MORPHSAE: RESERVED SPARSE SLOTS WITHOUT AUTOMATIC COMPONENT RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding a neural network may require separating its activations into components with different roles. Standard TopK sparse autoencoders learn one sparse feature inventory, so they cannot ensure that every proposed group receives selected features. We introduce **MorphSAE**, an additive sparse autoencoder inspired by **Morphological Component Analysis (MCA)**, a sparse-demixing method from the **Compressed Sensing** and sparse-representation literature. Each named dictionary has its own TopK budget, and the decoded components add to reconstruct the original activation. This design guarantees returned indices from every dictionary. In three matched comparisons, an optional cross-dictionary loss also reduced its sampled overlap statistic. In a controlled synthetic copy task, an explicit separator signal concentrates the task-relevant activation change in a controller dictionary. Compared with a selected conventional-SAE intervention, the combined named and supervised design improves target recovery while reducing collateral error. This demonstrates a useful named decomposition in this supervised synthetic task. A matched model with one shared budget performs similarly, so the quota alone does not explain the success. On natural activations, the tested quota provides no reliable reconstruction or unsupervised-specialization benefit. Because each decoded component remains in the original activation space, the architecture permits it to be passed into another MorphSAE. We hypothesize that future separator functions could support repeated decomposition into smaller, more interpretable dictionaries; this remains untested.

1 INTRODUCTION

A sparse autoencoder represents a neural activation with a few learned feature vectors, often called a **dictionary** (Huben et al., 2024; Gao et al., 2024). A TopK SAE gives all features one shared budget: if $K = 32$, every feature competes for the same 32 selected positions. This is sensible when the goal is one feature inventory. It is less suitable when an experiment requires separate lists from several named groups.

Suppose an audit asks for eight structural candidates and 24 lexical candidates before running two interventions. A global TopK model might spend almost all 32 slots on one group. An empty list would then be ambiguous: was the signal absent, or did that group lose the competition? This example motivates a stricter interface, although the present paper does not show that such an interface improves a natural downstream audit.

We introduce the **Morphological Sparse Autoencoder (MorphSAE)**. "Morphological" refers to **Morphological Component Analysis (MCA)**, a sparse-demixing method from the **Compressed Sensing** and sparse-representation literature, not linguistic morphology. MCA models an observation as a sum of parts represented in different dictionaries (Bobin et al., 2007; Donoho, 2006; Donoho & Huo, 2001; Peyré et al., 2010). MorphSAE borrows this additive form. Several encoders see the same activation, each encoder selects features from its own dictionary, and the decoded branch outputs add to the reconstruction.

The key change is simple but exact: MorphSAE gives every named dictionary its own TopK budget. A **named-global** SAE uses the same named dictionaries and exposes the same partial reconstructions, but all rows still compete under one shared TopK. The two models have the same combined linear decoder span. They differ in which sets of indices they may return.

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.

This distinction separates four questions:

1. **Allocation:** does every group receive its assigned number of returned indices?
2. **Geometry:** does an optional loss make decoder rows from different groups less aligned by its chosen measure?
3. **Reconstruction:** does the quota reduce held-out error?
4. **Function:** do the named groups acquire the roles intended by the researcher?

Only the first answer follows from the architecture. Our experiments show that the optional loss changes its own geometry statistic, but the quota does not reliably improve reconstruction and the tested unsupervised branches do not pass the full functional criterion. The one positive functional example is synthetic and explicitly supervised: named controller and background targets produce useful named components, while the direct quota-versus-named-global contrast is essentially zero.

Our contributions are:

- a grouped-TopK SAE with named additive outputs and fixed per-group returned-index budgets;
- a direct comparator, named-global TopK, that isolates the quota from naming and additive decoding;
- controlled evidence that separates allocation, geometry, reconstruction, and function; and
- a bounded positive example showing that an explicit identifiability signal can organize useful named synthetic components, without showing a quota-specific benefit.

The contribution is a controlled modeling primitive and a set of evidence boundaries, not a claim that MorphSAE should replace a conventional SAE.

2 From Morphological Component Analysis to MorphSAE

2.1 Additive sparse separation

Classical MCA describes an observation as a sum of components:

$$y = \sum_{g=1}^{G} \Phi_g \alpha_g + e,$$

where Φ_g is a dictionary for component g, α_g is a sparse code, and e is residual error. A common noisy sparse-demixing problem is

$$\min_{\{\alpha_g\}} \sum_g \|\alpha_g\|_1 \quad \text{s.t.} \quad \left\| y - \sum_g \Phi_g \alpha_g \right\|_2 \le \varepsilon.$$

In plain language, separation is possible only when the true parts are represented differently enough. A component should be sparse in its intended dictionary and not equally sparse in every other dictionary. MCA calls this **morphological diversity** (Bobin et al., 2007). For a fixed union of two orthonormal bases, Donoho and Huo define cross-basis coherence as $M = \max_{p,q} |\langle \phi_{1,p}, \phi_{2,q} \rangle|$ and give sufficient conditions for unique ℓ_0 and ℓ_1 recovery (Donoho & Huo, 2001). Those are fixed-dictionary, noiseless, solver-specific results. They do not assign semantic meaning to a component. Later MCA work learns dictionaries with a nonconvex procedure, which supplies a precedent for adaptivity but not automatic identifiability (Peyré et al., 2010).

