NF-SAE: TESTING AN NMF-INSPIRED SPARSE AUTOENCODER ACROSS LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks store information as long lists of numbers called activations. Sparse autoencoders try to rewrite each list using only a few learned features. Inspired by nonnegative matrix factorization (NMF), a method that builds examples by adding nonnegative parts, we test an autoencoder whose main feature strengths cannot be negative. We call it NF-SAE. We train the same overall recipe on internal activations from three language models and compare it with a signed magnitude-TopK control. NF-SAE looks promising on Gemma, but the result does not carry over to Pythia or Qwen. BatchTopK has lower mean reconstruction and output errors in every setting where we test it. The clearest practical concern is that one fixed recipe behaves very differently across models of different sizes and scales; this study cannot show which design choice caused that difference. NF-SAE therefore remains an idea to test, not a generally better method. The main lesson is simple: success on one model should be checked with matched controls before making a broad claim.

1 INTRODUCTION

A neural network activation is a list of numbers that records what is happening at one point inside a model. A sparse autoencoder (SAE) learns a large set of feature directions and tries to rebuild each activation using only a few of them. Researchers hope that these learned features will make model internals easier to study, although good reconstruction alone does not prove that a feature is meaningful (Karvonen et al., 2025).

Many commonly used SAE variants keep feature strengths nonnegative (Zhu et al., 2026). To test one-sided strengths against a simple alternative, our main control instead keeps the largest values by magnitude and allows their strengths to have either sign. A negative value flips the direction of its feature. We ask whether it helps to make the main NF-SAE feature strengths one-sided: they may be zero or positive, but not negative. The idea comes from nonnegative matrix factorization (NMF), which represents data by adding nonnegative building blocks (Lee & Seung, 1999). We call our NMF-inspired model *NF-SAE*.

Our original Gemma experiment looked encouraging. We therefore carried the same full recipe to Pythia and Qwen. The favorable pattern did not repeat. BatchTopK also had lower mean reconstruction and output errors than NF-SAE in every Gemma and Pythia setting where we tested it. Many features did not activate on the fixed evaluation packets, and training behaved differently across the three models.

These outcomes change the claim. This paper does not present NF-SAE as a winning architecture. Instead, it reports a careful transfer and failure audit: what worked once, what failed to repeat, and what should be controlled next. The evidence points first to a recipe that was not equally suited to all three models. It does not isolate nonnegativity as the cause.

Our contributions are:

- We explain the simple part of NMF that NF-SAE keeps: its main branch adds features with nonnegative strengths.
- We test the complete recipe on three language models and report the favorable, unfavorable, and unstable results together.

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

- We show why the current evidence is not a general method win: the models were not matched in relative size, training behavior, parameters, or compute, and many features did not activate on the fixed evaluation packets.

2 FROM NMF TO NF-SAE

2.1 THE IDEA FROM NMF

For a matrix of nonnegative data V, classical NMF looks for two nonnegative matrices whose product approximately rebuilds it:

$$\min_{H,U\geq 0} \|V - HU\|_F^2. \tag{1}$$

In plain language, each example is rebuilt by adding learned parts; it is not rebuilt by subtracting a part. This additive picture is the main inspiration for NF-SAE.

NMF has been useful for selected image and text representations, including faces and term–document matrices (Lee & Seung, 1999; Xu et al., 2003). It has also been applied to nonnegative CNN features and constructed protein-feature matrices (Collins et al., 2018; Okun & Priisalu, 2006). These examples show that NMF can be useful in several domains. They do *not* show that all vision or protein activations are naturally additive, or that language is a bad domain for NMF.

