Dimensional Collapse in Transformer Attention Outputs

A Challenge for Sparse Dictionary Learning

Paper: Junxuan Wang, Xuyang Ge, Wentao Shu,
Zhengfu He, and Xipeng Qiu

Reading seminar | arXiv:2508.16929v4 | 11 November 2026

Data geometry can make SAE capacity go unused

Central idea

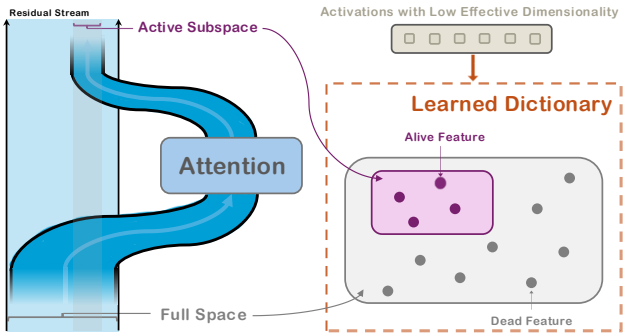
Initialize sparse features in the data's active subspace, then train normally.

- Attention outputs often have a concentrated singular spectrum.
- ASI improves reconstruction and reduces dead features.
- At 1M features, the 87% $\rightarrow$ <1% result uses ASI + SparseAdam.

Published at ICML 2026, PMLR 306. Official code: Llamascopium.
Read version: arXiv v4, 11 February 2026; links checked 11 November 2026.

Paper, Abstract and Secs. 4–5, PDF pp. 1, 4–8.

Why should random features fill the entire space?



Paper, Fig. 1 and Sec. 2.3, PDF pp. 2–3.

- Larger dictionaries can waste memory and training on inactive features.
- A dead feature never activates over 10M evaluation tokens.
- The proposed intervention changes the starting geometry.

The target is the output written to the residual stream

$$Z_i = \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i, \quad Z = [Z_1, \ldots, Z_H]$$

$$\boxed{O = ZW^O = \sum_{i=1}^{H} Z_i W_i^O}$$

- $Z_i \in \mathbb{R}^{n \times d_h}$; $Z, O \in \mathbb{R}^{n \times d}$; $W^O \in \mathbb{R}^{d \times d}$.
- O is the attention block's total write into the residual stream.
- Keep attention patterns, concatenated heads Z, and outputs O distinct.

Paper, Sec. 3.1, PDF p. 3. Standard MHSA notation: $Hd_h = d$.

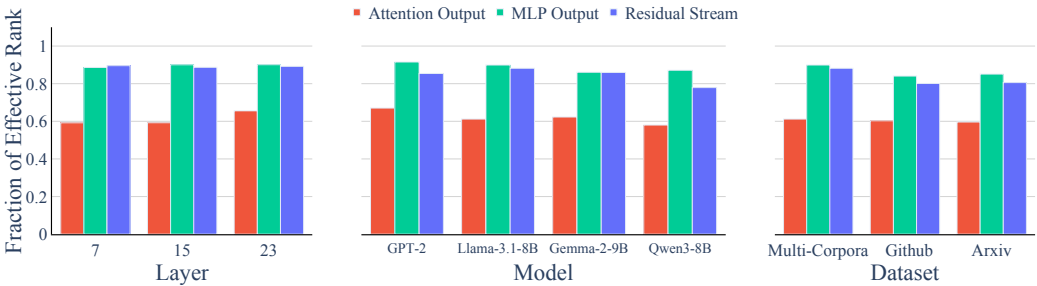
Effective rank summarizes how concentrated a spectrum is

$$\widetilde{A} = U\Sigma V^\top, \qquad p_j = \frac{\sigma_j}{\sum_k \sigma_k}, \qquad \operatorname{erank}(\widetilde{A}) = \exp\left(-\sum_j p_j \log p_j\right)$$

- $\widetilde{A}$: mean-centered token activations, shape $n \times d$.
- Uses singular values, rather than squared singular values.
- Equal singular values give the full rank; concentrated spectra lower it.
- erank $/ d \approx 0.6$ describes spectral concentration; it does not imply that the remaining directions are exactly zero.

Paper, Definition 4.1, Sec. 4.1, PDF p. 4.