2.2 Where the analogy stops

MorphSAE keeps the additive algebra. If decoder atoms are rows of D_g, then the MCA synthesis dictionary corresponds to $\Phi_g = D_g^\top$, and branch g reconstructs $D_g^\top z_g(x)$. The learning problem is different: MorphSAE learns overcomplete dictionaries and nonlinear encoders together, uses nonnegative coefficients, and forces a fixed selection quota even when a true component may be absent.

> **Theory boundary.** Classical MCA recovery guarantees do not transfer automatically. MorphSAE does not establish a sparse generative decomposition, and its learned encoder is not the optimizer assumed by the classical results. Its optional loss measures a sampled *mean squared cross-dictionary cosine*, not maximum coherence of the complete fixed dictionary. Lower sampled overlap is therefore evidence about one geometry statistic, not a certificate of unique recovery, correct components, or semantic identifiability.

SAEs, k-sparse autoencoders, multi-dictionary sparse models, and structured incoherent dictionary learning already provide the main ingredients around this method (Makhzani & Frey, 2013; Ramirez et al., 2010). Matryoshka SAEs train nested prefixes rather than simultaneous disjoint codes (Bussmann et al., 2025). SAE features can also change across seeds, absorb related concepts, and leave structured residual error (Paulo & Belrose, 2025; Chanin et al., 2025; Engels et al., 2025). MorphSAE's narrower contribution is to make a product of groupwise TopK returned-index constraints an explicit SAE interface. We use *MorphSAE* rather than *MSAE* to avoid confusion with "Matryoshka SAE."

3 METHOD

3.1 ONE BUDGET PER NAMED DICTIONARY

For activation $x \in \mathbb{R}^d$, a bias-free global TopK SAE computes

$$z(x) = \mathrm{TopK}_k(\mathrm{ReLU}(Wx)), \qquad \hat{x} = D^\top z(x).$$

MorphSAE declares groups $j = 1, \ldots, G$:

$$z_j(x) = \mathrm{TopK}_{k_j}(\mathrm{ReLU}(W_j x)), \qquad \hat{x}_j = D_j^\top z_j(x), \qquad \hat{x} = \sum_{j=1}^{G} \hat{x}_j.$$

The controlled study uses this bias-free form. The later natural study adds centering, scalar normalization, encoder biases, and a decoder bias; Appendix A.2 gives the exact form and configuration.

PyTorch `topk` returns exactly k_j indices, with ties resolved by the library implementation. Because TopK follows ReLU, a returned value may be exactly zero. MorphSAE therefore guarantees k_j **returned indices**, not k_j positive coefficients or a nonzero branch reconstruction. In the six original controlled evaluation cells, the observed mean positive counts were exactly 8 and 24, but this is an observation from those cells, not the architectural guarantee.

Concatenating the groups recovers the same linear decoder span as one dictionary. Let $\mathcal{I}_j$ contain the indices in group j, with $k = \sum_j k_j$. Global TopK permits

$$\mathcal{S}_{\mathrm{global}} = \{S \subseteq \cup_j \mathcal{I}_j : |S| = k\},$$

whereas MorphSAE permits only

$$\mathcal{S}_{\mathrm{Morph}} = \{\cup_j S_j : S_j \subseteq \mathcal{I}_j,\ |S_j| = k_j\ \forall j\}.$$

These are sets of returned indices. Their positive support can be smaller.

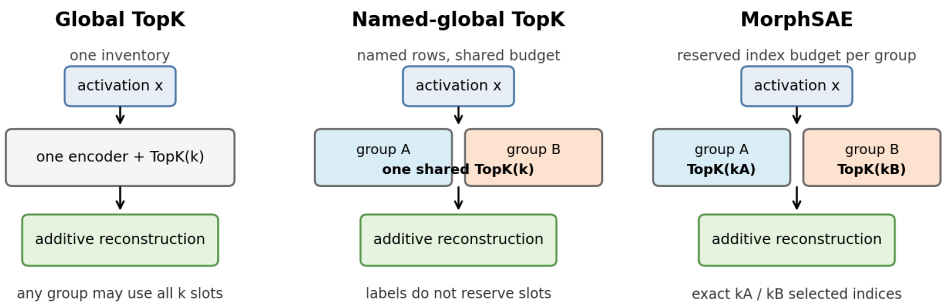


Figure 1: Three sparse-selection rules over the same named row groups. Global TopK uses one inventory. Named-global adds group labels and partial outputs but keeps one shared budget. MorphSAE applies TopK inside every group, so returned indices are reserved by construction.

Property	Global TopK	Named-global TopK	MorphSAE
Sparse competition	one shared budget	one shared budget	one budget per group
Reserved returned indices	no	no	yes
Named partial outputs	optional	yes	yes
Cross-group loss	requires a partition	available	available
Appropriate use	one inventory	named subsets	protected allocation

Table 1. MorphSAE's irreducible difference is protected groupwise competition.

3.2 Optional cross-group loss

For two groups, we optionally add

$$\mathcal{L}_{\text{cross}} = \lambda_{\text{cross}}\, \mathbb{E}_{i\sim I, j\sim J} \langle \bar{d}_{1,i}, \bar{d}_{2,j} \rangle^2,$$

where $\bar{d}$ is a unit-normalized decoder row and each minibatch samples at most 512 rows from each group. This loss penalizes average squared cosine between sampled cross-group rows. It does not measure the maximum, and it omits within-group pairs.

