversal. A four-point refit, post hoc by construction, gives -0.19 per decade at $R^2 = 0.45$ and puts the vanishing point four orders of magnitude away, which is a statement about the fit's uselessness rather than about 10^{11} parameters. The crossing is measured at four scales, moves toward the middle of the stack slowly and not log-linearly, and we do not know what sets it.

Matched compute: unmeasurable mid-stack, as at the other large runs. At their own best rates the sign-like and spectral arms carry 65–99% dead latents at every site but the last, so only the final site is measurable and the profile classifies as late-layer-only on measurable sites, the pattern of Appendix AA at 44.8M and 60.5M, here at every mid-stack site rather than six of nine. The floor-clearing reversal is therefore established at matched loss at all four scales and at matched compute at 13.9M only.

What this does and does not buy. It removes 60.5M as the ceiling and replaces it with 210M; it supplies the one scale-isolating pair at matched tokens per parameter; it adds a fourth inversion and a fourth reversal; and it converts the extrapolation objection from "cannot exclude" to "contradicted at the next point". It does not reach the regime the warning is about, three to four orders larger; it is $n = 3$; its matched-loss spectral cell sits 3.1% from target; and the 210M reversal survives family-wise correction on one negative site. Not measured here: JumpReLU, the functional metric, embedding dilution.

Cost. 5.6 h on 8× RTX PRO 6000 (44.6 card-hours), the per-step rates 58, 67 and 183 ms for Euclidean, sign-like and spectral, the Newton–Schulz iteration scaling worse than the dense compute. Checkpoints of the fifteen cells either readout names are kept in full precision; the twelve bracketing cells, which no readout uses, in bfloat16 transport copies with the originals' checksums recorded.

AA THE DEPTH SWEEP UNDER BOTH READOUTS

This appendix supports §4's statement that the two readouts agree on the signs and disagree on what can be measured. §2 commits us to reading every important comparison at matched loss and at matched compute, because across 12 transformer comparisons the two disagree on 4 signs and 3 verdicts; the depth sweep had been read at matched loss alone. This closes that gap: all three scales, both readouts, eight seeds, 600 SAE fits, with the 2% dead-latent exclusion applied to both sides. The two readouts genuinely select different models. Euclidean's best-loss rate is its matched-loss rate at every scale (so that arm's two readouts share checkpoints, a fact rather than a convenience) while sign-like moves ($0.000251 \to 0.001585$ at 13.9M; $0.000215 \to 0.000562$ at both larger scales) and spectral moves furthest ($0.001259 \to 0.005012$; $0.001 \to 0.01$, a factor of ten). Each retrained final-site gap reproduces Table 1's matched-compute column ($+0.0416$ against $+0.0414$, $+0.0686$ against $+0.0692$, $+0.0663$ against $+0.0663$) under a gate fixed beforehand at 0.0075.

Table 41: The Euclidean−spectral gap by site at three scales under both readouts, eight seeds; * marks floor-clearing. "dead" is the largest mean dead-latent fraction across arms at that layer at 60.5M, shown because it is what separates the two halves of the table.

	matched loss				matched compute			
Layer	13.9M	44.8M	60.5M	dead	13.9M	44.8M	60.5M	dead
0	+.0019	+.0003	+.0075	.03	+.0015	+.0006	+.0054	.03
1	−.0749*	−.0565*	−.0358*	.40	−.0776*	−.0444*	+.0022	.88
2	−.0819*	−.0752*	−.0409*	.23	−.0757*	−.0672*	−.0467*	.83
3	−.0853*	−.0870*	−.0365*	.02	−.0779*	−.0886*	−.0658*	.60
4	−.0547*	−.0620*	−.0051	.00	−.0468*	−.0707*	−.0333*	.25
5	−.0026	−.0235*	+.0247*	.00	+.0038	−.0377*	+.0025	.08
6	+.0300*	+.0088	+.0418*	.00	+.0416*	+.0002	+.0363*	.02
7	—	+.0311*	+.0538*	.00	—	+.0379*	+.0593*	.00
8	—	+.0453*	+.0583*	.00	—	+.0686*	+.0663*	.00

The gaps agree; the measurability does not. At every site the matched-compute gap carries the same sign as the matched-loss gap and is usually larger (Table 41): at 60.5M layer 3 deepens from -0.0365 to -0.0658 and layer 4 from a sub-floor -0.0051 to a floor-clearing -0.0333. Read as raw effects, the mid-stack reversal is corroborated by the second readout rather than contradicted by it. What changes is whether those sites are measurable. Dead-latent fractions rise sharply under matched compute and the rise is confined to one arm (Table 42): Euclidean's rate is identical under both readouts and it stays near zero dead throughout, while spectral at matched compute sits at

$\eta = 0.01$, ten times its matched-loss rate, and its mid-stack activations stop being fittable by a fixed-k SAE at the published budget. Applying our own 2% exclusion therefore removes six of nine sites at both larger scales, against two or three at matched loss, and it removes exactly the sites carrying the negative half of the profile.

Table 42: Dead-latent fraction by arm at the 60.5M mid-stack sites under matched compute.

Layer (60.5**M, matched compute**)	euclidean	sign-like	spectral
1	0.007	0.559	**0.839**
2	0.003	0.404	**0.833**
3	0.000	0.149	**0.599**
4	0.000	0.023	**0.255**

What we claim, and what we do not. Under the pre-registered rule the shape verdicts are: 13.9M agree (sign-reversing under both; the "weaker" is only that more early sites are excluded), 44.8M and 60.5M disagree (sign-reversing at matched loss, depth-invariant at matched compute once the exclusion is applied). So the floor-clearing sign reversal is established at matched loss at all three scales and at matched compute at 13.9M only; at the two larger scales the mid-stack under matched compute is not measurable at the published SAE budget, which is a different statement from the effect being absent, and the raw gaps say which one this is. We report both and lead with neither. This is the third sighting of one mechanism: Appendix J found the same thing from the opposite direction, where the matched-loss rule pushed two isolated arms above their optima and their SAEs came back 84–97% dead mid-stack. The common factor is not the readout's name but distance from an arm's own optimum, whichever convention drives an arm well past it produces activations a fixed-k SAE at a fixed budget cannot fit. That is a property of the instrument, one more uncontrolled dependence of the kind this paper is about, and whether a larger SAE budget or dead-latent resampling restores measurability at those sites is a methods question we have not tested and will not decide after seeing which answer would suit us.

A hardware-offset control, obtained at no extra cost. The 13.9M cells were retrained for this experiment rather than transferred, which puts the new matched-compute profile on different hardware from the stored matched-loss one. Because Euclidean's best rate is its matched rate, the retrained Euclidean arm and the stored one are the same models at the same rate and seeds, so their difference is hardware and nothing else: across all seven sites the worst discrepancy is 8.9×10^{-5}, inside the stored arm's own seed SD at every layer. The comparison carries no hardware artifact.

AB The reversal at a second token budget and under a second SAE architecture

This appendix supports §4's training-and-architecture paragraph. The depth sweeps above run at one token budget per scale, so budget and scale move together, and through TopK alone; Appendix Q makes the final site architecture-independent but reads no site above it. Two sweeps on the 13.9M model, at the published depth readout, vary each in turn.