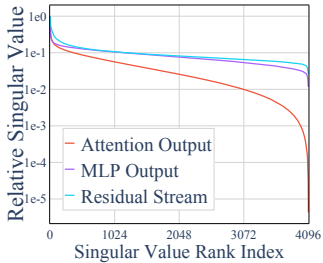
Attention outputs typically use fewer effective directions

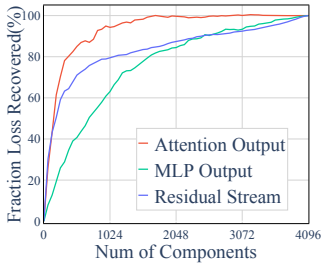


- Left: layers. Middle: models. Right: data domains.
- Main-text summary: attention $\sim 60\%$, MLP/residual $\sim 90\%$ effective rank.
- These are empirical trends; ratios vary by model, data, and layer.

Paper, Fig. 2, PDF p. 4; counterexamples to a strict ordering appear in Fig. 10, PDF p. 15.

Task-relevant attention contributions are also concentrated



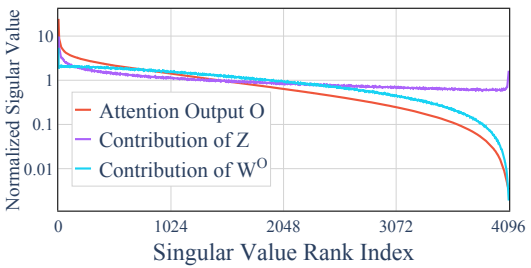


$$\mathrm{FLR}(m) = \frac{L_{\text{zero}} - L_{\text{recovered}}(m)}{L_{\text{zero}} - L_{\text{original}}}$$

For $>$ 99% recovery: attention needs 39.1% of dimensions, MLP 95.3%, residual 96.9%. Recovery is relative to the zero-ablation loss gap.

Paper, Fig. 3, Sec. 4.3, PDF p. 5; App. B, PDF p. 14. Llama-3.1-8B, layer 15, SlimPajama.

The output projection further suppresses the spectral tail



For a unit output direction $\hat{e}$, let

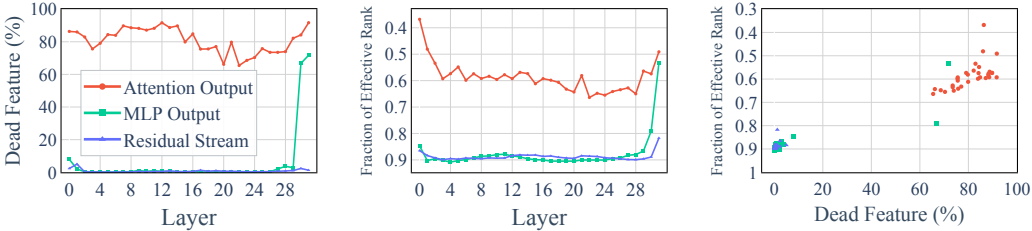
$$v = W^O \hat{e}, \qquad \hat{v} = v/\|v\|_2.$$

$$\mathrm{Var}(O\hat{e}) = \underbrace{\mathrm{Var}(Z\hat{v})}_{Z}\,\underbrace{\|v\|_2^2}_{W^O}.$$

Tail suppression tracks W^O strongly. Overlapping head write subspaces provide an intuitive explanation.

Paper, Sec. 4.4 and Fig. 4, PDF p. 5. Variance factors correspond to squared spectral scales; assume $v \neq 0$.

Lower effective rank co-occurs with more dead features



- LlamaScope: same dictionary size 32,768 and $L_0 = 50$ across sites.
- Attention outputs and late MLP outputs show the strongest co-occurrence.
- Capacity and sparsity sweeps preserve the broad trend.
- Sec. 6 leaves the precise causal dynamics unresolved.

Paper, Fig. 5, PDF p. 6; App. F, Tables 1–3, PDF pp. 18–19; Sec. 6, PDF p. 8.

ASI folds a data-dependent projection into initial weights

$$V_{\text{active}} = V_{:,:d_{\text{init}}}, \qquad P = V_{\text{active}} V_{\text{active}}^{\top}$$

$$\boxed{W_D \leftarrow W_D P, \qquad W_E \leftarrow W_D^{\top}}$$

1. Center representative activations and compute their SVD.
2. Select the leading d_{init} right singular directions.
3. Project the random dictionary; tie the encoder initialization.
4. Train the full SAE normally, without repeated projection.