The full non-Matryoshka objective is

$$\mathcal{L} = d^{-1}\|x - \hat{x}\|_2^2 + \mathcal{L}_{\text{cross}}.$$

The method contract is short: MorphSAE guarantees the returned-index allocation $(k_1, \ldots, k_G)$; the optional loss directly optimizes sampled cross-group squared cosine; neither mechanism guarantees better reconstruction, recurring atoms, or semantic ownership.

4 Evaluation

The main natural experiments use residual-stream vectors after block 3 of `Pythia-160m-deduped` (Biderman et al., 2023). The two groups contain 8,192 and 32,768 rows and receive quotas of 8 and 24, for 32 returned indices in total. Named-global uses the same row groups but selects all 32 entries jointly. Fraction of variance unexplained (FVU) is squared reconstruction error divided by activation variance; lower is better.

Question	Direct comparison	Evidence status
Does the optional loss change its target?	loss off versus on, three matched 100M-token pairs	retrospective and nonblind
Does the quota change reconstruction?	MorphSAE versus named-global, six matched 150M-token pairs, with cross loss and dead-row revival off	complete priority subset of an incomplete 45/64 program; fixed-budget and nonconverged
Do unsupervised groups acquire the intended roles?	12 evaluations over nine trained states and three exact relabelings	retrospective functional rule
Can an identifiability signal organize useful groups?	supervised MorphSAE, supervised named-global, and conventional SAE on a synthetic copy task	three fixed subjects by four SAE seeds

Table 2. The four questions use different comparisons and support different conclusions.

The later natural study trains on 128-token sequences from ArXiv, Github, Pile-CC, and PubMed Abstracts. Its in-domain test contains 2,048 documents and 248,596 scored token rows. Its out-of-domain set, called **C2**, contains 436 held-out UD English ESLSpok learner-transcript files, 1,144 inference units, and 9,614 scored rows (Nivre et al., 2020). The exact revisions, splits, optimizer settings, leakage checks, and bootstrap units appear in Appendix A.

The full natural plan stopped at 45 of 64 jobs. The six-pair selector comparison was chosen before its outcomes were inspected. A later baseline package was chosen after those outcomes and the synthetic results were visible, so it is descriptive and result-aware. The older geometry and functional studies are retrospective and nonblind. Intervals resample only the executed seeds, subjects, or experimental blocks; they describe sensitivity within this finite design, not a population of models.

5 Results

5.1 Allocation is exact only at the returned-index level

MorphSAE returns the requested number of indices from every group by definition. Named-global does not provide this guarantee, although whole-group starvation was not observed in the four historical matched pairs. The evidence therefore establishes a capability difference, not a demonstrated repair of starvation in the present experiments.

5.2 The optional loss changes the statistic it targets

With $\lambda_{\text{cross}} = 0.01$, sampled cross-group overlap falls in all three matched pairs (Table 3). The no-loss-to-loss ratios are **2.739, 2.205, and 2.328**. Reconstruction does not move consistently: loss-minus-no-loss FVU is $-0.135, -0.246, +0.036$.

Seed	Overlap, no loss	Overlap, with loss	Reduction	FVU, no loss / with loss
42	0.1640	0.0599	2.739-fold	0.322 / 0.187
43	0.1422	0.0645	2.205-fold	0.337 / 0.091
44	0.1554	0.0668	2.328-fold	0.062 / 0.097

Table 3. Three retrospective, nonblind, matched 100M-token pairs. No across-seed interval is claimed, and convergence is not established. This is an objective-action result, not evidence of component recovery.

5.3 THE QUOTA GIVES NO RELIABLE RECONSTRUCTION ADVANTAGE

With both cross loss and dead-row revival disabled, MorphSAE minus named-global FVU is **0.000023 [-0.000130, 0.000217]** in-domain; four of six fixed seeds favor MorphSAE. On C2, the difference is **0.000387 [0.000025, 0.000758]**, favoring named-global. The C2 interval is a post-hoc sensitivity analysis over 408 transcript-dependency components formed from 436 files; the locally frozen inference-unit interval was **[0.000028, 0.000780]**. None of the 12 runs meets our stability test across checkpoints. We therefore report results after a fixed amount of training, not results at convergence.

The tuned MorphSAE recipe also has higher in-domain FVU than conventional TopK in all six Pythia-160M pairs: **0.000660 [0.000276, 0.001071]**. This comparison was chosen after earlier outcomes were visible. An older comparison that replaces inactive rows favors MorphSAE by **0.040–0.273 FVU** in four of four pairs, but the selector changes which rows are replaced. Neither method shows whole-group starvation, both fail the functional composite, and the broader preference gate fails. This older result mixes the effect of feature selection with the effect of replacing inactive rows; it is not an intrinsic quota advantage.