Design, fixed before any fit. *Budget:* the 20.04 tokens-per-parameter run of §3 ($34{,}000$ steps, $4.3\times$ the published training, same model, frozen vocabulary), TopK at every site, both readouts, its three seeds, rates read by the stored selection rule. *Architecture:* Matryoshka (the objective changed, the selection kept (Appendix Q)) on the published matched-loss checkpoints, eight seeds, matched loss; matched compute was not run. Gates: each sweep's final site must reproduce its published gap within 0.0075 (it does, three times: $+0.0422$ against the stored $+0.0422$, $+0.0412$ against $+0.0414$, $+0.0316$ against $+0.0317$); the 2% dead-latent rule per site, which at 4.71 tokens per parameter and matched loss excludes sites 1–2 and rests the reversal on 3–4. Four predictions: the budget run is sign-reversing at matched loss; its largest-magnitude site is layer 2 or 3 with at least half the -0.0853 seen at 4.71; Matryoshka agrees with TopK at every admitted site, within 25% where TopK clears the floor; the budget run is sign-reversing at matched compute on the sites the rule admits.

Table 43: The Euclidean−spectral gap by residual site at 13.9M. Bold: floor-clearing with the interval off zero; brackets: sites the dead-latent rule excludes (their raw gaps are shown, not read); †: three seeds, intervals are ranges. The crossing is the interpolated zero as a fraction of depth. The 20.04 tok/param run at matched compute is omitted: the rule excludes every site from 1 to 5 (dead latents 12–85%), leaving +0.0066 at site 0 and +0.0412 at site 6; the excluded raw gaps run −0.0423 to −0.0594.

	TopK, 4.71 **tok/param**		**TopK,** 20.04 **tok/param**	**Matryoshka,** 4.71
site	matched loss	matched compute	matched loss†	matched loss
0	+0.0019	+0.0015	+0.0027	[+0.0002]
1	[−0.0749]	[−0.0776]	[−0.0430]	[−0.0824]
2	[−0.0819]	[−0.0757]	[−0.0422]	[−0.0856]
3	**−0.0853**	[−0.0779]	**−0.0450**	[−0.0837]
4	**−0.0547**	**−0.0468**	**−0.0502**	**−0.0541**
5	−0.0026	+0.0038	**−0.0109**	−0.0033
6	**+0.0300**	**+0.0416**	**+0.0422**	**+0.0316**
crossing	0.85	—	0.87	0.85

Budget. At 20.04 tokens per parameter the profile is sign-reversing at matched loss (the first prediction held): negative and floor-clearing at sites 3–5, positive at 6, the same two early sites excluded (Table 43). It is flatter than at 4.71: layer 3 halves (−0.0853 → −0.0450), layer 4 holds (−0.0547 → −0.0502), layer 5 crosses the floor (−0.0026 → −0.0109) and the final site grows (+0.0300 → +0.0422). The largest-magnitude site is therefore layer 4, not 2–3: the second prediction missed. The negative region is wider and shallower, five of six blocks against four, and the crossing moves from 0.85 to 0.87 of depth, +0.02 along the training axis against −0.13 to −0.33 along the scale axis. At 13.9M, training length is not what moves the crossing; scale is (and at 60.5M the question cannot be asked, Appendix L). At matched compute the outcome is the one Appendix AA reports for the three larger scales, now at 13.9M with more training: spectral at its own best rate leaves 12–85% of the dictionary dead at every mid-stack site, the rule admits sites 0 and 6 only, and the fourth prediction missed. The mid-stack is unmeasurable at this SAE budget, not un-reversed, the excluded raw gaps all carrying the reversal's sign. So the matched-compute reversal is established at 13.9M at its published budget and nowhere else we have looked.

Architecture. Matryoshka's nested objective leaves more latents dead at early sites (12%, 37%, 21%, and 2.8% at layer 3, just over the rule where TopK had 0.1%), so four sites are excluded against TopK's two. On the three it admits, Matryoshka is TopK: layer 4 −0.0541 against −0.0547 (0.99×), layer 5 sub-floor in both, layer 6 +0.0316 against +0.0300 (1.05×); on the excluded sites its raw gaps (−0.082 to −0.086) are TopK's (−0.075 to −0.085). Sign-reversing, the third prediction held, resting under the rule on one negative site, which we say plainly.

Cost. 273 fits on one RTX 3090, six processes sharing the card, 1.5 h of occupancy.

AC THE MECHANISM CONTROLS AT LAYER 3, WHERE THE SIGN IS OPPOSITE

This appendix supports §4's statement that the two sites are dissociated on every mechanism axis. Every mechanism control of Part II sits at the final residual stream, and the paper's point is that the final layer is not special. Layer 3 is the reversal's largest admitted site (−0.0853; sites 1–2 are excluded by the dead-latent rule). The three controls were run there on the published matched-loss checkpoints, eight seeds, through the same code paths: the token-frequency stratification and its capacity check (Appendix U; TopK at $k = 16, 32, 64$ on the same activations, deciles and position mask), the covariance surrogate (Appendix T, the same protocol exactly), and PCA at the EV-matched rank with the dense autoencoder. Because the layer-3 gap is negative, the frequency statistic is read on the magnitude: |gap| on the rarest decile minus the commonest, as a fraction of the mean |gap|. Four predictions were fixed before any fit: the layer-3 gap is not tail-concentrated (fraction below 0.33); the Gaussian surrogate does not carry it ($r_{\text{synth}} \leq 0.25$); PCA sees at most a third of it; capacity closes it at most 2× faster on rare tokens. Gates: the $k = 32$ gap reproduces the

depth sweep's -0.0853 (it gives -0.0852); dead latents at most 0.16% at matched loss; at matched compute the sign-like arm leaves 8.9% dead at this site, so that block is shown and not read.

Table 44: The three mechanism controls at the two sites, 13.9M, matched loss, eight seeds. Every control turns over between the sites: under the appendices' own rules the final layer reads supported / beyond second order / sparse-specific and layer 3 reads refuted-flat / second-order / partial. The final-layer numbers are those of Appendices U and T; the layer-3 numbers come from the same code paths on the same checkpoints.

control	**final layer** ($+0.0300$)	**layer 3** (-0.0853)
frequency: |gap| rarest − commonest, over the mean	$+0.65$	-0.08
Spearman ρ, |gap| against frequency rank	-0.90	-0.07
capacity: |gap| closed, k=16 → 64, rarest / commonest	$6.4\times$	$1.5\times$
per unit of headroom	$3.9\times$	$1.07\times$
covariance surrogate: r_{synth}	0.18	1.11
r_{white}	1.61	0.62
PCA at the EV-matched rank: share of the SAE gap	0.20	0.73
linear gap at rank 32, PCA / dense AE	$+0.014$ / $+0.014$	-0.180 / -0.178

Two of four predictions held, and the two misses are the finding. The frequency profile at layer 3 is flat: -0.057 to -0.080 across the ten deciles at $k = 32$, |gap| rarest minus commonest -0.0056 $[-0.0094, -0.0015]$, refuted-flat under the rule of Appendix U, and capacity closes it almost uniformly ($1.5\times$, $1.07\times$ per headroom); the first and fourth predictions held, and the tail signature belongs to the final layer alone. The second and third missed, in the same direction: a Gaussian carrying each arm's exact layer-3 covariance reproduces the whole reversal (-0.0943 against -0.0847 real; whitening leaves -0.0528, so the two conditions disagree by the corroboration test of Appendix T and the reversal is over-determined there), and PCA at the matched rank sees 73% of it, with a linear gap of -0.180 at rank 32 because Euclidean's layer-3 stream is less compressible at every rank (0.37 against 0.55) (Table 44). The predictions were derived from the dilution account of Appendix X on the reasoning that a lookup-table remnant is higher-order structure a Gaussian destroys; it is not, the amount of embedding variance an arm's blocks leave standing is a covariance property, and the surrogate carries it.

What it means. The two sites are dissociated on every axis the paper has: the mid-stack reversal is second-order, largely linear and frequency-flat, the signature of embedding dilution, now measured three ways rather than inferred from one, and the final-site deficit is beyond second order, sparse-specific and concentrated in the tail. "The reversal is embedding dilution; the final-site deficit is not" therefore holds at the site where the sign flips, and "sparse-specific" is a statement about the final site, where it was measured. One scale, one site on each side of the crossing, one SAE seed; 192 fits and 24 PCA cells on one RTX 3090, six processes sharing the card, 0.8 h of occupancy.