Row-vector convention: $W_D \in \mathbb{R}^{h \times d}$, $W_E \in \mathbb{R}^{d \times h}$; h is dictionary size.

Paper, Sec. 5.2, PDF p. 6; App. K, PDF p. 22. The displayed matrix update governs the whole dictionary.

Initialization aligns features without fixing a permanent bottleneck

$$z = \mathrm{TopK}(xW_E + b_E), \qquad \hat{x} = zW_D + b_D$$

$$\mathcal{L} = \|x - \hat{x}\|_2^2 + \alpha \mathcal{L}_{\mathrm{aux}}$$

- Ordinary TopK uses only the reconstruction term; AuxK adds an auxiliary loss.
- ASI changes the starting directions; weights may subsequently leave the subspace.
- $d_{\mathrm{init}} = d$ recovers ordinary initialization.
- Main setting: $d = 4096$, $d_{\mathrm{init}} = 768$. The latter is an optimization choice, not the effective rank.

Paper, Secs. 3.2, 5.2 and App. E, PDF pp. 3, 6, 16–18.

The main comparison holds capacity and training data fixed

Target	Llama-3.1-8B, attention output, layer 15 (zero-based)
Data / training budget	SlimPajama / 800M tokens per SAE
Dictionary / sparsity	32,768 features / $k = 50$, unless swept
Batch / context	4096 / 1024 tokens
Optimizer	Adam; SparseAdam added in scaling experiments
ASI / AuxK	$d_{\text{init}} = 768$ / $\alpha = 1/32$, $k_{\text{aux}} = 512$

- Evaluate dead features, Normalized MSE, and Δ LM Loss; all lower is better.
- Three seeds for sparsity curves; a separate 15-seed significance test.

Paper, App. E–H, PDF pp. 16–20. Delta LM Loss: evaluated over 1M tokens.

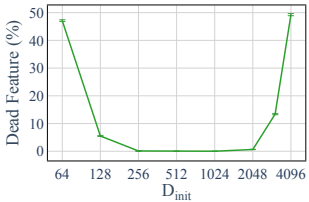
At 32K features, ASI improves utilization and reconstruction

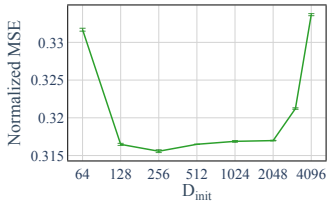
Method	Dead count ↓	NMSE ↓	Δ LM Loss ↓ ($\times 10^{-3}$)
TopK	16144.33	0.33367	8.375
AuxK	54.67	0.32241	7.958
ASI	4.00	0.31680	7.847

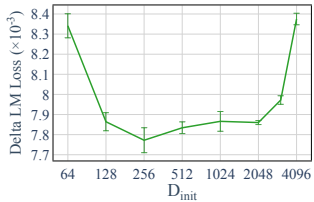
- $k = 50$; means over three seeds. ASI dead-count std: 1.73.
- NMSE improves by 5.06% relative to TopK, 1.74% relative to AuxK.
- The LM-loss increase falls by 0.000528 vs. TopK; this is a much smaller absolute effect.

Paper, Table 4, PDF p. 20; relative percentages calculated from reported means.

Moderately restricting the initial subspace works best



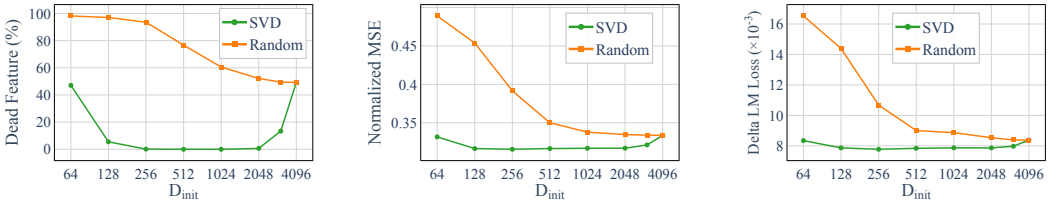




- Horizontal axis: initial subspace dimension; all vertical metrics lower is better.
- Full-space initialization leaves many features dead.
- Very small subspaces damage reconstruction; 256–2048 is broadly usable.
- Experiments use 768 by default, rather than a fixed fraction of effective rank.