2.2 WHAT NF-SAE KEEPS—AND WHAT IT CHANGES

Transformer activations contain both positive and negative numbers, so classical NMF does not apply directly. NF-SAE keeps only one idea: the strengths in its main feature branch are nonnegative. The learned feature directions themselves may contain either sign, and a smaller signed branch can repair the reconstruction. A simple view is

$$\hat{x} = b + \sum_{i\in A} z_i q_i + \text{signed correction}, \qquad z_i \geq 0. \tag{2}$$

A main-branch feature cannot flip its own direction, but different features can still cancel each other. NF-SAE is therefore *inspired by* NMF; it is not classical NMF and does not inherit NMF guarantees. Appendix A gives the complete equations and training objective.

Semi-NMF also allows signed directions with nonnegative coefficients (Ding et al., 2010), and nonnegative sparse coding combines reconstruction with sparse nonnegative codes (Hoyer, 2002). Recent work has applied semi-NMF to language-model activations (Shafran et al., 2026). Our components are not claimed as individually novel. The contribution here is the cross-model audit of one complete recipe and its failures.

3 WHAT WE TESTED

We trained NF-SAE on one internal layer from each of three language models: Gemma-2-2B, Pythia-410M, and Qwen2.5-7B-Instruct. Each model had three paired training runs. Within a pair, every method saw the same training examples in the same order. We tested reconstruction on held-out FineWeb and WikiText text.

The main comparison is a signed magnitude-TopK (AbsTopK-style) control with the same number of learned features and the same target number of active features. Gemma and Pythia also include nonnegative TopK and BatchTopK baselines (Gao et al., 2025; Bussmann et al., 2024; Zhu et al., 2026). BatchTopK allocates a shared activity budget across a batch and has its own training recipe.

We ask three basic questions:

1. How closely does the SAE rebuild the original activation?
2. How much does replacing the activation change the language model's output distribution?
3. How much does it worsen next-token prediction loss?

Table 1: Plain-language summary of NF-SAE versus the equal-width signed magnitude-TopK control. "Output measures" are changes to the language model's predictions and loss.

Model	Test text	What happened?
Gemma	FineWeb	NF-SAE was better on reconstruction and both output measures.
Gemma	WikiText	Reconstruction was better; output differences were uncertain.
Pythia	FineWeb	Results were similar or slightly worse for NF-SAE.
Pythia	WikiText	NF-SAE was worse on all three measures.
Qwen	FineWeb/WikiText	NF-SAE was worse and reconstruction varied greatly by seed.

Lower is better for all three. We also record unused features, training time, and memory. Exact definitions, uncertainty intervals, model layers, optimizer settings, and all selected learning rates are in Appendix B.

An important fairness warning. The comparison is useful but not perfectly matched. A fixed absolute width and activity target mean different relative capacities for models of different sizes. NF-SAE also has more matrix parameters than the equal-width signed control. The Gemma outcome influenced recipe development, and the final plan was internally frozen only after some runs were already under way. We therefore compare complete observed configurations; we do not claim that nonnegativity alone caused any difference.

4 Results

Table 1 gives the central result without the full metric table. Exact values, seed-level plots, and uncertainty intervals appear in Appendix C.

The Gemma result did not transfer. NF-SAE improved all three average measures on Gemma/FineWeb and improved reconstruction on Gemma/WikiText. On Pythia, that pattern disappeared. On Qwen, NF-SAE was worse on average and one run failed especially badly. We kept that run rather than hiding or replacing it.

BatchTopK had the lowest tested mean errors. On Gemma and Pythia, BatchTopK had lower mean reconstruction error and lower mean changes to predictions and loss than NF-SAE in every tested model–dataset pair. The comparison is not perfectly controlled because its activity level shifted on some test data and its training recipe differs. It is therefore an informative operating-point comparison, not proof that BatchTopK is best under every resource or sparsity match.

Many features were inactive on the test packets. More than half of NF-SAE's features never activated on an evaluation packet; the signed magnitude-TopK control had the same broad problem. BatchTopK used its dictionary much more fully. A model can reconstruct reasonably well while still learning a poorly used dictionary, so these experiments do not establish interpretability.