AD DICTIONARY COMPARISON: MATCHING, CHANCE, AND PRECISION

This appendix supports §5's dictionary numbers: how two dictionaries are compared, what chance is, whether the matching rule is conservative, and how many decimals the levels carry.

The statistic. Every decoder column of one dictionary is matched to its best cosine partner in the other, and the median of that statistic is reported. The matching is greedy and not bijective, several columns may select the same partner, and this is not a small effect: across our 51 pairs a greedy match uses only $1{,}974$ of $3{,}072$ available partners on average, so a third of the columns are double-claimed. The chance level of 0.1780 is the same statistic computed between two matrices of $3{,}072$ random unit columns in $\mathbb{R}^{384}$, our exact dictionary shape, over six independent draws; maximum-cosine chance levels depend on both width and dictionary size, so this baseline does not transfer to a different configuration. The reading this appendix and Appendix AE reach, that a seed changes the coordinates a dictionary is written in more than what its features do, has a precedent outside SAEs, where independently trained networks were found to share low-dimensional subspaces expressed in different bases (Li et al., 2015), and Gerasimov et al. (2026) reach it from the other side: the

features that fail to reproduce across SAE seeds concentrate in reproducible lower-rank subspaces, basis ambiguity within a shared region of activation space rather than noise.

Whether greedy matching is conservative, measured. Greedy matching inflates both sides of the ratio the dissociation rests on, and a bijective assignment should tighten it; both halves deserved a number, so the same 51 pairs were re-matched with a Hungarian assignment (Kuhn, 1955) on the same cosine matrices, with no SAE refitted (Table 45). Every level is inflated, and the mechanism is as expected: the drop scales with how much structure a pair has, because a structured dictionary contains popular columns that many partners claim at once and a bijection has to give them up. The consequence runs the other way. Both orderings the dissociation needs survive, base-model seed stays above chance (0.1790 against 0.1725) and cross-geometry stays below the SAE-seed floor (0.3141 against 0.3806), and the ratio rises, from $27.5\times$ to $31.9\times$, because the near-chance denominator loses proportionally more than the structured numerator does. So the greedy rule was conservative for this comparison after all, and the greedy numbers are kept as published, having now been checked rather than asserted. One thing gets worse and one caveat is new: matched bijectively the base-seed alignment falls to 0.1790, which widens the unexplained disagreement with McCann (2026) from $4.9\times$ to $5.1\times$ and removes our ability to blame the matching rule for it, since this now is their rule; and the base-seed excess over chance falls from $+0.0090$ $[+0.0073, +0.0109]$ to $+0.0065$ $[+0.0046, +0.0088]$, whose lower end sits below the 0.5% convention that Appendix AI identifies as the one place our floor constant touches a stated claim. Under bijective matching that claim is marginal rather than clear.

Table 45: The four dictionary-alignment levels under greedy and bijective (Hungarian) matching, the same cosine matrices.

Level	n	**greedy**	**bijective**	**drop**
chance	6	0.1780	0.1725	-0.0055
base-model seed	9	0.1871	0.1790	-0.0080
cross-geometry	18	0.3317	0.3141	-0.0176
SAE-seed floor	18	0.4257	0.3806	-0.0451

How many decimals these carry. Printing dictionary quantities to four decimals on pair counts of 18, 18, 9 and 6 needs justifying, so an interval was put on each level (bootstrap over pairs, 10,000 replicates) and allowed to set the precision (Table 46). The chance baseline is reproducible to ± 0.0001 across draws and genuinely carries its fourth decimal; the other three levels carry three. The four-decimal forms are kept elsewhere because they are what the code emits and they make the arithmetic checkable, but no argument turns on a digit past the third, and the one that looked as though it might, the $27\times$ ratio, is restated with its interval in Appendix AI.

Table 46: Each alignment level with its pair-bootstrap interval and the number of decimals it supports.

Level	n	**mean**	95% **interval**	**decimals**
chance	6	0.1780	$[0.1779, 0.1782]$	4
base-model seed	9	0.187	$[0.1853, 0.1890]$	3
cross-geometry	18	0.332	$[0.3239, 0.3398]$	3
SAE-seed floor	18	0.426	$[0.4178, 0.4329]$	3

AE SEED AND GEOMETRY ACT ON DIFFERENT THINGS, IN FULL

This appendix supports §5's dissociation and records the one number in the paper that conflicts with published work. The two questions an SAE answers: *which* features are recovered, and *how well* they reconstruct, turn out to be governed by different parts of the training setup, and both are measured on the same axis in the same runs, so this is one experiment.

The two axes. Seed dominates the coordinates in which features are expressed; geometry dominates reconstruction quality; on which features fire, the two act equally. Across 24 independently

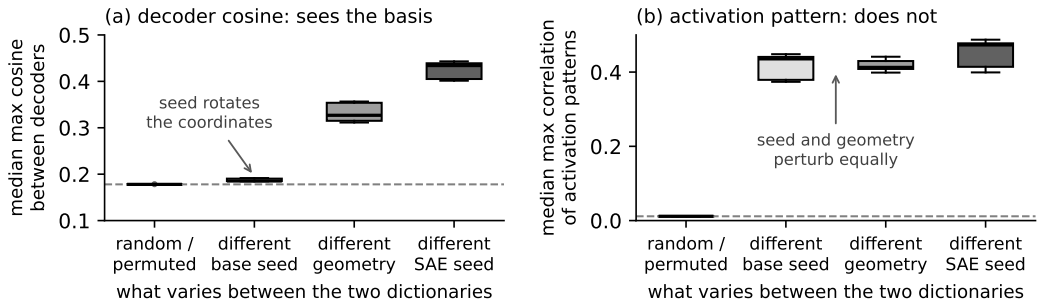

Figure 7: Agreement between two SAE dictionaries as a function of what differs between the models that produced them, on two criteria. **(a)** Decoder cosine, which depends on the residual basis: two SAEs on identical activations already disagree substantially (0.4256, 66% of features unmatched at cosine 0.5), an instability established at much larger scale (30% of features shared at 131K latents on Llama-3-8B (Paulo & Belrose, 2026)); changing the base model's seed leaves dictionaries $+0.0090$ above random vectors, below the reporting floor; changing the update geometry moves them less (0.3317). **(b)** The same pairs matched by the correlation of their features' activation patterns over the same frozen tokens, which does not depend on the basis: seed and geometry now perturb the feature set equally (0.4187 and 0.4180), both a floor-clearing 0.036 below the SAE-seed level (0.4539), and all three sit 0.41 above a permuted-token chance of 0.012. What (a) shows as seed destroying correspondence, (b) shows as seed rotating coordinates.

initialized base models per geometry (eight in each of the three headline runs, SAE seed pinned), seed-to-seed spread in EV is at most 0.0023 while the geometry gap is 18–$36\times$ that spread; the two larger runs were originally read at $n = 3$ and gave 18–$44\times$, and eight seeds narrowed the range rather than confirming it (Appendix AF). Neither knob is irrelevant to the other quantity; the ratios differ by more than an order of magnitude in opposite directions. The first consequence is uncomfortable for us: seed is the larger effect on dictionary geometry, so any claim that two models "find different features" must be read against a baseline in which two seeds already produce unrelated bases, and our geometry result is valid only because it is measured at matched seed. The second is general: a study reporting that an intervention changed an SAE's dictionary has, without a stability baseline, reported a quantity whose floor it never measured.