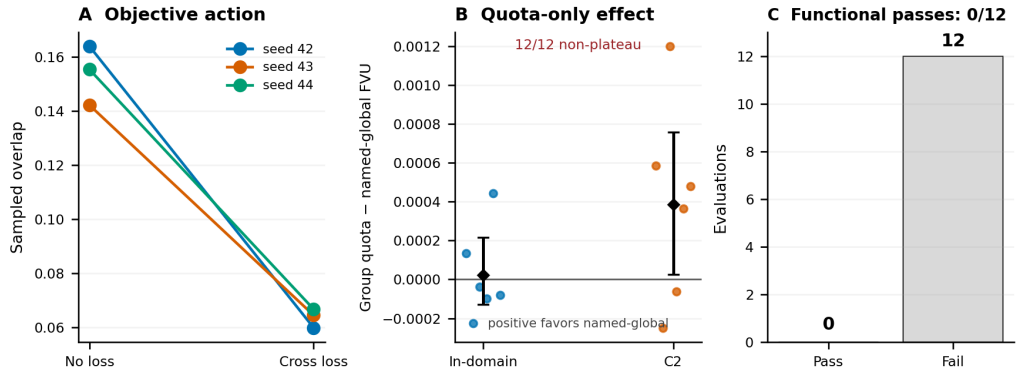


Figure 2: Summary of the main evidence. Left: the optional loss consistently reduces its target. Center: revival-free quota effects on FVU are near zero in-domain and adverse on C2; points are fixed-seed differences and bars are finite-design intervals, with C2 using the post-hoc transcript-cluster sensitivity. Right: none of 12 unsupervised evaluations over nine trained states passes the retrospective nominal composite.

5.4 NAMES AND QUOTAS DO NOT IDENTIFY FUNCTION

The full unsupervised functional criterion passes in **0/12 evaluations**: 0/6 original regularized or unregularized cells, 0/3 independently trained equal-allocation cells, and 0/3 deterministic branch relabelings. These are nine distinct trained states plus three relabelings, not 12 independent training runs. Individual endpoints sometimes pass, but their signs vary across seeds and no configuration passes the conjunction. Exact label swaps leave reconstruction unchanged, showing directly that a branch name does not determine its function.

This is a bounded negative result. It says that asymmetric capacity plus the tested geometry loss is insufficient under this retrospective nominal test. It does not show that semantic decomposition is impossible.

5.5 EXPLICIT SUPERVISION ORGANIZES USEFUL NAMED SYNTHETIC COMPONENTS; THE QUOTA ADDS ALMOST NOTHING

The synthetic copy task provides the paper's one positive functional example. The model has a known controller signal that determines which source token to copy. Training targets explicitly ask one SAE branch to capture the clean-minus-corrupt controller change and ask the other branch to remain stable.

The direct quota comparison comes first. MorphSAE minus branch-supervised named-global target recovery is 2.98×10^{-7} **[**-3.70×10^{-6}**,** 4.10×10^{-6}**]**, with 7 of 12 signs positive. The frozen quota-use gate fails. Thus the supervised named decomposition is useful, but its target-recovery benefit is not attributable to the quota.

Relative to a conventional SAE selected-atom intervention, the combined named/structured and weakly supervised recipe improves target recovery by **0.3966 [0.3648, 0.4290]**. This broader comparison changes named structure, branch supervision, losses, selection, and intervention size together. It is not a supervision-only estimate. Mean intervention L_0 is 29.78 for MorphSAE and 43.19 for named-global, so even the direct quota contrast is not intervention-budget matched. Appendix C gives the task, equations, controls, and all three reported metrics.

Question	Complete evidence	Result	Supported conclusion
Does per-group TopK reserve indices?	method definition	exactly k_j returned indices	allocation guarantee only
Does the optional loss affect overlap?	3 matched pairs	2.205–2.739-fold reduction	yes, for its sampled target
Does quota-only selection lower FVU?	6 revival-free pairs	tied in-domain; worse on C2	no robust advantage
Does tuned MorphSAE beat conventional TopK?	6 Pythia-160M pairs	worse in 6/6	no general quality claim
Do unsupervised groups pass the nominal composite?	12 evaluations / 9 trained states	0/12 retrospective passes	not under tested signals and criterion
Does the named/structured supervised recipe help?	3 subjects $\times$ 4 SAE seeds	+0.3966 target recovery	the bundled recipe helps
Is that benefit quota-specific?	same 12 cells	approximately zero	no detected quota benefit

Table 4. Claim ledger. Each conclusion is limited to the evidence in the second column.

6 USE, INTERPRETATION, AND FUTURE WORK

Use MorphSAE when a fixed number of returned indices from every named group is part of the scientific interface. Use named-global when named partial reconstructions are useful but empty or imbalanced groups are acceptable. Use a conventional TopK SAE when one feature inventory is enough. None of these choices identifies semantic components by itself.

The two-channel audit from the introduction illustrates the conditional value of MorphSAE. It makes a null intervention unambiguous with respect to **allocation**: the requested indices were returned. It does not show that the indices were positive, informative, or correctly named. The present experiments do not demonstrate a natural analysis whose conclusion changes because of this guarantee.

Why publish this finding now? Because it opens a concrete scientific program. The synthetic study shows that an explicit training signal can learn a useful two-part decomposition. The architecture is also compositional: every decoded branch is a vector in the original activation space, so that branch can be passed to another MorphSAE. Repeating this step creates a chain or tree of additive splits.

We hope future work will find and test more **separator functions**: training signals or rules that assign different parts of an activation to different branches. Candidates could include task-specific supervised losses, generative assumptions, or other independently validated signals. If repeated splits isolate distinct feature families, later dictionaries might be smaller and easier to interpret. This paper tests one supervised separator in one synthetic task. It does not train a multi-stage chain, show that dictionaries become smaller, or measure an interpretability gain. Those outcomes are the goal of future work, not claims about the current experiments.

The broader lesson is to report allocation, realized activity, geometry, reconstruction, and function separately. A quota proves allocation. A geometry loss can show action on a geometry measure. Functional ownership still requires an additional identifiability signal and a direct test.

