Null-Calibrated Evaluation of Sparse Autoencoder Decoder Reproducibility

Anonymous Authors[1]

Abstract

Sparse autoencoders (SAEs) are often evaluated by reconstruction loss, but interpretability workflows also require that learned dictionaries be reproducible across random seeds and robust to evaluation artifacts. We study SAE decoder reproducibility as a benchmark-design problem: every stability score is reported against a metric-specific random-dictionary null, pairwise seed statistics are treated as dependent, and decoder geometry is audited with assignment-based, activation-level, firing-overlap, causal, streaming, and synthetic-ground-truth controls. In compute-limited cached-activation regimes, reconstruction can appear converged while decoder-column similarity remains within 1.5% of the geometric null; longer training raises decoder agreement, but activation and functional diagnostics lag. These results argue that SAE benchmarks should report reconstruction, null-calibrated decoder matching, held-out activation agreement, and ground-truth or downstream checks together rather than treating reconstruction or a single stability metric as sufficient.

1. Introduction

Sparse autoencoders (SAEs) are a central tool for mechanistic interpretability: they decompose neural network activations into sparse latent variables whose decoder columns are treated as candidate feature directions (Cunningham et al., 2023; Bricken et al., 2023; Gao et al., 2024; Templeton et al., 2024). The standard selection criterion is reconstruction quality. For benchmark design, however, reconstruction is not enough. If two runs with the same data and hyperparameters learn different dictionaries, then a single-run explanation may reflect an arbitrary decomposition rather than a reproducible property of the model.

This paper frames SAE stability as a reliable-evaluation problem. Recent work reports substantial cross-seed instability (Paulo & Belrose, 2025), argues for feature consistency as a primary criterion (Song et al., 2025), and shows that reconstruction quality can fail to predict feature quality in synthetic benchmarks (Chanin & Garriga-Alonso, 2026). We ask what a benchmark should measure when the object under evaluation is itself a learned, overcomplete dictionary with permutation, splitting, and redundancy ambiguities.

Our answer is a null-calibrated multi-metric protocol. We evaluate SAEs on an algorithmic transformer and selected Pythia layers (Biderman et al., 2023); report every stability metric against a random-dictionary or random-SAE baseline; distinguish nearest-neighbor decoder geometry from one-to-one assignment; use held-out activation, firing-overlap, causal, streaming, and synthetic-ground-truth checks; and document failure modes such as dead-neuron resampling. The contribution is not a new SAE architecture. It is a benchmark discipline for deciding when reconstruction, decoder reproducibility, and feature correctness should be kept separate.

Contributions.

1. We derive and use a calibrated random-dictionary null for decoder nearest-neighbor similarity, matching empirical random baselines to four decimals in the main settings.
2. We show a reproducible timescale separation: short-budget SAEs reconstruct well while decoder-column similarity remains practically at the null; longer training raises decoder agreement.
3. We audit the same checkpoints with assignment-based, activation-pattern, firing-overlap, causal, streaming, and synthetic-ground-truth controls, exposing where decoder geometry is and is not informative.
4. We package the reported numbers with frozen manifests and a validator, making the paper values auditable from the supplement.

2. Evaluation Protocol

Objects being evaluated. An SAE maps an activation $x \in \mathbb{R}^{d_{\text{model}}}$ to sparse latents $z \in \mathbb{R}^{d_{\text{sae}}}$ and reconstructs

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

$\hat{x} = W_{\text{dec}} z + b$. We evaluate the decoder columns of W_{dec} across independently trained SAEs. This is intentionally narrower than semantic feature identity: decoder agreement does not by itself imply that two latents fire on the same tokens or have the same causal effect.

Decoder MMCS. Mean maximum cosine similarity (MMCS) computes, for each decoder column in one SAE, the maximum absolute cosine similarity to any decoder column in another SAE, symmetrized across directions. MMCS is useful for detecting whether two overcomplete dictionaries occupy nearby decoder directions, but it is many-to-one and can over-credit splitting or redundancy.

Assignment, activation, and functional diagnostics. We therefore pair MMCS with stricter and more functional checks. Decoder Hungarian matching enforces one-to-one assignment over decoder cosine similarities. Activation-pattern Hungarian matching assigns features by held-out latent activation correlation. Firing-overlap Hungarian assigns features by held-out binary firing Jaccard similarity. Representational similarity analysis (RSA) compares latent-space sample-similarity matrices. A mod-113 causal-ablation check matches features by decoder Hungarian matching and compares the per-example true-logit drop after ablating each feature's decoded contribution.