Paper, Fig. 6, PDF p. 6; App. J.3, PDF p. 22. Curves show three-seed mean and standard deviation.

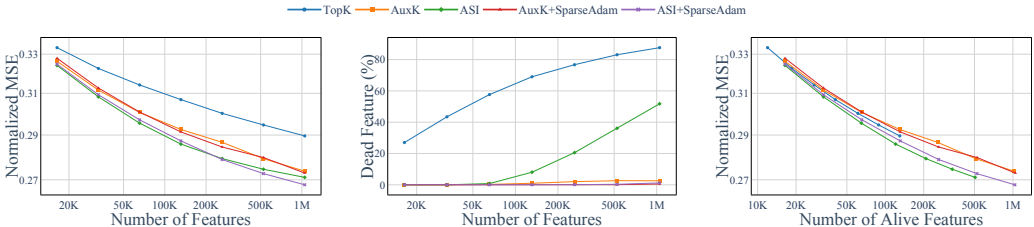
The data-supported directions matter, not dimensionality alone



- Green: leading singular directions. Orange: random subspace.
- Random restrictions worsen dead features, reconstruction, and LM loss.
- Near-full-rank residual streams gain little from ASI.

Paper, App. J, Fig. 12, PDF p. 21; residual-stream control: Fig. 13, PDF p. 22.

At large capacity, alive-feature counts do not tell the full story



(a) Loss vs. #Total Features. *(b)* #Dead Features vs. #Total Features *(c)* Loss vs. #Alive Features.

- Left: ASI improves reconstruction across 16K–1M total features.
- Middle: ASI alone retains many dead features at 1M; AuxK keeps more alive.
- Right: ASI uses alive features more efficiently than AuxK.

Paper, Sec. 5.3 and Fig. 8, PDF pp. 7–8. Shared legend retained from the original figure.

SparseAdam complements geometry at the million-feature scale

The headline result has two ingredients

Standard TopK: $\sim 87\%$ dead at 1M features.
ASI + SparseAdam: $< 1\%$ dead, with the lowest reconstruction error.

- ASI addresses the initial mismatch with data directions.
- SparseAdam masks updates of parameters and moments to nonzero-gradient positions.
- The paper links this benefit to stale-momentum effects in sparse training.
- Optimizer learning rates were tuned separately: Adam 4×10^{-5}, SparseAdam 6×10^{-5}.

Paper, Sec. 5.3, Fig. 8, App. E and M, PDF pp. 7–8, 16, 24–25.

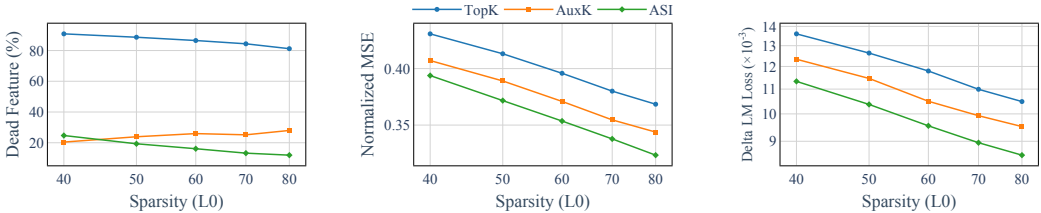
Reconstruction gains transfer more consistently than minimum death

Target	Dead: AuxK / ASI	NMSE: AuxK	NMSE: ASI
Llama / layer 7	98 / 5308	0.29800	**0.28882**
Llama / layer 23	332 / 27	0.20235	**0.20141**
Qwen / layer 8	4 / 7	0.28849	**0.28533**
Qwen / layer 26	66 / 566	0.28127	**0.28087**

- Llama-3.1-8B: SlimPajama; Qwen3-8B: fineweb-edu.
- ASI has lower NMSE and LM-loss degradation in all four configurations.
- AuxK has fewer dead features in three of the four configurations.

Paper, App. I, Tables 6–7, PDF p. 21. All targets are attention outputs.