7 LIMITATIONS

- **Incomplete and nonconverged natural program.** The study stopped at 45/64 confirmation jobs, leaving the revival factorial and complete Matryoshka comparison unavailable. All 12 clean selector runs fail the plateau criterion.
- **Outcome visibility.** The objective-action study, historical selector comparison, and functional thresholds are retrospective. Stage 1 was chosen before its own results; Stage 2 was chosen after earlier results were visible. The C2 transcript-cluster correction is post hoc.
- **Small and narrow design.** Intervals condition on a small number of fixed seeds or subjects. Most evidence comes from Pythia-160M block 3, with limited Pythia-70M evidence and poor untuned transfer.
- **Bounded activity evidence.** Positive-support and zero-output distributions were not audited across every headline study, so the guarantee remains at the returned-index level.
- **Bounded functional result.** The 0/12 result uses a retrospective nominal conjunction and only nine independently trained states. It is not an impossibility result.
- **Confounded synthetic comparison.** The conventional contrast bundles structure, supervision, selector, losses, and unmatched intervention size. The quota comparison is also not intervention-L_0 matched.
- **No transferred theorem.** Classical MCA guarantees do not apply to these learned overcomplete dictionaries, nonlinear amortized encoders, or sampled mean-overlap loss.

We do not claim general reconstruction superiority, semantic decomposition, one-to-one feature identity, natural behavioral mediation, or a universal advantage over other structured sparse models.

8 CONCLUSION

MorphSAE replaces one global TopK budget with a separate budget for every named additive dictionary. This guarantees returned indices from every group and allows direct control of a sampled cross-group geometry statistic. In our experiments, those controls do not reliably improve reconstruction or produce unsupervised functional specialization. Explicit supervision does organize useful named components in one synthetic task, but the direct quota effect is approximately zero. MorphSAE should therefore be used for experiments that require protected allocation, while component meaning must be supplied and tested separately.

REPRODUCIBILITY STATEMENT

The compact result object and claim ledger bind the reported numbers to saved, hash-checked artifacts. The build performs no training or model inference. The controlled matrix completed 25/25 jobs; the named-global comparator completed 8/8; the trained-copy study completed 36/36; and the two natural priority analyses cover 12 and 24 complete jobs. The broader natural program is frozen as incomplete at 45/64 confirmation jobs. Representative software is Python 3.12.3, PyTorch 2.7.0, CUDA 12.8, and NumPy 1.26.4 with deterministic PyTorch algorithms enabled. Artifact regeneration verifies saved results, not independent checkpoint replication.

ETHICS STATEMENT

The study analyzes public language-model activations and public dependency-treebank text. It makes no deployment or model-control claim. Spoken-learner transcript groups may contain sensitive linguistic attributes; only aggregate representation measurements are reported.

AI USE STATEMENT

Writing. Yes. AI assistants helped draft and polish prose, reorganize explanations, and improve clarity. The human authors reviewed and edited the resulting text.

Retrieval and discovery. Yes. AI assistants helped search for relevant literature and identify related work. The human authors checked the cited sources and their metadata against the original records.

Research ideation and execution. Yes. AI assistants supported brainstorming, experiment and analysis code, artifact checks, figures, and independent-context critiques. The human authors reviewed the AI-assisted work and retain responsibility for the methods, evidence, claims, and submission.

REFERENCES

Stella Biderman, Hailey Schoelkopf, Quentin 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 *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Jerome Bobin, Jean-Luc Starck, Jalal M. Fadili, Yassir Moudden, and David L. Donoho. Morphological Component Analysis: An adaptive thresholding strategy. *IEEE Transactions on Image Processing*, 16(11):2675–2681, 2007. doi: 10.1109/TIP.2007.907073.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders, 2025.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2025.

David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006. doi: 10.1109/TIT.2006.871582.

David L. Donoho and Xiaoming Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001. doi: 10.1109/18.959265.

Joshua Engels, Logan Smith, and Max Tegmark. Decomposing the dark matter of sparse autoencoders, 2025.

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, 2024.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Alireza Makhzani and Brendan Frey. k-sparse autoencoders, 2013.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4034–4043, 2020.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025.

Gabriel Peyré, Jalal Fadili, and Jean-Luc Starck. Learning the morphological diversity. *SIAM Journal on Imaging Sciences*, 3(3):646–669, 2010. doi: 10.1137/090770783.

Ignacio Ramirez, Pablo Sprechmann, and Guillermo Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3501–3508, 2010. doi: 10.1109/CVPR.2010.5539964.

A Evidence and execution details

A.1 Natural-program timing and completion matrix

The natural configuration and frozen protocol have SHA-256 prefixes `ea56dd19b082` and `1ad731270ac2`; the protocol records status `frozen_before_scientific_training`. The Stage-1 decision was recorded at 2026-09-25 02:20:48 UTC and prohibited Stage 2 until Stage-1 inspection. The N5/N6 analysis was ready at 12:42:45 UTC. At 12:45:51 UTC, after its near-zero in-domain estimate and the failed synthetic quota gate were recorded, Stage 2 selected the stable conventional baseline and second-site scope. These local receipts are hash-bound but not externally timestamped; exact paths appear in the artifact guide.

The locally planned Pythia-160M program contains six seeds for each of eight methods. The Pythia-70M second site contains four seeds for four methods.