NF-SAE also cost more. The equal-width NF-SAE had more dominant matrix parameters, used more memory, and took longer to train than signed magnitude-TopK. These are implementation-specific measurements, but they rule out an equal-resource advantage in this study. Full utilization and cost tables are in Appendix D.

5 What may explain the mixed result

The leading concern is recipe mismatch. The same absolute width, activity target, regularization, exposure, branch split, and clipping threshold were reused across models with different activation sizes and scales. NF-SAE also triggered gradient clipping much more often outside Gemma. Every method family received the same small validation-only learning-rate sweep, so the issue is not simply that learning rate was never tuned. Rather, the broader recipe was not scaled or matched to each model. This is the most immediate evidence-based concern, but it is not a proven cause; clipping, for example, may be a symptom rather than the source of failure.

This is not a "vision good, text bad" result. NMF already has useful text applications when the input is a nonnegative word-count matrix (Lee & Seung, 1999; Xu et al., 2003). The more careful future question is whether a particular internal representation is well described by adding one-sided features. Vision and protein models could provide useful new test cases, but we have no results showing that NF-SAE will work better there.

Geometry is a future hypothesis, not a finding. One-sided feature strengths may struggle when the same idea often appears in opposite directions, or when useful reconstructions require cancellation. We did not measure those properties. Future work should define simple geometry tests in advance and ask whether they predict NF-SAE's relative performance. The technical versions of these hypotheses are collected in Appendix E.

What should be tested next. The next study should match parameter count, compute, relative width, and activity level; tune all methods with equal budgets; test several sparsity levels; and use stronger dictionary-health checks. Only then should it isolate which NF-SAE component matters. Tests of feature meaning and causal usefulness should come after the dictionaries are healthy.

6 Limitations and conclusion

This study covers one layer in each of three language models, two English-text test sets, and three training seeds. It does not sample a broad population of layers, documents, or models. The methods are not matched for parameters, compute, or every training choice. The recipe was shaped by the earlier Gemma result, and the final completion plan was not a public preregistration. No vision, protein, or vision-language experiment is reported.

NF-SAE asks a reasonable question inspired by NMF: can an SAE benefit when its main feature strengths cannot reverse direction? In our tests, the complete recipe looked favorable on Gemma but did not transfer to Pythia or Qwen, and BatchTopK had lower mean reconstruction and output errors where tested. The evidence does not show that the one-sided idea is wrong; it shows that the current recipe is not yet a reliable general method. A fairer, healthier, and more carefully scaled experiment is needed before making a stronger claim.

AI use statement

AI was used for the following purposes:

- Yes, to aid or polish writing. Details are described in the paper.
- Yes, for retrieval and discovery (e.g., finding related work). Details are described in the paper.
- Yes, for research ideation or execution. Details are described in the paper.

Ethics statement

The study uses model activations and text datasets, not human-subject interaction. Dataset licenses, privacy, redistribution rights, and model terms remain relevant to any release. The anonymous evidence bundle therefore excludes raw text, activations, model weights, SAE checkpoint bytes, machine identifiers, and identity-bearing paths.

Reproducibility statement

The supplement contains the final configurations, derived results, model-selection receipts, environment locks, analysis code, and tests. Its standalone verifier checks file hashes and rebuilds the reported comparisons. This supports analysis-level reproduction, not a complete rerun of training or evaluation: raw text, activations, model weights, and SAE checkpoints are withheld pending rights review. Appendix F gives the exact boundary.

REFERENCES

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024. doi: 10.48550/arXiv.2412.06410. URL https://arxiv.org/abs/2412.06410.

Edo Collins, Radhakrishna Achanta, and Sabine Süsstrunk. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision*, pp. 352–368, 2018. doi: 10.1007/978-3-030-01264-9_21. URL https://doi.org/10.1007/978-3-030-01264-9_21.