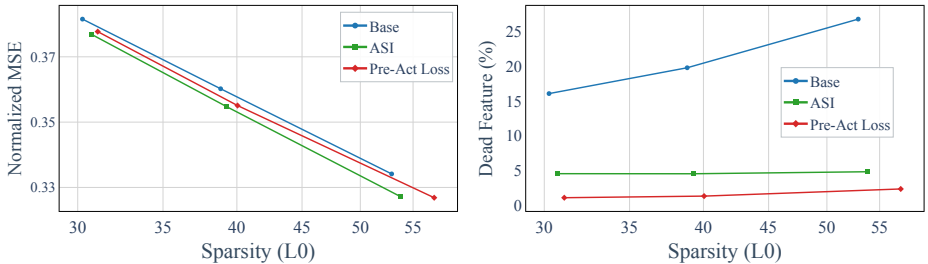
Lorsa benefits from head-aware subspace initialization



- Group sparse Lorsa units by their original MHSA head.
- Initialize input/output directions in the corresponding active subspaces; copy original Q/K parameters.
- With 32,768 units, ASI improves reconstruction and dead-unit rate.
- This evaluates the combined Lorsa initialization design.

Paper, Sec. 5.4, Fig. 9, PDF p. 8; implementation details in App. K/O, PDF pp. 23, 26–27.

JumpReLU benefits when training starts with low activity

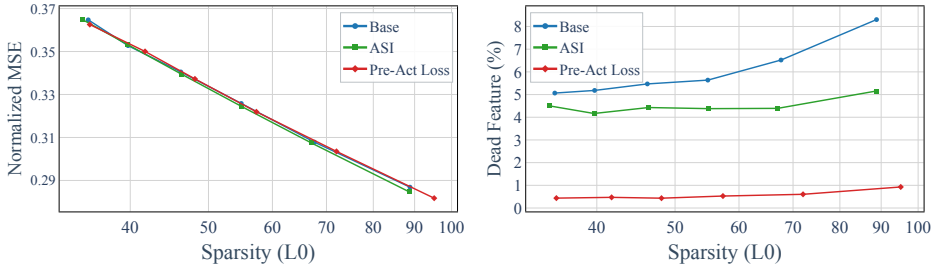


Blue: base; green: ASI; red: the paper's Pre-Act Loss comparator.

- Left: normalized reconstruction error. Right: dead-feature fraction.
- ASI improves both metrics when L_0 starts low and stays low.
- Larger L_0 means more active features per token.

Paper, App. E.4/L, Fig. 14, PDF pp. 18, 24. JumpReLU, Llama-3.1-8B middle-layer attention output.

A dense training start can erase the initialization advantage



Blue: base; green: ASI; red: the paper's Pre-Act Loss comparator.

- High initial L_0, then reduced: ASI and base reconstruction nearly coincide.
- High-K annealing in TopK also removes the observed advantage.
- Authors' explanation: early dense updates move features out of the subspace.

Paper, App. L, Figs. 15–16, PDF pp. 24–25. High-K TopK is a separate control.

Feature quality looks similar in a small blinded evaluation

Semantic consistency

100 sampled features total; 50 per method. Scores range from 1 to 5.

Method	Score 5	Mean*
ASI	17/50	3.54
TopK	18/50	3.52

Low-activity directions

Reconstruction error projected onto the paper's dead subspace:

TopK	0.00350
ASI	0.00334

Similar errors support retained coverage in this comparison.

More alive features do not establish better monosemanticity.

Paper, App. N, Tables 8–10, PDF pp. 25–26. *Means calculated from score counts; authors report no significant difference.

Geometry-aware initialization is useful when the geometry persists

1. **Measure the target site.** Attention outputs often concentrate strongly; residual streams may not.
2. **Choose directions from data.** ASI improves the starting dictionary, with a one-time SVD/projection cost.
3. **Track three outcomes.** Feature survival, reconstruction, and downstream LM loss need separate checks.
4. **Respect the training regime.** High initial L_0 can erase gains; very large dictionaries benefit from SparseAdam.

Open question: How do low-rank geometry and feature competition jointly produce dead features?

Paper, Sec. 6, PDF p. 8; App. J–N. Final interpretation based on the reported controls.