Site and method	Meaning	Complete / planned
P160 N1	group quota, cross loss, revival on	6/6
P160 N2	named-global, cross loss, revival on	0/6
P160 N3	group quota, cross loss, revival off	1/6
P160 N4	named-global, cross loss, revival off	0/6
P160 N5	group quota, no cross loss, revival off	6/6
P160 N6	named-global, no cross loss, revival off	6/6
P160 N7	conventional TopK	6/6
P160 N8	Matryoshka SAE	4/6
P70 N1/N2/N7/N8	four methods, four seeds each	16/16

No incomplete method enters a reported aggregate. Stage 1 uses P160 N5/N6. Stage 2 uses P160 N1/N7 and P70 N1/N2/N7. P70 N8 was excluded because the result-aware Stage-2 rule enumerated only the stable conventional baseline and N1/N2 second-site contrasts; the later completed N8 cells were outside that fixed package. The remaining jobs were paused for the paper deadline, not imputed.

A.2 Core training and analysis configuration

Natural data and estimands. The natural token cache uses `ArmelR/the-pile-splitted` revision `970296c12832ff2131ea13a04c35d8949cd7ee1e`, source configurations ArXiv, Github, Pile-CC, and PubMed Abstracts, and the Pythia tokenizer revision `582159a2dfe3e712a8d47ae83dec95ae3bde8e7e`. A seeded document-level partition forms 128-token sequences; the first position of each sequence is not scored. Training contains 150M valid token activations. Separate held-out targets contain 1,024 normalization documents, 512 learning-rate-tuning documents, 512 stopping-development documents (62,542 scored rows), and 2,048 in-domain test documents (248,596 rows). One document is one bootstrap block.

C2 uses UD English ESLSpok release 2.18 at revision `f4d0ebe5eae0bc179a988934e15f018dcacfe73b`: 436 held-out transcript files yield 1,144 inference units and 9,614 activation rows. The locally frozen analysis treated inference units as bootstrap blocks. Prompted by review, the reported C2 sensitivity instead joins the two transcript files used by each entity-substitution unit and aggregates all units over the resulting 408 transcript-dependency components before 20,000 paired bootstrap resamples. This correction was post hoc; its interval is similar to the frozen interval and does not turn the result into a prospective test. Thus "in-domain" means unseen documents from the four-component training corpus, whereas C2 means learner-English transcripts from a different source. The hash-based leakage audit compares training documents with every held-out cache: it found zero normalized-document intersections, zero complete token-sequence intersections, and zero shared contiguous token windows of length at least 16. It cannot rule out paraphrases or other semantic duplication.

Setting	Controlled 100M study	Natural priority study
Main model/site	Pythia-160M-deduped, block 3	same
Branch widths	8,192 / 32,768	same
Branch quotas	8 / 24	same
Total selected L_0	32	32
Learned weights	62,914,560	62,914,560
Learned biases	none	40,960 encoder + 768 decoder
Input coordinates	raw residual stream	fixed training mean / scalar RMS
Initialization	separate encoder and unit-row Gaussian decoder	unit-row Gaussian decoder copied to encoder
Learning rate	7×10^{-5}	tuned 3×10^{-4}
Cross-loss weight when enabled	0.01	0.01
Checkpoints	terminal 100M	50M, 75M, 100M, 125M, 150M
Primary held-out role	out-of-domain C2	in-domain test

The controlled cache contains 100,001,097 valid tokens in 822,400 sequences. Both non-Matryoshka phases use Adam with $(\beta_1, \beta_2) = (0, 0.999)$, $\epsilon = 10^{-8}$, 1,000 warmup steps, effective batch size 4,096, and unit decoder rows. Paired cells load identical saved tensors and token schedules. A row is called dead after more than 2,000 optimizer steps without selection. Revival, when enabled, replaces at most 64 rows per group (or per full conventional dictionary) with normalized current-batch examples and resets their optimizer state; the N5/N6 priority comparison disables it.

The natural learning rate and revival allowance were chosen on one 20M-token tuning seed from $\{3 \times 10^{-5}, 7 \times 10^{-5}, 1.5 \times 10^{-4}, 3 \times 10^{-4}\}$ and allowances $\{64, 256\}$ using development FVU, then frozen before confirmation. In-domain and stopping-development intervals use 2,000 paired resamples of fixed seeds and within-job document blocks; the post-hoc C2 sensitivity uses 20,000 resamples over transcript-dependency components. The plateau diagnostic fits FVU against the five checkpoints in units of 10M tokens; it passes only if the entire bootstrap interval for the stopping-development slope lies within $[-0.0002, 0.0002]$. Controlled objective-action results report three fixed pairs without a seed-population interval.

A.3 Historical revival-enabled selector comparison

Seed	MorphSAE FVU	Named-global FVU	MorphSAE minus named-global	Named-global small-group slots
42	0.187	0.278	-0.090	6.82
43	0.091	0.363	-0.273	5.18
44	0.097	0.138	-0.040	6.75
45	0.270	0.329	-0.059	9.77

MorphSAE selects exactly 8/24 slots in every row. The paired methods otherwise share initialization, schedule, width, total L_0, loss, and evaluation, but selection changes the subsequent per-group revival trajectory. Whole-group starvation is zero for both methods, overlap signs are mixed, and both fail the functional composite. The predeclared broader-preference gate therefore fails.

A.4 Natural priority estimates

The selector-only contrast is MorphSAE minus named-global; positive FVU favors named-global. In-domain and stopping-development intervals use the locally frozen inference-unit analysis. The C2 row uses the post-hoc transcript-dependency-component sensitivity analysis described above.