Chris H. Q. Ding, Tao Li, and Michael I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010. doi: 10.1109/TPAMI.2008.277. URL https://pubmed.ncbi.nlm.nih.gov/19926898/.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *Advances in Neural Information Processing Systems*, volume 16, 2003. URL https://proceedings.neurips.cc/paper/2003/hash/1843e35d41ccf6e63273495ba42df3c1-Abstract.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arXiv.2406.04093. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Patrik O. Hoyer. Non-negative sparse coding. *arXiv preprint cs/0202009*, 2002. doi: 10.48550/arXiv.cs/0202009. URL https://arxiv.org/abs/cs/0202009.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29223–29264, 2025. doi: 10.48550/arXiv.2503.09532. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999. doi: 10.1038/44565. URL https://www.nature.com/articles/44565.

Oleg Okun and Helen Priisalu. Fast nonnegative matrix factorization and its application for protein fold recognition. *EURASIP Journal on Advances in Signal Processing*, 2006:071817, 2006. doi: 10.1155/ASP/2006/71817. URL https://doi.org/10.1155/ASP/2006/71817.

Or Shafran, Atticus Geiger, and Mor Geva. Constructing interpretable features from compositional neuron groups. *arXiv preprint arXiv:2506.10920*, 2026. doi: 10.48550/arXiv.2506.10920. URL https://arxiv.org/abs/2506.10920.

Wei Xu, Xin Liu, and Yihong Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–273, 2003. doi: 10.1145/860435.860485. URL https://doi.org/10.1145/860435.860485.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. AbsTopK: Rethinking sparse autoencoders for bidirectional features. In *International Conference on Learning Representations*, 2026. doi: 10.48550/arXiv.2510.00404. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/527360cae239e340eb056e32b01f1caf-Abstract-Conference.html.

A FULL NF-SAE DEFINITION AND IMPLEMENTATION

For residual activation $x \in \mathbb{R}^d$ and learned pre-bias b, let $c = x - b$ and use the lossless sign-split lift

$$\phi(c) = [c_+, (-c)_+] \in \mathbb{R}^{2d}_+, \qquad c = c_+ - (-c)_+. \tag{3}$$

NF-SAE encodes a nonnegative factor branch and signed correction branch:

$$z_N = \mathrm{TopK}_+(\phi(c)W_N + a_N, 40), \tag{4}$$

$$z_S = \mathrm{TopK}_{|\cdot|}(cW_S + a_S, 16), \tag{5}$$

$$\hat{x} = b + (z_N \odot g)Q + z_S D_S. \tag{6}$$

TopK_+ applies ReLU and retains the 40 largest values. $\mathrm{TopK}_{|\cdot|}$ selects 16 entries by magnitude while preserving their signs. The branches have $k_N = 4{,}096$ and $k_S = 12{,}288$ coordinates, with $W_N \in \mathbb{R}^{2d \times k_N}$, $a_N \in \mathbb{R}^{k_N}$, $W_S \in \mathbb{R}^{d \times k_S}$, $a_S \in \mathbb{R}^{k_S}$, and $D_S \in \mathbb{R}^{k_S \times d}$. Let the learned lifted decoder be $D_N = [D_N^{(1)}, D_N^{(2)}] \in \mathbb{R}^{k_N \times 2d}$. The ith row of $Q \in \mathbb{R}^{k_N \times d}$ is the unit-normalized fold $D_{N,i}^{(1)} - D_{N,i}^{(2)}$, and $g_i = \mathrm{softplus}(r_i) > 0$. Encoders and decoders are learned independently after initialization. Thus z_N contributes only positive multiples of each signed Q_i; other atoms and $z_S D_S$ may still cancel it coordinatewise. The artifact name for NF-SAE is C6.