Random baselines. For two independent random unit vectors in $\mathbb{R}^d$, the inner product obeys $(z+1)/2 \sim \text{Beta}((d-1)/2, (d-1)/2)$. The expected maximum over m dictionary elements gives a calibrated reference for MMCS. For $d = 128, m = 1024$, the predicted null is 0.2989 and the observed random baseline is 0.2990; for Pythia-410M with $d = 1024, m = 1024$, predicted and observed are 0.1073 and 0.1074. Activation, firing, RSA, and assignment baselines are computed by running the same metric on randomly initialized SAEs with the same architecture and data split.

Statistical dependence. Pairwise seed comparisons are not independent because each SAE appears in multiple pairs. We use pairwise summaries descriptively and reserve confirmatory statements for whole-seed resampling and seed-label permutation in the 20-seed short-budget study.

3. Experiments

Algorithmic transformer. We train a two-layer transformer ($d_{\text{model}} = 128$) on modular addition modulo 113 to 100% accuracy through grokking (Power et al., 2022; Nanda et al., 2023). SAEs are trained on layer-1 residual-stream activations with seeds $\{42, 123, 456, 789, 1011\}$ unless otherwise stated.

Table 1. Recommended SAE reproducibility reporting ladder. Each row controls a different benchmark failure mode; none alone establishes semantic feature correctness.

Metric	Null/control	What it supports
Recon. EV	Held-out split	Basic reconstruction quality
Decoder MMCS	Random dictionary	Many-to-one decoder geometry
Decoder Hungarian	Random SAE	Exclusive decoder assignment
Activation / firing	Held-out random SAE	Functional-use similarity
Causal / GT / task	Active-random or planted GT	Task-relevant correctness

Table 2. Short-budget mod-113 architecture comparison at 30 epochs. All architectures remain near the MMCS random null. ReLU-family L0 values are not sparsity-matched to TopK/BatchTopK, so this is not an architecture-quality ranking.

Architecture	MMCS	Rand. ratio	L0	Loss
TopK	0.304 ± 0.001	$1.015\times$	32	21.3
BatchTopK	0.302 ± 0.001	$1.008\times$	32	8.9
ReLU	0.300 ± 0.001	$1.002\times$	$\sim$328	0.31
JumpReLU	0.300 ± 0.001	$1.002\times$	$\sim$365	0.42

Language model scope checks. We train TopK and ReLU-family SAEs on selected Pythia-70M/160M/410M layers extracted from Wikitext-103 tokens (Biderman et al., 2023). These are selected-layer scope checks, not frontier-scale scaling laws.

Synthetic positive control. We generate $x = W_{\text{gt}} z + \epsilon$ with $W_{\text{gt}} \in \mathbb{R}^{128\times256}$, positive 5%-sparse coefficients, and Gaussian noise $\sigma = 0.01$. Because planted decoder columns are known, we can compare cross-seed MMCS with ground-truth MCC (GT-MCC) computed by Hungarian matching against W_{gt}. A regression test verifies that a column-permuted overcomplete dictionary scores GT-MCC $= 1.0$.

4. Results

4.1. Reconstruction Can Converge Before Decoder Reproducibility

At 30 epochs, four SAE architectures reduce reconstruction loss on mod-113 while decoder MMCS remains near the random-dictionary null (Table 2). For TopK, trained MMCS is 0.3040 versus 0.3000 for random dictionaries, an absolute gap of 0.0040. In a 20-seed robustness run, a whole-seed bootstrap gives a 95% interval of $[0.0037, 0.0043]$ for the trained-random gap; the gap is detectable but within the practical-equivalence margin $\epsilon = 0.005$.

Longer training changes the conclusion for decoder geome-

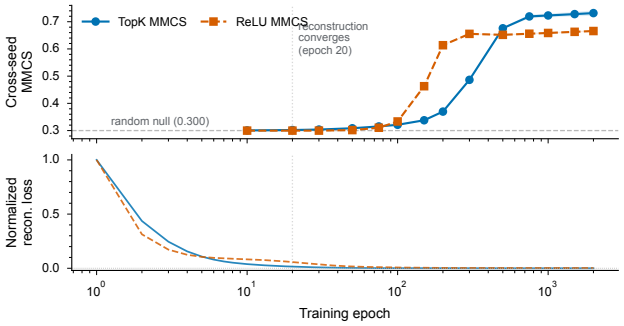

Figure 1. Reconstruction converges before decoder reproducibility. Decoder MMCS remains near the random null early, then rises long after normalized reconstruction loss has saturated.