Split	Estimate	95% finite-design interval	Fixed-seed signs (+ / -)
Stopping development	0.000015	[-0.000143, 0.000201]	2 / 4
In-domain test	0.000023	[-0.000130, 0.000217]	2 / 4
C2	0.000387	[0.000025, 0.000758]	4 / 2

The six Pythia-160M conventional-baseline FVU values are 0.021070, 0.021180, 0.021169, 0.021270, 0.021267, and 0.021147. MorphSAE N1 minus conventional N7 is 0.000660 [0.000276, 0.001071]. At Pythia-70M, N1 minus N7 is 0.000952 [-0.003209, 0.006209], N2 minus N7 is 0.000637 [-0.002296, 0.002463], and N1 minus N2 is 0.000315 [-0.004573, 0.005615]. These second-site intervals cross zero.

B FUNCTIONAL EVALUATION

B.1 WHAT THE COMPOSITE TESTS

Representations and data roles. Probes operate on each decoded branch output $\hat{x}_j \in \mathbb{R}^{768}$, not on sparse codes; the raw block-3 residual vector is the normalization baseline. Each representation is standardized using its own discovery-set mean and coordinatewise scale. Multiclass ridge regression is fit to one-hot labels on UD English EWT (discovery), with $\alpha \in \{0.1, 1, 10, 100\}$ chosen by macro-F1 on UD English GUM (calibration; larger α breaks ties). C1 and C2 are disjoint UD English ESLSpok transcript-file groups. C1 checks that the raw signal is measurable; every pass/fail endpoint below is computed on C2. Discovery/calibration/C1/C2 contain 15,639/7,751/6,477/5,930 structural-task rows and 4,494/2,061/2,040/1,888 token-identity rows. C2 contains 436 transcript groups.

The structural targets are sentence-relative position quartile (four classes) and signed dependency-head distance (`ROOT`, four left-distance bins, and four right-distance bins). The lexical target is identity among the 24 fixed tokens listed in the prepared manifest. Strict absolute position is excluded because the model architecture makes it ineligible. For each target, the same frozen row population is used for raw and both branch representations.

Nominal assignment and controls. In the original cells, the branch named `pos` (width 8,192; TopK 8) is assigned the structural probes and document-context change; `content` (width 32,768; TopK 24) is assigned token identity and entity substitution. Equal-allocation cells keep that assignment but give both branches width 20,480 and TopK 16. Each relabel control is an exact tensor permutation of a regularized original state: the original `content` tensors become `pos` (width 32,768; TopK 24), the original `pos` tensors become `content` (width 8,192; TopK 8), and the same nominal assignment is then evaluated. The summed reconstruction and FVU are unchanged. These three relabel evaluations therefore add no trained states; they test whether the names alone determine the conclusion.

Structural/context specialization requires relative-position and dependency-distance localization plus a document-context counterfactual. Lexical specialization requires token-identity localization plus an entity-substitution counterfactual. Probes use macro-F1. For a C-class task and representation r, define

$$R(r) = \frac{F1(r) - 1/C}{F1(\text{raw}) - 1/C}, \qquad S = R(\text{assigned}) - R(\text{other}).$$

A localization endpoint passes only when the 2.5th percentile of 500 fixed transcript-group bootstraps has $R(\text{assigned}) \geq 0.65$ and $S > 0$. For counterfactual c, let $\rho_b(c)$ be branch b's mean cosine-distance response divided by the raw-activation response. With assigned branch a, other branch o, and opposite-family control c', the lower bounds must satisfy $\rho_a(c) - \rho_o(c) \geq 0.05$ and $\rho_a(c) - \rho_a(c') > 0$. A configuration passes only if every structural, context, lexical, and entity requirement passes.

Counterfactual construction. C2 provides 203 document-context pairs and 28 entity pairs. A context pair compares a later transcript sentence alone with the same sentence preceded by the closest earlier sentence from that transcript; responses average aligned first-subtoken branch vectors in the later sentence. An entity pair replaces the first eligible single-token proper noun with a proper noun from a different transcript while preserving the entire token layout; the response is measured at the replaced position. Transcript groups used for entity pairs are excluded from context pairs. For either construct, ρ_b is the cosine distance between the paired mean branch vectors divided by the corresponding raw-vector distance. This makes the test a branch-response comparison, not a claim that a probe coefficient is causal.

Every endpoint uses held-out C2 transcript groups. These thresholds are retrospective: C2 and predecessor outputs were visible, and the corrected v2.5 rules were then inherited without cell-specific tuning. The composite evaluates one nominal assignment rather than searching over label permutations.

Condition	Evaluations	Distinct trained states	Complete passes
Original regularized/unregularized cells	6	6	0
Equal-allocation controls	3	3	0
Exact branch relabelings	3	0 new	0
Total	**12**	**9**	**0**

Individual endpoints are heterogeneous. Across the three matched regularized-minus-unregularized pairs, changes in relative-position selectivity are +0.280, +0.050, and -0.020; head-distance changes are +0.029, -0.450, and +0.255; token-identity changes are +0.094, +0.541, and -0.331. These mixed signs motivate the conjunction rather than selective endpoint reporting.

B.2 Recurrence and intervention checks

Across regularized seed pairs, content-branch absolute-cosine nearest-line means are 0.738–0.757 and reciprocal-match rates are 0.321–0.333. Position means are 0.574–0.613 and reciprocal rates are 0.241–0.263. Isotropic Gaussian nearest-neighbor means are 0.149 and 0.137. Decoder directions therefore recur above this limited null, but not one-to-one; cross-branch controls can be similar.