The training objective is

$$\mathcal{L} = \mathrm{MSE}(x, \hat{x}) + \lambda_N \|z_N\|_1 / k_N + \lambda_S \|z_S\|_1 / k_S + \gamma \operatorname*{mean}_{i,j} (Q_i^\top \bar{D}_{S,j})^2, \tag{7}$$

where $\bar{D}_{S,j}$ is a unit signed-branch direction, $\lambda_N = 4 \times 10^{-4}$, $\lambda_S = 8 \times 10^{-4}$, and $\gamma = 10^{-3}$. This is an SAE objective, not NMF optimization. The final term discourages branch redundancy but does not enforce orthogonality. Signed decoder rows are renormalized after every update. Lifted decoder atoms are unconstrained; only their folded directions are normalized at decode, with scale represented by g. The 40/16 activity split and 1:3 branch-width split came from exploratory Gemma development and are not ablated here.

Projected anchor initialization. Let $A = \phi(C)$ be the sign-split initialization buffer. For each pass p, draw a deterministic Gaussian projection $R_p \in \mathbb{R}^{2d \times r}/\sqrt{r}$ and set $Y = AR_p$. Repeatedly select the unselected row of Y with greatest squared norm, then replace every row y by $y - (y^\top u/\|u\|_2^2)u$ for selected row u. Each pass selects at most r anchors. Here $r = \max(64, \min(k_N, \max(64, 2d/4)))$ and the number of passes is $\lceil k_N/r \rceil$. If degeneracy leaves slots unfilled, the largest-norm unselected rows of A fill them. This is an initialization heuristic, not evidence that the data meet an NMF separability condition (Donoho & Stodden, 2003).

Reseeding. Cumulative absolute activity is updated after each batch. Once 300,000 rows have been seen, a branch is eligible when the fraction of coordinates with cumulative activity at most 10^{-10} exceeds its trigger (0.20 for N, 0.85 for S). Up to 512 eligible coordinates, ordered by index, are reset from deterministic batch rows. BatchTopK instead uses its own ten-million-row dead threshold and AuxK path.

B DETAILED EXPERIMENTAL DESIGN

Representations and scale. We extract post-block residuals before the next block at layer 12 of Gemma-2-2B and Pythia-410M-deduped, and layer 14 of Qwen2.5-7B-Instruct. Their residual dimensions d are 2,304, 1,024, and 3,584. Fixed width 16,384 therefore gives expansion ratios 7.11, 16.00, and 4.57; factor-branch width 4,096 gives 1.78, 4.00, and 1.14 times d; and nominal L_0/d is 2.43%, 5.47%, and 1.56%. These are important cross-model confounders. Each SAE consumes 49,999,872 sampled activation-row exposures from balanced ArXiv, FineWeb-Edu, GitHub, Pile-CC, and PubMed Abstract packets.

Paired runs. One paired replicate is a seed in $\{47, 48, 49\}$. The seed changes the training packet, initialization, and row order. Within a model and seed, every method receives the same packet and ordered rows. Each model/domain has one fixed test packet shared by all seeds and methods. Thus $n = 3$ measures paired training variation, not independent variation over documents, layers, or models.

Optimization and tuning. All non-BatchTopK configurations use Adam, 500-step linear warmup, no decay, and gradient clipping at 1. Every model–method family received the same frozen, validation-only, one-seed learning-rate sweep over $\{10^{-4}, 2 \times 10^{-4}, 4 \times 10^{-4}\}$; the selector primarily minimized validation MSE. It selected NF-SAE/signed magnitude-TopK rates of $4/4$, $4/2$, and $2/2 \times 10^{-4}$ on Gemma, Pythia, and Qwen. BatchTopK uses a 1,000-step warmup, no L_1 penalty, AuxK weight .03125, and no generic reseeding. Mean clipped steps per run for NF-SAE versus signed magnitude-TopK were 463/194 on Gemma, 1,945/526 on Pythia, and 8,289/6,074 on Qwen, out of 12,207 steps. No nonfinite step was recorded.

Table 2: Chronology and post-selection boundary.