Why cross-geometry pairs agree more than cross-seed pairs. Figure 7(a) shows cross-geometry dictionaries aligning better (0.3133–0.3550) than cross-seed ones (0.1870), which reads as counter-intuitive until one notices what the seed controls. The seed is set immediately before the model is constructed, so at a given seed all three arms begin from byte-identical initial weights and consume the same data order; two seeds share neither. Cross-geometry pairs therefore differ only in the update rule applied to a common starting point, and the natural reading is that initialization largely fixes the basis and $8{,}000$ steps of a different geometry do not relocate it. We cannot fully isolate that: the seed sets initialization and data order together, and attributing the effect to initialization alone would need a run varying one at a time, which we did not do.

Geometrically unrelated, functionally corresponding. Decoder cosine is one criterion and not the only one. Because every arm and seed is evaluated on the same frozen validation tokens, each feature also has an activation pattern indexed by the same token positions, and two dictionaries can be matched by the correlation of those patterns, asking whether they respond to the same inputs rather than whether they point the same way. On the 13.9M Euclidean arm, base seeds 0 and 1 reach a median best activation-pattern correlation of 0.4478 against a measured chance level of 0.0118 (the same computation with one side's token order permuted, which destroys correspondence and preserves every marginal); a dictionary matched against itself scores 0.9997. So the same pair of models that sits $+0.0090$ above chance geometrically sits $+0.4360$ above chance functionally. This is why our result and reports of cross-seed feature universality (Bricken et al., 2023) can both be right: universality claims are made on interpretive, hence functional, criteria, and on a functional criterion we reproduce agreement. What does not survive is the stronger reading of our own earlier

framing, that seed changes which features exist. It changes the coordinate system they are expressed in, and that earlier claim is withdrawn.

Table 47: Every pair the decoder-cosine levels rest on, re-matched by activation pattern: 18 SAE-seed, 18 cross-geometry and 9 base-seed pairs, chance measured per pair by permuting one side's token order.

		decoder cosine		activation pattern	
what differs	n	greedy	above chance	greedy	above chance
SAE seed (same activations)	18	0.4257	+0.2477	0.4539	+0.4424
geometry (same base seed)	18	0.3317	+0.1537	0.4180	+0.4065
base seed (same geometry)	9	0.1871	+0.0090	0.4187	+0.4071

Every pair, not one. One arm and one seed pair do not license that sentence, so the activation-pattern matching was run on all 45 pairs the decoder-cosine levels rest on, with both greedy and bijective matching (Table 47); the Euclidean $0/1$ pair reproduces the stored 0.4478 (0.4482), and a dictionary against itself scores 1.0000. The seed–geometry asymmetry was coordinates. Geometrically, changing the seed costs 0.145 more alignment than changing the geometry does; functionally the two cost the same to four decimals (0.4187 against 0.4180). Both sit 0.036 below the SAE-seed level, which clears the floor, so the inequality the dissociation rests on, geometry moves the feature set by more than SAE-training noise does, holds on the invariant criterion too, at a third of its geometric margin. The ratio Appendix AI quotes, SAE-seed excess over base-seed excess, is $27\times$ in decoder cosine and $1.09\times$ in activation pattern; it is a statement about the basis and is labeled as one. The bijective match lowers every level by 0.09–0.10 and changes no ordering. Per arm, base-seed correspondence is 0.441 (Euclidean), 0.439 (sign-like) and 0.376 (spectral): the spectral models are functionally the least consistent across seeds, on three pairs per arm and one seed pair, so this is reported and not built on.

Where this sits among ways of comparing representations. Comparing what two trained networks have learned is an older problem than SAEs, and the field's measures encode different invariances. Matched units and shared subspaces (Li et al., 2015) are the nearest ancestor of our decoder-cosine matching; CCA-based similarity (Raghu et al., 2017; Morcos et al., 2018) and centered kernel alignment (Kornblith et al., 2019) compare subspaces or kernels and are invariant to rotations that our decoder cosine is not; model stitching (Lenc & Vedaldi, 2019; Bansal et al., 2021; Csiszárik et al., 2021) asks whether one network's representation can be mapped into another's and used, which is closer to our activation-pattern matching and to the transfer that McCann (2026) measure. These measures disagree with one another on the basic question of whether two random initializations learn the same representation (Ding et al., 2021), and initialization alone produces substantial representational differences at matched task performance (Mehrer et al., 2020). Our two criteria are chosen with that in mind: decoder cosine is deliberately coordinate-sensitive, since a practitioner reading a dictionary reads coordinates, and activation-pattern correlation is deliberately invariant to the basis; reporting both, against an SAE-seed floor and a random-vector chance level, is what lets the seed–geometry asymmetry be located in coordinates rather than asserted in general. The SAE literature has its own instability findings, which §7 lists: Paulo & Belrose (2026) recover different features from SAEs differing only in SAE seed; Gerasimov et al. (2026) locate the unstable features in reproducible low-rank subspaces; Fel et al. (2025) treat the instability as a problem to fix and Brzozowski & Chung (2026) reply that the gains are artifacts of initialization and metric design; Leask et al. (2025) show latents are neither complete nor atomic. Our same-activation floor reproduces the instability at our scale.

Cross-model SAE work asks a different question. A growing line fits dictionaries across models rather than comparing per-model reconstruction: Lan et al. (2025) find SAE feature spaces of different language models similar under rotation-invariant transformations; Thasarathan et al. (2025) and Nasiri-Sarvi et al. (2026) learn one concept space across models (and, for the latter, modalities); Chen et al. (2025) transfer features between residual streams by affine maps; Ognev et al. (2026) train a single shared SAE in place of per-model ones. None of these qualifies for the audit of Appendix AG, which counts only comparisons of a reconstruction quantity across base models, and

none varies the base optimizer; they are cited as the context in which a cross-model reconstruction number is read, and our result says that such a number carries the optimizer with it.

The conclusion is not ours first, and the geometric halves do not agree. McCann (2026) argue the same thing far more directly: independently trained transformers compute the same function in residual-stream bases differing by a rotation, so a seed changes coordinates rather than content. We reached the statement from the opposite direction. Our functional measurement (matched activation-pattern correlation) and theirs (transferring one seed's SAE onto another seed's activations, which collapses to negative explained variance and recovers to within 0.025 EV after a Procrustes rotation (Schönemann, 1966)) are consistent under that account, since a rotation preserves the correspondence we measure while destroying the transfer they measure. On nominally the same geometric statistic: SAE decoder columns, base models differing in initialization and data order, and $8\times$ expansion, they report mean maximum cosine 0.91–0.93 on Pythia-70m where we report 0.1870 against a chance level of 0.1780. The matching rule does not explain it: re-matched with the same bijective assignment they use, ours reads 0.1790, so the gap widens to $5.1\times$ (Appendix AD).

Training length is not why. What differs is the base models. Ours train for $8{,}000$ steps at 4.71 tokens per parameter; Pythia-70m is trained to convergence on three orders of magnitude more data. The candidate we named first was therefore that a canonical basis is something a model acquires, and that ours are stopped before it does, which predicted that the 20.04 tokens-per-parameter run, $4.3\times$ the training, would sit measurably above 0.1870. It does not. The test was fixed before any fit: the published recipe on three base seeds and three SAE seeds at both budgets, every pair under both criteria, the change paired over the 27 identical (arm, base pair, SAE seed) triples, a floor of 0.0100 on it, and for scale what the account implies: a basis acquired log-linearly in tokens per parameter from 0.187 at 4.71 to 0.92 at Pythia-70m's roughly $4{,}300$ would read 0.34 at 20.04. Measured, the base-seed level moves from 0.1867 to 0.1893: $+0.0025$ $[+0.0019, +0.0031]$, a quarter of the floor and sixty times short of the account; bijectively $+0.0031$; by activation pattern $+0.0003$ $[-0.0044, +0.0053]$. Flat under every criterion (Table 48). The other two levels are worth reading. The SAE-seed floor did not move ($0.4259 \to 0.4212$), so SAE training stochasticity does not depend on how long the base model trained. The cross-geometry level fell, $0.3317 \to 0.2984$ in decoder cosine (-0.0333 $[-0.0366, -0.0302]$) while by activation pattern it stayed at $0.419 \to 0.417$: longer training moves the three geometries' dictionaries further apart in coordinates and not at all in what fires, so the inequality behind the dissociation holds at the second budget with a wider margin (0.123 below the SAE-seed floor against 0.094), and the dissociation itself is reproduced there (same seeds, different geometry 0.417; same geometry, different seed 0.418). What remains: training length over the $4.3\times$ we can test is not the explanation, and the conflict stands. The candidates left are what a seed varies (ours changes initialization and data order together), scale, and the $900\times$ of training we cannot afford. A reader should treat our seed-level geometric figure as specific to small, briefly trained models.