Subtracting either supporting long-run branch increases first-subtoken NLL. Position subtraction gives mean ΔNLL 1.833–2.223 and content subtraction 3.014–5.308. These off-manifold interventions show that the additive outputs affect the model, not that a branch owns a natural behavior.

C Trained-copy positive control

Synthetic task and panels. The trained-copy study uses three fixed synthetic-transformer checkpoints and four SAE seeds. Each transformer receives eight source values from a 32-symbol vocabulary plus a query in $\{0, \ldots, 7\}$ and is trained to output the value at the queried slot. It has learned position, query, and value embeddings; a single factorized attention head of width 64; layer normalization; and a 32-way linear readout. Each subject was trained for 4,000 steps on a separate balanced 8,192-row panel and retained only after clean and corrupt accuracy exceeded 0.95 and routing controls passed.

The SAE panels are disjoint from subject training and from one another. Each row fixes a query q and target y, samples the other seven values, and creates three inputs: clean has y at slot q, corrupt replaces it by a distinct contrast y', and sham uses a third value. Fit has 2,048 rows (64 blocks of 32); development and confirmation each have 256 rows (32 blocks of 8) and enumerate every (q, y) pair once. Row IDs, seeds, blocks, and $(q, y, y', \mathrm{sham})$ tuples are disjoint across panels. Each learned SAE has width 256 and total TopK 32.

Training objectives. Let x^+ and x^- be the clean and corrupt pre-layer-normalization head states, $\delta = x^+ - x^-$, and $\hat{x}_c^{\pm}, \hat{x}_b^{\pm}$ the decoded controller and background branches. All methods minimize normalized reconstruction MSE on both states. MorphSAE and named-global add

$$\mathcal{L}_{\text{named}} = \frac{\|(\hat{x}_c^+ - \hat{x}_c^-) - \delta\|_2^2}{\text{Var}(\delta)} + \frac{\|\hat{x}_b^+ - \hat{x}_b^-\|_2^2}{\text{Var}(x^+, x^-)},$$

where $\text{Var}(x^+, x^-)$ is the scalar variance pooled over both state sets. This loss assigns the paired causal change to the controller branch and asks the background branch to remain invariant. The conventional SAE instead adds $\|(\hat{x}^+ - \hat{x}^-) - \delta\|_2^2 / \text{Var}(\delta)$ to its reconstruction loss. Both coefficients are 1. Models use Adam for 3,000 steps at learning rate 10^{-3}. These paired-delta targets are the synthetic identifiability signal; there are no analogous labels in the natural experiments.

Frozen intervention. For named methods, the predicted change is $\hat{x}_c^+ - \hat{x}_c^-$. The conventional dictionary has no controller branch, so development data select exactly 128 atoms by descending mean absolute clean–corrupt coefficient change, breaking ties by lower index; confirmation uses only those atoms. We add the predicted change to the corrupt head state and recompute logits. Atom selection never uses confirmation rows.

Let $m(\ell, y)$ be the target logit minus the corrupted alternative logit. **Target recovery** is

$$\frac{m(\ell_{\text{patch}}, y) - m(\ell_{\text{corrupt}}, y)}{m(\ell_{\text{clean}}, y) - m(\ell_{\text{corrupt}}, y)}.$$

Full-vocabulary recovery is one minus centered-logit RMSE from clean, normalized by the clean-corrupt RMSE. **Collateral error** is the same normalized RMSE after excluding the target and corrupted-alternative logits. Recovery 1 and collateral error 0 are ideal.

MorphSAE and named-global use identical controller/background supervision; the conventional contrast therefore changes named structure, its supervision target, and the intervention rule. The primary quota contrast also has unequal realized intervention size: mean L_0 is 29.78 for MorphSAE and 43.19 for named-global. Intervals use 5,000 crossed resamples of the three subjects, four SAE seeds, paired confirmation blocks, and within-block rows; they are finite-design sensitivity intervals.

Contrast	Target recovery	Full-vocabulary recovery	Collateral error
MorphSAE minus named-global	0.0000003 [-0.0000037, 0.0000041]	0.000106 [0.000045, 0.000183]	-0.000123 [-0.000208, -0.000056]
Named/structured supervised minus conventional	0.3966 [0.3648, 0.4290]	0.4950 [0.4702, 0.5207]	-0.3612 [-0.3760, -0.3464]

The quota-specific simultaneous-benefit gate and 10/12 sign gate fail. Exact-oracle, shuffled-control, collateral-safety, and per-subject reliability gates pass. Realized intervention L_0 means are 29.78 for MorphSAE, 43.19 for named-global, 37.15 for the selected conventional intervention, and 59.61 for the unrestricted conventional diagnostic. These differences further caution against interpreting either between-method contrast as a pure selector effect.

D SUPPORTING DEVELOPMENT AND TRANSFER EVIDENCE

Three selected 1B-token regularized checkpoints explain more than 92% of tail-training variance and have sampled overlap 0.00281–0.00464. A matched unregularized checkpoint has overlap 0.01520, but the selected candidate universe and stream cursor were not preserved; this is supporting development evidence.

Untuned transfer performs poorly: FVU is 1.844–3.664 at Pythia-160M layer 1, 9.804–22.438 at layer 7, 0.424–0.724 at Pythia-70M layer 3, and 1.790–3.734 at Pythia-410M layer 3. These cells change hidden dimension and were not retuned, so they are condition-sensitivity results rather than a scaling law.