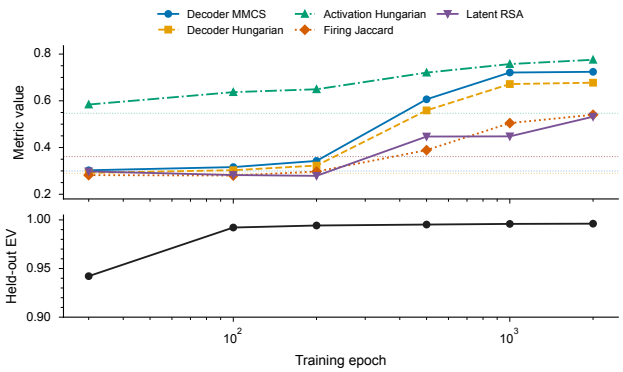

Figure 2. Held-out multi-metric audit. Decoder metrics rise strongly by 2,000 epochs; activation-pattern and firing-overlap agreement rise later and more weakly. Dotted lines are metric-specific random-SAE baselines.

try. On mod-113, reconstruction loss saturates within about 20 epochs, but TopK decoder MMCS continues rising until approximately 1,000 epochs, asymptoting near 0.73. ReLU plateaus lower, near 0.66. Figure 1 shows the benchmark failure mode: selecting by reconstruction alone would declare the SAE converged long before decoder reproducibility stabilizes.

4.2. A Single Stability Metric Is Not Sufficient

Figure 2 evaluates the same TopK checkpoints with held-out metrics. At epoch 2,000, decoder MMCS is 0.724 versus 0.300 random, and decoder Hungarian matching is 0.678 versus 0.289 random. Activation-pattern matching also rises (0.776 versus 0.547 random), but firing overlap is weaker (0.540 versus 0.361 random) and improves later.

The causal audit gives the same warning. For the 24 most active held-out reference features, decoder-Hungarian matches produce higher causal-effect correlations than active-random target features, but the absolute correlations remain modest: 0.037 versus 0.003 at epoch 30, 0.183 versus 0.001 at epoch 500, and 0.290 versus -0.006 at epoch 2,000. Thus geometric matching contains some functional signal, but does not establish causal feature identity.

4.3. Language-Model Scope Checks

The same evaluation discipline transfers to selected language-model layers. In the 500-epoch Pythia-410M layer-21 TopK audit with 5 seeds and a held-out activation split, explained variance is 0.947, decoder MMCS is 0.521 versus 0.107 random, and decoder Hungarian is 0.453 versus 0.104 random. Activation-pattern and firing-overlap Hungarian scores also exceed their baselines (0.560 versus 0.166; 0.618 versus 0.091). Additional selected-layer extensions through Pythia-1B/1.4B show decoder metrics above random, but weak firing overlap in some layers; they should be read as scope checks, not a scaling law.

Table 3. Additional selected held-out Pythia TopK checks/extensions. Cells report absolute value / random baseline. Rows differ in layer, epoch, activation count, and SAE width, so they are not monotonic scaling-law points.

Model	L	Ep.	Acts	SAE	EV	Dec. MMCS	Dec. Hung.	Act. Hung.
70M	4	1000	200K	1024/32	0.679	0.393/0.151	0.328/0.146	0.481/0.134
160M	10	1000	200K	1024/32	0.662	0.354/0.124	0.299/0.120	0.516/0.139
410M	21	1000	200K	1024/32	0.947	0.608/0.107	0.556/0.104	0.589/0.166
1B	12	500	100K	2048/32	0.957	0.547/0.080	0.470/0.077	0.385/0.177
1.4B	12	500	100K	2048/32	0.972	0.421/0.080	0.358/0.077	0.278/0.170

4.4. Controls for Benchmark Artifacts

Two controls illustrate why reproducibility metrics need stress tests. First, a streaming-data control on Pythia-410M compares repeated cached activations against online activation extraction. At 200 step-equivalents, cached training reaches MMCS 0.458 ($4.26\times$ random), while streaming reaches 0.418 ($3.89\times$ random). Cache repetition therefore explains part, but not most, of the above-random signal.