Table 48: Dictionary alignment (median maximum cosine, or maximum activation-pattern correlation) at the two training budgets, 13.9M, final residual stream, 27 pairs per level and budget, the same three base seeds and three SAE seeds at both. The base-seed level does not move with $4.3\times$ the training; the cross-geometry level falls geometrically and not functionally.

	decoder cosine		**bijective**		**activation pattern**	
tokens per parameter	4.71	20.04	4.71	20.04	4.71	20.04
same arm, different base seed	0.1867	0.1893	0.1788	0.1819	0.4178	0.4181
same model, different SAE seed	0.4259	0.4212	0.3808	0.3703	0.4533	0.4501
same seeds, different geometry	0.3317	**0.2984**	0.3139	**0.2817**	0.4190	0.4173
chance	0.1780		0.1725		0.0116	

AF VARIANCE ATTRIBUTABLE TO THE SAE'S OWN SEED

This appendix supports the ratios of §5. Table 1 and every seed ratio in Appendix AE hold the SAE seed fixed, so every quoted spread is base-model-seed variance; since re-seeding the SAE produces substantially different dictionaries, that choice needs justifying rather than assuming: if the SAE

seed moved EV comparably, the ratios would be inflated. It does not. Refitting the SAE under five seeds on a fixed base model per arm (identical activations, so the only varying quantity is the SAE's initialization) gives EV standard deviations of 0.00032 (Euclidean), 0.00043 (sign-like) and 0.00053 (spectral). The largest is between three and five times smaller than the base-model-seed SDs of 0.00167, 0.00134 and 0.00163; folding both noise sources together in quadrature moves the ratios from 18.0, 33.7, 35.8 to 17.4, 32.1, 34.4; the same order of magnitude, in the conservative direction, so pinning the SAE seed did not manufacture them.

Eight seeds at every scale, and the range narrowed rather than confirmed. The two larger runs originally rested on $n = 3$, which indicates an order of magnitude rather than three significant figures. The depth retraining of Appendix Y supplies eight seeds at both, so all three ratios now rest on $n = 8$ and the final-site gaps replicate the stored ones to 0.0008 and 0.0010 (Table 49). The ratios move, and not in our favor: the headline range narrows from 18–$44\times$ to 18–$36\times$, because at $n = 3$ the 44.8M run's Euclidean arm had an underestimated spread ($0.00073 \rightarrow 0.00121$). Three seeds were not enough to characterize a standard deviation, which is now demonstrated rather than conceded; the claim the ratios support, that geometry moves reconstruction by more than an order of magnitude beyond what seed does, is unaffected, and its upper end is not.

Table 49: The geometry-gap-to-seed-spread ratio at each scale, before and after the eight-seed retraining.

	n	**pooled seed SD**	**ratio**	
13.9M	$8 \rightarrow 8$	$0.00167 \rightarrow 0.00167$	$18.1 \rightarrow 18.0$	unchanged
44.8M	$3 \rightarrow 8$	$0.00106 \rightarrow 0.00134$	$43.9 \rightarrow \mathbf{33.7}$	**falls**
60.5M	$3 \rightarrow 8$	$0.00200 \rightarrow 0.00163$	$28.6 \rightarrow \mathbf{35.8}$	rises

The dictionary is unidentified where the metric is not. Two facts sit together: re-seeding the SAE leaves 66% of features without a partner above cosine 0.5, yet changes explained variance by less than 0.0005. Many materially different dictionaries achieve the same reconstruction, so the SAE objective is badly under-determined at this width and sparsity even though its performance is reproducible. This is the sharpest version of why dictionary-level comparisons need a stability floor and reconstruction-level ones do not, and it is a property of the instrument rather than of any geometry.

AG THE CHECKLIST, APPLIED TO PUBLISHED COMPARISONS

This appendix supports the audit sentence of §5 and cites the premise the paper rests on: that SAE reconstruction metrics are used to compare base models. It applies the checklist of Appendix C as far as it can be applied from published text alone.

What was counted, fixed before searching. A qualifying comparison compares an SAE reconstruction quantity: EV, FVU, and loss recovered or achieved L_0, across two or more base models, where that includes a model against an intervened version of itself, and draws a conclusion from the comparison. Papers comparing SAE architectures on one base model do not qualify, since that is not the practice being described. Six candidates were examined and three qualified; the three excluded are named below, because a survey that reports only its hits is not a survey.

What the audit shows, and what it does not. Zero of three report the base optimizer or its learning rate (Table 50), which is the control this paper argues is load-bearing. Three of three report the readout depth: it is reported, it is simply not treated as something the answer's sign could depend on. Three of three hold SAE hyperparameters fixed across the models being compared, which §5 shows is a distinct problem, identical settings having given a $4.08\times$ spread in achieved L_0 across our arms. Two of the three compare a model against an intervened version of itself (pruning, compression) so the base optimizer is shared between their arms and our specific confound does not bite them; we say so rather than let the table imply otherwise. Li et al. (2026) compares across families (Qwen and Llama), where the optimizers and their schedules genuinely differ and are not accounted for.

Table 50: Three published comparisons audited against the checklist. Each row is what the paper states; "no" means we did not find it in the text, not that the authors were unaware of it. The base optimizer is often recoverable from a model's own release notes; the point is that the comparison does not account for it.

Comparison	base optimizer	its rate	readout depth	SAE hyperparams per model?
Borobia et al. (2026) pruned vs. dense	no	no	yes	shared
Gupte et al. (2025) compressed vs. dense	no	no	yes	shared
Li et al. (2026) across model families	no	no	yes	shared
this paper	yes	yes	yes	swept per arm

Excluded, and why. Gao et al. (2025) is cited elsewhere for the TopK architecture; we could not confirm from its text that it draws a conclusion from a cross-base-model reconstruction comparison, so it is not counted. Llama Scope (He et al., 2024) compares its release against others in a table but reports no cross-base-model reconstruction metric. Korznikov et al. (2026) compare trained SAEs against random baselines on one base model, an SAE-level comparison rather than a base-model one.

Scope, stated plainly. Three papers is an illustration, not a systematic review, and the sample was assembled by search rather than by enumerating a venue. It establishes that the practice exists and that the control is absent from the instances we checked; it does not establish a rate, which would need a sampling frame we did not build.

AH STATISTICAL REPORTING: WHAT EACH INTERVAL IS, AND WHAT IT LEAVES OUT

This appendix supports §2's statistical protocol and §6's statistical scope: which n each result rests on, the seed spreads behind the intervals, the resampling unit, the components of variation the printed $\pm$ does not include, and the interval on the loss-adjusted gap.