Stage	Choices or observations	Status
Exploration	Favorable Gemma result; branch activity/widths; initialization and tuning choices	Outcome-aware development
Final plan	Models, layers, seeds, methods, test packets, endpoints, and effect direction fixed after 21 partial and five completed runs existed	Hash-fixed completion plan; not preregistered
Execution	21 runs continued, four started fresh, five completed runs reused; all 30 received unified evaluation	Configurations verified
Paper	Transfer-audit framing, NF-SAE name, diagnostic emphasis, and geometry/vision/protein hypotheses	Retrospective

Evaluation and intervals. Every checkpoint is evaluated on fixed 2,048-sequence FineWeb and WikiText packets at context length 128 (about 259,000 eligible rows each). We report elementwise residual MSE; token-weighted $D_{\mathrm{KL}}(p_{\mathrm{original}} \| p_{\mathrm{recon}})$ divided by mean zero-ablation KL; reconstructed minus original next-token cross-entropy; realized L_0; dead-feature fraction; fraction of variance explained (FVE); time; and memory. For N rows of dimension d, FVE is

$$1 - \frac{\sum_{n,j}(x_{nj} - \hat{x}_{nj})^2}{\sum_{n,j}(x_{nj} - \bar{x}_{\mathrm{scalar}})^2}, \quad \bar{x}_{\mathrm{scalar}} = (Nd)^{-1} \sum_{n,j} x_{nj}. \tag{8}$$

A feature is dead when its code is zero for every eligible evaluation row. For each lower-is-better endpoint, the paired relative effect is $100(y_{\mathrm{signed}} - y_{\mathrm{NF}})/y_{\mathrm{signed}}$. We report the mean over three seeds and a descriptive two-sided Student-t interval with two degrees of freedom. No row is imputed; all 60 evaluation records and six BatchTopK calibration records are present.

C Exact reconstruction and output results

Table 3: Means over three paired training replicates. NF denotes NF-SAE; STK denotes the equal-width signed magnitude-TopK control. Loss increase is nats/token. Lower is better except FVE.

Model	Domain	Method	MSE	Norm. KL	Loss inc.	FVE
Gemma	FineWeb	NF	2.5932	.0453	.4663	.7703
		STK	2.6813	.0491	.5096	.7625
Gemma	WikiText	NF	2.6964	.0808	.7504	.7621
		STK	2.7507	.0836	.7824	.7573
Pythia	FineWeb	NF	.1071	.0833	.6175	.9871
		STK	.1057	.0795	.5886	.9872
Pythia	WikiText	NF	.1760	.1177	.9571	.9882
		STK	.1478	.1106	.8917	.9901
Qwen	FineWeb	NF	4.2782	.0923	.8294	.9240
		STK	.8134	.0827	.7358	.9856
Qwen	WikiText	NF	25.2040	.1644	1.3420	.9403
		STK	2.1437	.1406	1.1178	.9949

Qwen MSE values across seeds were 7.249, 2.092, and 3.493 on FineWeb for NF-SAE, versus 0.905, 0.800, and 0.735 for signed magnitude-TopK; WikiText values were 50.289, 11.686, and 13.637, versus 2.733, 2.099, and 1.598. The extreme seed is retained.

D Baselines, utilization, and cost

NF-SAE nearly reaches its overall per-example activity cap, but that total does not establish balance between its branches. Mean training time for NF-SAE versus signed magnitude-TopK is 2.41/1.54 hours on Gemma, 1.25/.85 on Pythia, and 2.14/1.52 on Qwen. Mean per-run peak allocated memory is 4.37/2.77 GB, 3.17/2.02 GB, and 5.17/3.21 GB. NF-SAE has $40960d$ dominant encoder/decoder matrix parameters, compared with $32768d$ for equal-width signed magnitude-TopK, a 25% difference. A width-20,480 signed parameter-count control was not run.