References and implementation entry points

- Wang, Ge, Shu, He, and Qiu.
  Dimensional Collapse in Transformer Attention Outputs: A Challenge for Sparse Dictionary Learning.
  arXiv:2508.16929v4, 11 February 2026.
- ICML 2026 proceedings, PMLR 306:126304–126331.
- OpenMOSS/Llamascopium: official public implementation.
  Initializer: active-subspace switch and dimension configuration.

Figures are reused from the authors' arXiv v4 source or cropped from its PDF. Calculated comparisons are labeled explicitly.

Publication and code checked 11 November 2026. This deck is a reading seminar, not an author presentation.

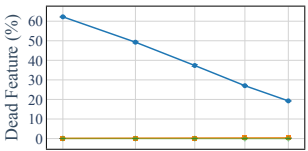
Backup: ASI in row-vector pseudocode

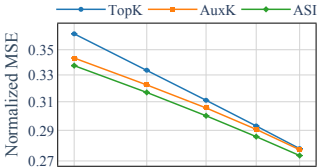
$$
\begin{aligned}
\widetilde{X} &= X - \mathrm{mean}_{\mathrm{tokens}}(X), \\
U, \Sigma, V^{\top} &= \mathrm{SVD}(\widetilde{X}), \\
B &= V_{:,:d_{\mathrm{init}}}, \\
W_D &\leftarrow (W_D B) B^{\top}, \\
W_E &\leftarrow W_D^{\top}.
\end{aligned}
$$

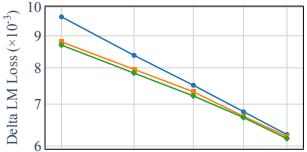
- X: $n \times d$; B: $d \times d_{\mathrm{init}}$; W_D: $h \times d$.
- The factored update avoids materializing the $d \times d$ projection.
- Subsequent SGD/Adam training is unrestricted.

Paper, Sec. 5.2/App. K, PDF pp. 6, 22. Variable-name and prose inconsistencies resolved using the matrix update.

Backup: gains persist across the tested sparsity frontier



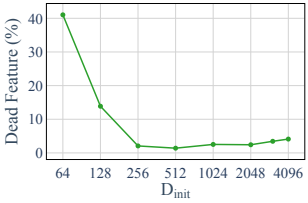


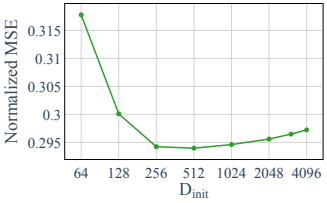


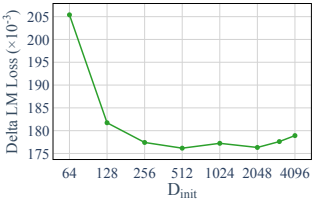
- Fixed 32,768-feature dictionary, $L_0 \in \{40, 50, 60, 70, 80\}$.
- ASI has the best mean values for all three metrics in Table 4.
- Increasing L_0 reduces base TopK death and narrows reconstruction gaps.
- Separate 15-seed test: ASI vs. AuxK $p = 1.27 \times 10^{-6}$ for Δ LM Loss.

Paper, Fig. 7, PDF p. 7; Tables 4–5, PDF p. 20. Test: one-tailed Welch, $k = 50$.

Backup: a near-full-rank site offers little room for ASI







- Target: post-layer-15 residual stream of Llama-3.1-8B.
- ASI provides minimal gains relative to ordinary initialization.
- Over-restricting directions damages reconstruction; assess the spectrum first.

Paper, App. J.2 and Fig. 13, PDF p. 22. These curves describe a different activation site from the main experiment.

Backup: evaluation conventions determine the interpretation

- **Spectra:** 10M activations per site; mean-centered; exclude extreme activation norms. Five samples give effective rank 2523.165 ± 0.404 in App. C.
- **Dead features:** no activation in 10M evaluation tokens.
- **NMSE:** retain the paper's reported metric; v4 does not spell out its normalization denominator.
- **Δ LM Loss:** positive degradation, lower is better in tables; interpreted here as $L_{\mathrm{SAE}} - L_{\mathrm{original}}$.
- **Optimizer comparison:** learning rates swept separately; one-time SVD is additional preprocessing.

Paper, Secs. 2.3, 4.2, 5.2; App. C/E, PDF pp. 3–6, 14, 16–18.