Which n each result rests on. Not every seed-derived quantity is read at $n = 8$; the central table is not (Table 51). The split follows from what was retrained: eight seeds exist at the frozen matched-loss rates and nowhere else, so everything computed from the depth sweep is $n = 8$ while everything computed from the original final grids stays at $n = 3$, and the 210M run is $n = 3$ throughout. In the $n = 3$ cells a percentile bootstrap over three units cannot exceed the observed range; those intervals are ranges, not calibrated coverage, and should be read as such wherever they carry a $\pm$.

Table 51: The seed count behind each reported quantity, by scale.

quantity	13.9M	44.8M	60.5M	210M
Table 1, matched loss	8	3	3	3
Table 1, matched compute	8	3	3	3
Pairwise table (Appendix H)	8	3	3	3
Loss adjustment, both readouts	8	3	3	—
Depth sweep, every site	8	**8**	**8**	3
Seed SDs and geometry/seed ratios	8	**8**	**8**	—

Per-arm seed standard deviations at the matched-loss cells. Table 52 reports them because the interval half-widths cannot be reconciled with the seed spread if the 13.9M values are used as a proxy for the larger runs, and they are not a good proxy. At 13.9M the reported half-width and the normal-theory one agree. At the two smaller-n runs the reported interval is tighter, for the reason stated: a percentile bootstrap over three units cannot exceed the observed seed range. That is why the † rows are read as ranges rather than as calibrated 95% coverage, and why the absolute floor rather than the interval is what the verdicts turn on.

Table 52: Per-arm seed standard deviations of EV at the matched-loss cells, the gap, and the interval half-width from normal theory against the one reported.

Run	n	Eucl. SD	sign SD	spec. SD	**gap**	normal-theory	reported
13.9M	8	0.00151	0.00185	0.00164	$+0.0300$	±0.0015	±0.0013
44.8M†	3	0.00073	0.00133	0.00111	$+0.0461$	±0.0015	±0.0016
60.5M†	3	0.00245	0.00135	0.00221	$+0.0573$	±0.0037	±0.0035

The resampling unit is the seed pair, and it used to be the seed. The arms share initialization and data order at each seed, so a seed is one unit carrying three values; an earlier version of the bootstrap (Efron, 1979) drew the arms independently, discarding whatever covariance the shared seed induces. All 18 comparisons were re-derived under both estimators (Table 53 gives the four extremes). The point estimate cannot move, the mean of differences being the difference of means, so only half-widths change: across all 18, no verdict changes, half-widths move by 0.49 to $1.46\times$ (median 1.03), and ten get wider against eight narrower. We adopt the paired estimator because it is the one the design licenses, not because it helped. What the correlations say is that the pairing buys nothing systematic: across arms the seed-to-seed correlation runs from -0.870 to $+0.888$ with median -0.076, and on the only run with $n = 8$ the median is $+0.032$. Sharing an initialization and a data order does not give two arms a common reconstruction component; the large $|r|$ values all sit at $n = 3$, where a correlation estimated from three points is noise, which is also why those rows move most.

Table 53: The interval half-width under the unpaired and the paired seed-block bootstrap for the four comparisons that moved most, with the seed-to-seed correlation between arms. The 44.8M and 60.5M rows rest on three seeds†.

Run	**Readout**	**pair**	**unpaired**	**paired**	r
13.9M	matched loss	Eucl.$-$spectral	±0.0014	±0.0013	$+0.196$
44.8M	matched loss	Eucl.$-$spectral	±0.0011	±0.0016	-0.627
60.5M	matched loss	sign$-$spectral	±0.0024	±0.0012	$+0.882$
44.8M	matched compute	Eucl.$-$spectral	±0.0014	±0.0007	$+0.888$

The interval is seed-only, and four larger components sit outside it. Every uncertainty component we have measured at 13.9M is in Table 54. Random rows have a sampling distribution; systematic rows are design choices held fixed, where there is no distribution to average over and the number is how far the gap moves under a defensible alternative. The two kinds must not be added; a systematic row is a span the reader applies, not a variance to fold in. Four components exceed the interval we print. That does not make the headline gap unsafe ($+0.0300$ survives every one of them, which is the whole point of running those controls) but it does mean the $\pm$ is not a total uncertainty and should never be read as one; it also sets the resolution at which any agreement between two configurations can be asserted, since below about 0.006 two numbers agreeing says more about the instrument than about the quantity. One correction of our own is recorded here: an earlier statement attributed a 0.0020 shift to the evaluation slice; Appendix P decomposes it into $+0.0017$ of SAE learning rate and $+0.0003$ of slice, and measuring the slice directly (same run, same configuration, two held-out slices) gives 0.0001: the component named was not the component measured.

Evaluation-set uncertainty, measured rather than bounded. Every primary EV comes from one fixed slice of 8,192 activations (32 sequences of 256 positions), so every interval conditions on a single token sample. We rebuilt all 24 matched-loss cells at 13.9M from their stored checkpoints, fitted the SAE on exactly the published pool, and scored that fixed dictionary on 40 further slices of the same size, disjoint from each other and from anything the SAE saw (Table 55). We had pre-registered the expectation that this would be material, and it is not. The reasoning was not wrong about the arms: a single arm's EV really does move by about 0.0030 between slices, twice the seed SD. What it missed is that the arms are scored on the same slice, and a slice that is hard to reconstruct is hard for all three, across slices the arms correlate at $+0.975$ (Euclidean against spectral) and $+0.990$ (Euclidean against sign-like), so roughly 78% of the per-arm slice variance cancels in the difference and what survives, 0.0007, sits below the ±0.0014 the paper prints; every one of the 40

Table 54: Every uncertainty component measured at 13.9M, by kind and size.

Component	Kind	Size	Resampled by us?
base-model seed	random	0.0013	**yes** — this is the printed ±
evaluation slice	random	0.0007	no — measured below
SAE seed	random	0.0006	no; fixed at 0 throughout
SAE configuration (8 cells)	systematic	0.0063	—
activation pool size	systematic	0.0025	—
SAE learning rate	systematic	0.0017	—
activation source	systematic	0.0016	—
weight decay	systematic	0.0009	—
evaluation slice (grid)	systematic	0.0001	—
refit on other hardware	systematic	0.0001	—

slices reproduces the verdict. This is the mirror image of the pairing result above: base-model seeds are paired by design and induce no common component, so pairing them is correct and gains nothing, while evaluation slices are shared and induce a very strong one, which is why conditioning on one cost almost nothing. Neither was knowable without measuring. Two limits and one by-product: this is 13.9M only, since WikiText-103's validation split holds 247K tokens and cannot furnish 40 disjoint slices beyond the SAE's own pool; it addresses which sample is scored, not whether positions within a sample recur in the training pool, which Appendix B measures separately; and the refits happened on a local machine's GPU where the published cells ran on CUDA, and the per-arm eight-seed means agree to 0.0001.

Table 55: The gap over 40 unseen evaluation slices, 13.9M, eight seeds.

Quantity (8 seeds × 40 slices, Eucl. − spectral)	
gap over the 40 unseen slices	+0.0299
against the published readout	+0.0300
slice-to-slice SD of *each arm*	0.0029–0.0031
slice-to-slice SD of **the gap**	**0.0007**
seed-to-seed SD of the gap	0.0019
residual	0.0005
slices whose gap clears the 0.0100 floor	**40/40**

An interval on the loss-adjusted gap. Resampling seeds within each learning-rate cell and rerunning the identical interpolation 10,000 times gives the common-reference adjusted gap of Appendix E as $+0.0326$ with 95% CI $[+0.0314, +0.0332]$; the bootstrap mean sits 0.0002 below the point estimate, so the estimator is close to unbiased here, and the entire interval clears the floor. The reference loss and the per-arm fit method are held fixed, both being design choices fixed before the data, so this is the uncertainty given those choices. For the sign−spectral pair the same procedure is not well behaved: the bootstrap mean sits 0.0016 above the point estimate and the interval does not cover it, because that arm has the worst linearity in the study ($R^2 = 0.22$ on the extended grid, Appendix E), so which measured cells bracket the reference changes under resampling and the local-interpolation estimator is biased. We report no interval for that pair rather than print an uncalibrated one.