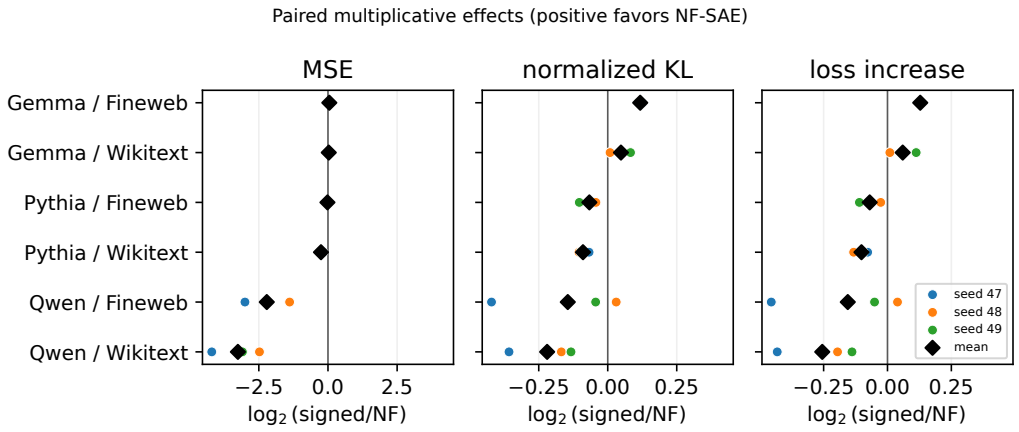

Figure 1: Paired seed effects, shown as $\log_2(\mathrm{STK}/\mathrm{NF})$: positive favors NF-SAE. Each point is one paired training replicate; diamonds are means. All seeds share a fixed test packet within each model/domain.

Table 4: NF-SAE versus signed magnitude-TopK: mean paired relative effects (percent) with descriptive 95% seed intervals. Positive favors NF-SAE; $n = 3$.

Model	Domain	MSE	Norm. KL	Loss increase
Gemma	FineWeb	+3.29 [3.01,3.56]	+7.83 [6.82,8.84]	+8.50 [7.93,9.07]
Gemma	WikiText	+1.98 [1.10,2.85]	+3.24 [−2.98,9.47]	+4.01 [−4.49,12.51]
Pythia	FineWeb	−1.35 [−4.32,1.61]	−4.74 [−10.55,1.08]	−4.97 [−12.52,2.57]
Pythia	WikiText	−19.20 [−30.62,−7.77]	−6.42 [−9.94,−2.89]	−7.33 [−12.52,−2.14]
Qwen	FineWeb	−412.56 [−1087.38,262.26]	−11.62 [−59.96,36.71]	−12.65 [−65.75,40.45]
Qwen	WikiText	−983.19 [−2651.89,685.52]	−16.72 [−41.51,8.08]	−19.83 [−52.65,12.99]

Against equal-width nonnegative TopK, NF-SAE has favorable mean effects on Gemma: MSE +5.45% (FineWeb) and +3.66% (WikiText), and normalized KL +29.35% and +25.70%. It is unfavorable or unresolved on Pythia: MSE −5.85% and −2.55% with intervals crossing zero, and normalized KL −26.76% and −26.41%. Exact seed effects are in the machine-readable supplement.

E Technical interpretation and future tests

Geometry hypotheses. Signedness alone cannot explain the pattern: all three sites are signed, yet Gemma is favorable. NF-SAE uses signed factor directions and a signed correction, so the technical question is whether the centered distribution is efficiently covered by a sparse positive cone of signed atoms. The sign split does not create parts, separability, or a privileged conic basis. Contextual residuals combine token-, position-, and computation-dependent information, while transformer blocks may add and cancel signed updates. A one-sided code may require separate atoms for opposite orientations, and the correction branch permits cancellation. Positive-cone coverage, antipodal-atom demand, and cross-branch cancellation were not measured.

Even an elementwise nonnegative raw site can become signed after learned centering, while NF-SAE's decoder and correction remain signed. A valid cross-domain study should therefore vary centering and decoder constraints and match backbone, data, dimension or expansion ratio, scale, tuning budget, and evaluation metric. An improvement on an unrelated vision or protein task would otherwise mix geometry with architecture, data, and task.