Second, dead-neuron resampling can manufacture misleading agreement. In the cached mod-113 protocol, resampling spikes decoder MMCS to 0.888 at epoch 10, then decays below the unregularized baseline by epoch 2,000. A Pythia streaming-resampling stress test is more severe: online resampling produces catastrophic optimization instability while decoder MMCS can still increase. Stability metrics must therefore be interpreted together with reconstruction and held-out diagnostics.

4.5. Synthetic Ground-Truth Calibration

In the recoverable synthetic sparse-superposition regime, MMCS and GT-MCC rise together (Table 4). At epoch 1,000, MMCS is $3.21\times$ random and GT-MCC is $3.49\times$ random, with Pearson correlation 0.997 across training horizons. This is a positive control for the metric pipeline, not proof that MMCS establishes correctness on real model

Table 4. Synthetic positive control. MMCS and GT-MCC rise together when planted decoder columns are recoverable. Random baselines: MMCS 0.2655, GT-MCC 0.2539.

Epoch	MMCS	GT-MCC	MMCS/rand	GT/rand
5	0.334	0.424	1.26×	1.67×
20	0.583	0.680	2.20×	2.68×
100	0.839	0.880	3.16×	3.46×
1000	0.852	0.887	3.21×	3.49×

activations where the ground truth is unknown.

5. Benchmarking Lessons

Report nulls, not only raw stability. Overcomplete dictionaries have substantial random nearest neighbors. A raw MMCS of 0.30 is near random for $d = 128, m = 1024$, but far above random for other dimensions. Dimension- and dictionary-size-specific nulls are necessary for interpretation.

Separate geometry from feature correctness. Decoder MMCS, decoder Hungarian, activation matching, firing overlap, causal effects, and GT-MCC ask different questions. In our runs, decoder geometry improves earlier and more strongly than firing or causal agreement. A benchmark that reports only reconstruction and decoder similarity can overstate feature reliability.

Treat pairwise seed scores carefully. Seed-pair metrics are dependent. They are useful descriptive summaries, but confidence statements should resample whole seeds or use permutation tests over seed labels.

Stress-test training-pipeline artifacts. Dead-neuron resampling can increase apparent stability while harming reconstruction. Streaming and held-out controls are cheap compared with the risk of mistaking a training artifact for reproducible structure.

6. Limitations

The primary controlled setting is mod-113 with $d_{\text{model}} = 128$; Pythia experiments are selected-layer checks through 1B/1.4B, below frontier-scale SAE practice and not a harmonized scaling law. The causal diagnostic is mod-113-only and modest in absolute correlation. We do not include human-label, auto-interpretability, downstream-task, or Pythia causal validation. The synthetic benchmark is a positive control in one recoverable regime and does not establish general correctness. These limitations are why we present the work as a benchmark/evaluation protocol rather than a claim that long training recovers semantic features.

7. Conclusion

SAE reconstruction can converge before decoder reproducibility, and decoder reproducibility can improve before functional or semantic identity is established. For benchmark design, the implication is direct: SAE evaluations should be null-calibrated, multi-metric, held out, and explicit about which notion of reproducibility they measure. This discipline turns seed instability from a vague concern into an auditable empirical object.

Generative AI Assistance

Generative AI tools assisted with editing, code/reproducibility auditing, workshop triage, and compliance checks. All reported numerical results were produced by experiment scripts and frozen evidence manifests.

References

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *ICML*, 2023.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Chanin, D. and Garriga-Alonso, A. SynthSAEBench: Evaluating sparse autoencoders on scalable realistic synthetic data. *arXiv preprint arXiv:2602.14687*, 2026.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders. *arXiv preprint arXiv:2406.04093*, 2024.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *ICLR*, 2023.

Paulo, G. and Belrose, N. Sparse autoencoders trained on the same data learn different features. *arXiv preprint arXiv:2501.16615*, 2025.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfit-

ting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Song, X., Muhamed, A., Zheng, Y., Kong, L., Tang, Z., Diab, M. T., Smith, V., and Zhang, K. Position: Mechanistic interpretability should prioritize feature consistency in sparse autoencoders. *arXiv preprint arXiv:2505.20254*, 2025.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

A. Supplementary Reproducibility Note

The anonymous supplement includes frozen JSON evidence manifests, the figure-generation script, the paper-value validator, selected experiment entry points, and minimal source code. Large Pythia activation caches and checkpoints are omitted for size; manifests record model, layer, activation-count, seed, and metric provenance. The paper-value validator checks high-risk numbers against frozen manifests rather than mutable rolling outputs.