Effective rank against EV, partialling out loss. Since $\mathrm{dEV}/\mathrm{d}L > 0$ within arms, the negative effective-rank/EV correlation could partly reflect effective rank proxying for "trained further". Partial correlations at 13.9M ($n = 8$, controlling for validation loss) strengthen rather than weaken: $-0.949 \to -0.924$ (Euclidean), $-0.642 \to -0.937$ (sign-like), $-0.780 \to -0.766$ (spectral), and the sign-like arm moves from $p = 0.086$ to $p = 0.002$. The association is not mediated by loss. The control is weak in one respect: within a cell, loss varies only by seed noise, and $r(\text{effective rank}, \text{loss})$ is itself inconsistent in sign across arms ($+0.57$, -0.42, -0.29), which is the more direct reason the mediation account does not hold.

AI FLOORS: PER-METRIC DERIVATION, AND WHAT A DIFFERENT CONSTANT WOULD CHANGE

This appendix supports §2's absolute floor. The rule sets floors at 1% of a metric's bounded range, and two things need stating that the main text has no room for: the bound for each metric (Table 56), and an honest account of which verdicts turn on the constant. The 1% is our convention and nothing forces it, so it should be possible to check what it is doing. The need for some such floor is not ours alone: Chanin (2026) finds that the field's SAE benchmark metrics carry more reseed noise and less discriminability than is assumed, two of them enough to be unusable at their canonical settings.

Table 56: Where each floor comes from, and what three conventions imply.

Metric	Bounded range, and why	0.5%	1%	2%
explained variance	$[0, 1]$ *attainable*; see below	0.005	**0.010**	0.020
median max-cosine	$[0, 1]$ *attainable*; see below	0.005	**0.010**	0.020
effective rank	$[1, d_{\text{model}}] = [1, 384]$	1.92	**3.84**	7.68
interference Ψ	*unbounded above*	different rule, below		
validation loss	*unbounded above*	no floor defined		

Neither of the first two metrics is bounded the way one might assume. $\text{EV} = 1 - \text{MSE}/\text{Var}$ is bounded above by 1 but unbounded below, a reconstruction worse than predicting the mean gives a negative value, and ordinary cosine lies in $[-1, 1]$. Calling either range $[0, 1]$ and deriving the floor as "1% of a bounded range" was wrong as stated, and we correct it rather than restate it. What survives is the constant, not the derivation. $[0, 1]$ is the attainable range for both quantities in any regime where the instrument is working: a dictionary that reconstructs worse than the mean, or a decoder anti-aligned with every counterpart, is a broken fit rather than a measurement to compare. Empirically, across all 894 EV values this paper reports the observed range is $[0.694, 0.988]$, none below 0 or above 1, and the lowest value anywhere (the covariance-matched surrogate of Appendix T at about 0.59) is a deliberately degraded condition. So 0.0100 should be read as a declared smallest effect of interest (in the sense of Lakens, 2017), chosen by us as 1% of the attainable range and defended by the sensitivity analysis below rather than by a bound; the constant is ours, and nothing in that framework selects it. Two metrics have no upper bound, and we do not pretend otherwise: interference Ψ is normalized so an orthogonal dictionary gives 0 but has no maximum, so it alone uses $\max(1\%$ of the mean, 10% of the observed span$)$, which is how the 0.1539 of Appendix AJ arises; validation loss is likewise unbounded and we define no floor for it, where the paper compares arms by loss it reports the nat differences and relies on the EV side for floor-clearing.

Sensitivity. Of the 18 EV comparisons at the three headline scales (three pairs, two readouts, three runs, a different 18 from the first transformer run's six metrics times three pairs, of which 11 excluded zero and only 7 cleared their floor, the count that first showed the floor was doing work), 14 clear at 0.5%, 13 at 1% and 12 at 2%. All 12 comparisons involving the spectral arm clear at every convention, and so does the entire depth verdict of §4: the same five sites clear at 0.5%, 1% and 2%, so sign-reversing does not depend on the constant. The three verdicts that move are Euclidean−sign-like at 13.9M matched compute (0.0092) and at 60.5M matched loss (0.0144), both middle-pair comparisons we already decline to claim, and the base-seed dictionary alignment ($+0.0090$ above chance), which is sub-floor at 1% and 2% but would be reportable at 0.5%. That is the one place the constant touches a stated claim, and it is more marginal than that under bijective matching, where the excess falls to $+0.0065$ $[+0.0046, +0.0088]$ and the interval no longer clears 0.5% even though the point estimate does (Appendix AD). On that one the honest reading is that the verdict label moves and the conclusion does not: what the dissociation rests on is not the gap between base-seed alignment and chance but its size relative to the SAE-seed stability floor, which sits 0.248 above chance against the base seed's 0.009. We previously called that ratio $27\times$ and it does not support two significant figures, resampling both levels gives $27.5\times$ with a 95% interval of $[22.7\times, 33.9\times]$, because a ratio whose denominator is 0.009 ± 0.002 inherits that denominator's instability. The claim is therefore stated as more than $20\times$, which the interval supports, which survives bijective matching ($31.9\times$), and which is all the argument needs. It is a statement about decoder coordinates: matched by activation pattern instead, the same ratio is $1.09\times$, because seed

rotates the basis far more than it changes which features fire (Appendix AE). Two base seeds land nearly as far apart as two random dictionaries and vastly further apart than two SAEs on identical activations, and no choice of constant in this range disturbs that ordering.

AJ THE TOY-MODEL TRANSFER TEST, WHICH FAILED

This appendix supports §6's second failed test. The toy model is a two-layer model of superposition (Elhage et al., 2022). Over a pre-registered dose–response across six feature-correlation levels, the Euclidean-minus-spectral difference in interference varies monotonically with correlation (Spearman $+1.000$ over a common sparsity band) and changes sign between $\rho = 0.263$ and $\rho = 0.529$. We pre-registered the transfer test before measuring the transformer, fixing the mapping from measured correlation to predicted sign in advance. Feature correlation in the transformer, measured on SAE codes with the coupled-pair fraction matched to the toy model's, is $\rho = 0.150$ (resolved at $8.1\times$ a permutation null and below the crossing point in all three arms) so the toy model predicts that spectral superposes more. At matched loss the measured difference is -0.0905 against a floor of 0.1539: the predicted sign, but sub-floor, so by our own rule it carries no reportable sign. At matched compute it is $+1.3579$, which clears its floor and carries the opposite sign. The one readout that resolves contradicts the prediction. The test could not have been won either way: the interference contrast at matched loss is $0.59\times$ its floor, and no number of seeds moves an absolute floor. We record that as a design fault of the test rather than as a result about transformers, and the transformer findings stand on their own measurements.

AK COMPUTE

This appendix supports the reproducibility statement. The toy-model study ran entirely on CPU (about 1.7×10^5 model trainings, made affordable by treating seeds as a leading tensor axis so that 20 seeds cost roughly $1.6\times$ a single seed). The transformer study used $\approx$132 h on one RTX 4090, 5.6 h on $8\times$ RTX PRO 6000 (44.6 card-hours) and 4.0 h on one RTX 3090, plus 21 CPU-hours and 7 local machine GPU hours of SAE fitting. Hours on different cards are not the same unit and are not summed; the statement names each. Table 57 is by hardware. Its rows are rebuilt from the per-step times measured on the training GPU and the run and step counts read out of the stored sweep rows for the sweeps whose rates were calibrated, and from each sweep's recorded session wall time for the rest, with the exceptions noted beneath it. Every stage but the last two groups ran on one RTX 4090: the run records name that card, and no stored sweep shows a per-step rate faster than the 4090's for its model shape, which a larger card would.