Decisive next tests. A prospective study should add a width-20,480 signed parameter-count control; broader symmetric tuning; multiple sparsity points with BatchTopK brackets; matched expansion and relative-sparsity points; and stronger utilization controls. A factorial study should isolate sign splitting, coefficient sign, centering, decoder signs, branch allocation, initialization, gains, regularization, and reseeding. It should pre-specify positive-cone coverage, antipodal demand, and cross-branch cancellation diagnostics and test whether relative performance follows them across layers and domains. Vision-transformer sites and protein single/pair or contact-probability representations can extend the test only after within-domain controls are established.

Table 5: NF-SAE (N) and BatchTopK (B) means. Loss is nats/token. BatchTopK has lower error in all cells, but test L_0 shifts.

		MSE		Norm. KL		Loss		L_0	
Model	Domain	N	B	N	B	N	B	N	B
Gemma	FineWeb	2.593	2.474	.0453	.0357	.466	.360	55.99	56.14
Gemma	WikiText	2.696	2.594	.0808	.0673	.750	.610	55.99	53.39
Pythia	FineWeb	.1071	.0888	.0833	.0465	.617	.340	56.00	56.11
Pythia	WikiText	.1760	.1404	.1177	.0693	.957	.552	56.00	65.81

Table 6: Dead-feature fraction: mean percentage of coordinates exactly zero throughout an evaluation packet.

Model	Domain	NF-SAE	signed magnitude-TopK	BatchTopK
Gemma	FineWeb	57.0	67.9	2.3
Gemma	WikiText	58.8	68.4	8.6
Pythia	FineWeb	86.3	87.3	1.6
Pythia	WikiText	87.1	88.4	3.6
Qwen	FineWeb	75.8	81.6	–
Qwen	WikiText	79.4	82.9	–

Precursor anchor audit. A precursor, nonconfirmatory anchor audit supplied no positive $k = 4{,}000$ separability evidence: successive-projection anchors did not clearly beat random anchors and were often weaker than rotated controls. Both registered conditions remained false and the formal decision stayed `PENDING_RATINGS`. This does not prove nonseparability or test NMF generally; it prevents the initialization heuristic from being mistaken for evidence of an NMF-identifiable geometry.

F Reproducibility and claim boundary

Pinned representations. Gemma uses `google/gemma-2-2b` revision `c5ebcd40d208330abc697524c919956e692655cf` at layer 12; Pythia uses `EleutherAI/pythia-410m-deduped` revision `c4fc8d586d62df497f1f9b69d66d3ca419992d3e` at layer 12; Qwen uses `Qwen/Qwen2.5-7B-Instruct` revision `a09a35458c702b33eeacc393d103063234e8bc28` at layer 14. Each hook is the post-block output before the next block. Context length is 128 and seeds are 47–49.

Evidence completeness. The final v47 study contains 30 training runs, 60 logical evaluation records, six BatchTopK calibration records, and 51 unique settled physical-meter records; its analysis receipt is `SCIENTIFIC_COMPLETE`. No expected result is missing, duplicated, or imputed. Project accounting records 224.257 GPU-hours including prior development; the 51 final-study meters account for 32.855 GPU-hours. These are accounting totals, not a matched-compute comparison.

Artifact scope. The anonymous supplement verifies its manifest, recomputes all 42 paired effects, reproduces all 36 tuning-candidate eligibility decisions and 12 selected winners, and checks sanitized scientific receipts. It supports the paper's arithmetic and analysis, not a clean-room rerun of training or model splicing. Raw text, activation packets, model weights, and SAE checkpoints are hash-identified but withheld pending rights review.

Supported claims are restricted to observed within-model contrasts at the specified layers, packets, width, nominal sparsity, exposure, and seeds. Unsupported claims include a test of NMF theory; causal attribution to one-sided codes; a Pareto or equal-parameter/FLOP/time advantage; improved feature semantics or interpretability; universal transfer; and protein, vision, or vision-language performance.