Per-arm cost, and what "matched compute" does and does not mean. The three update rules do not cost the same per step: the Newton–Schulz iteration adds five matmuls per matrix parameter per step, and the exact-SVD arm more. Marginal per-step times, measured on the training GPU by differencing two probe lengths so that fixed setup cancels, are in Table 58. So our second readout is matched steps and tokens, not matched FLOPs or wall clock, and calling it "matched compute" overstates it; we keep the name for continuity and a reader should read it as matched steps.

The FLOPs, counted, and why they disagree with the clock. Counting from the checkpoint configurations (6 FLOPs per matmul parameter per token for forward and backward, attention products at full T^2, the tied head, and the optimizer step itself) the Newton–Schulz iteration costs $5 \times (2m^2n + m^3)$ MACs per matrix per step on the six matrices per block that receive the geometry, and nothing else does anything but elementwise work (Table 59). In arithmetic the spectral arm receives 14–27% more per step, not 65–95%; the rest of the wall-clock gap is launch overhead on 384- and 640-wide matmuls, which a 4090 does not fill. The Newton–Schulz share of a spectral step is 12–21% and is largest at 44.8M, where the block width is 640 against a vocabulary of only 8,192; the overhead scales with d^3 per matrix against d^2 per token, so it grows with width at fixed batch. Adam's elementwise cost is invisible at this precision.

What a FLOP-matched budget would cost, bounded from stored curves. Under equal FLOPs the spectral arm gets $8{,}000/1.14 = 7{,}025$ steps at 13.9M, 6,293 at 44.8M and 6,648 at 60.5M. Reading its validation loss at those steps off the stored trajectory of its matched-compute cell, and

Table 57: Compute by stage and hardware. The depth retraining has no sweep record and is taken from the calibration measured on its GPU. The dose–response ran two to four processes on one card, so session wall times would multiply-count; its figure is the card's occupancy from first to last checkpoint. The fourth scale point's wall-clock hours are the run's stage timestamps (5.57 h on all eight cards, plus a three-minute calibration on one); card-hours are that occupancy summed over cards. The SAE-architecture, second-budget, layer-3 and 60.5M-budget rows are each card's occupancy from its first to its last process, part of it spent waiting for checkpoints to arrive. The retracted 4,000-step exact-SVD run (Appendix K) is not counted; it left no wall-time record.

Stage	**runs**	**steps**	**hours**
one RTX 4090			
13.9M learning-rate sweep	18	4,000	0.45
13.9M final sweep	72	8,000	3.60
44.8M learning-rate sweep	21	4,000	1.86
44.8M final sweep	27	8,000	4.78
60.5M learning-rate sweep	21	4,000	2.18
60.5M final sweep	27	8,000	5.60
token-budget run, learning-rate sweep	18	16,992	1.91
token-budget run, final sweep	27	34,000	5.73
grid extension for the loss correction	72	8,000	3.60
replication of the original cells	24	8,000	1.20
SAE fits run on GPU	591	—	3.4
weight decay, four attempts (§2.1)	147	4,000–8,000	7.8
exact-SVD instantiation, re-run at full budget (Appendix K)	30	4,000–8,000	26.7
depth retraining at 44.8M and 60.5M, 8 seeds (Appendix Y)	48	8,000	7.4
isolated family (§2.1)	99	4,000–8,000	3.2
depth at both readouts (Appendix AA)	72	8,000	5.6
functional metric (Appendix V)	88	8,000	5.3
interpolative overlap (Appendix F)	26	8,000	4.3
tied-aux counterfactual (Appendix R)	66	8,000	7.0
batch size 4× (Appendix N)	66	2,000	2.1
dose–response, 116 runs sharing the card (Appendix S)	116	8,000	5.5
SAE architectures, six processes sharing the card (Appendix Q)	289	—	1.8
JumpReLU under three protocols (Appendix Q.2)	348	—	1.0
seed-level alignment at 20.04 tok/param (Appendix AE)	33	—	0.3
60.5M at 4.71 tok/param (Appendix L)	42	17,400–34,800	19.9
one RTX 4090			**132.2**
eight RTX PRO 6000 Blackwell, one process per card			
fourth scale point, 210M (Appendix Z)	48	18,500–37,000	5.6
8× **RTX PRO 6000, wall-clock hours**			**5.6**
8× **RTX PRO 6000, card-hours**			**44.6**
one RTX 3090			
depth at a second budget and SAE (Appendix AB)	273	—	1.5
mechanism controls at layer 3 (Appendix AC)	192	—	0.8
operating-point brackets at full duration (Appendix A)	14	8,000	1.7
one RTX 3090			**4.0**

Table 58: Marginal milliseconds per step by arm, measured on the training GPU.

run	Euclidean	sign-like	spectral	**spectral / Euclidean**
44.8M	38.0	40.3	62.7	1.65×
60.5M	39.3	47.4	76.5	1.95×

applying the within-arm $\mathrm{dEV}/\mathrm{d}L$ of Appendix E, gives Table 60. The direction is the one we predicted, against us, and the size is a tenth to half of the floor; every matched-compute spectral comparison clears the floor by 4× or more under FLOP matching. Two things make this a bound rather than a result: the loss is read mid-schedule, where the cosine decay has not finished, so a dedicated shorter run would anneal fully and reach a lower loss than the value used, making the shift

Table 59: Training FLOPs per step by arm, the Newton–Schulz share, and the ratio in arithmetic against the ratio on the clock.

run	train GFLOP	NS5 GFLOP	**spec. / Eucl., FLOPs**	sign / Eucl.	wall-clock
13.9M	734	102	$\mathbf{1.14}\times$	$1.00\times$	—
44.8M	2,319	629	$\mathbf{1.27}\times$	$1.00\times$	$1.65\times$
60.5M	3,092	629	$\mathbf{1.20}\times$	$1.00\times$	$1.95\times$

an over-estimate; and the slope is across learning-rate cells, not along a trajectory. A FLOP-matched run is the real answer and needs a GPU; this says one is not worth the budget. The matched-loss readout, which carries the headline, is untouched throughout, because it equalizes the endpoint rather than the budget, and Appendix F shows the gap flat across a 0.055-nat band.

Table 60: The matched-compute gap re-read at a FLOP-matched step count, bounded from the stored loss trajectories and the within-arm slope.

run	steps	ΔL (nats)	$\mathrm{dEV}/\mathrm{d}L$	EV shift	gap, matched steps	**gap, matched FLOPs**
13.9M	7,025	$+0.0054$	0.092	$+0.0005$	$+0.0414$	$+\mathbf{0.0409}$
44.8M	6,293	$+0.0231$	0.204	$+0.0047$	$+0.0692$	$+\mathbf{0.0645}$
60.5M	6,648	$+0.0194$	0.040	$+0.0008$	$+0.0663$	$+\mathbf{0.0655}$

What is measured and what is estimated. Base-model training, which is 97% of the total, is derived from measurement. The SAE-on-GPU row is the one estimate (about 0.35 min per fit); at 3.4 of 132 hours, even a factor-of-two error in it moves the total by 3%. A further 21 CPU-hours of SAE fitting: the spectrum surrogates, the depth sweep, the SAE-seed and rescaling controls, ran on a local machine and is reported separately because it needs no GPU to reproduce; 7 more hours (the token-frequency cells of Appendix U, the Gated calibration of Appendix Q and the residual-stream statistics of Appendices A and L) ran on that local machine's own GPU and are stated beside them. The study is cheap enough for independent replication; an earlier version of this accounting printed a single cross-hardware total that was simply wrong, and the by-hardware form is what replaced it.