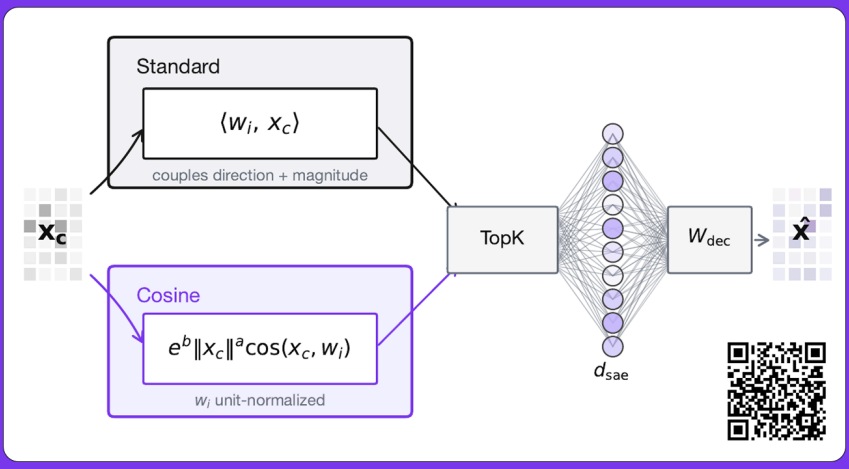
Spotlight

SAEs waste 74% of their features, cosine scoring finds the missing ones



Size Doesn't Matter: Cosine-Scored Sparse Autoencoders

Silen Naihin (Experiential Labs), Lev Stambler (Tear Labs)

arXiv: 2606.15054

+21.9% relative sparse probing accuracy

requires 10× less training data

Motivation

a) SAEs detect features via inner product, so activations scale with both directional alignment and input norm.

b) But layer normalization discards magnitude, so the encoder detects a quantity the model does not read.

c) Sparsity makes this worse: norm-inflated features win TopK slots, starving content features.

Solution + Results

We replace inner product scoring with a learned cosine magnitude blend, letting each feature decide how much norm to use.

Standard: $s_i(x) = \langle x_c, w_i \rangle + b_{\text{enc},i},$

Global: one shared exponent *a* for all features.

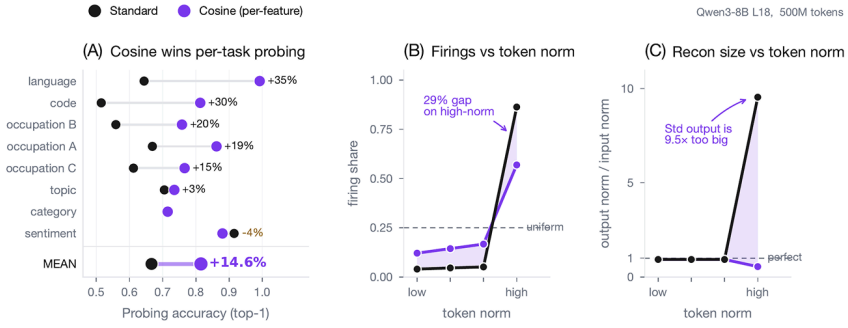
$$s_i(x) = \underbrace{\exp\big(a \log \|x_c\| + b\big)}_{e^b \, \|x_c\|^a} \cdot \cos(x_c, w_i) \; + \; b_{\text{enc},i},$$

Per-feature: $a_i = a_{base} + \delta_i$ *which* lets individual features deviate stably.

$$s_i(x) = e^{b_i} \, \|x_c\|^{\,a_{\text{base}}+\delta_i} \, \cos(x_c, w_i) + b_{\text{enc},i}.$$

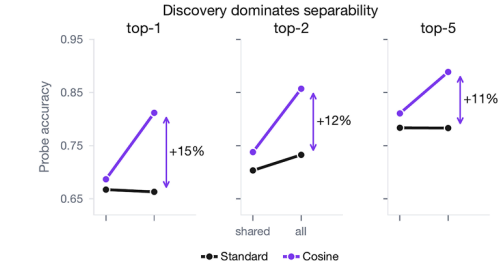
	Standard	Global a	Per-feature
FVE	0.770	0.769	**0.771**
Probing top-1	0.667 $\pm$0.003	0.808 $\pm$0.006	**0.813** $\pm$0.013
Q4 FVE	−184	**+0.33**	**+0.25**
Dead %	0.0	0.0	0.0
Learned a	—	0.258	mean 0.076

Interpretabilty

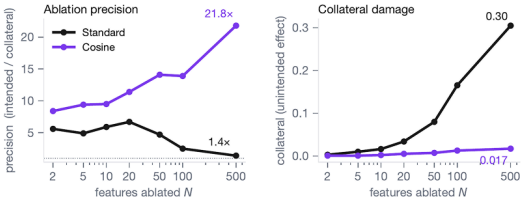


- Cosine improves 7/8 tasks at matched FVE.
- Standard unmatched features fire 22× more on the highest-norm quartile than the lowest, versus 4.7× for cosine.
- On high-norm tokens, standard reconstructs at 9.5× the input norm, while cosine stays near input scale.

Cosine discovers features standard SAEs miss
Restricting both to shared features nearly closes the gap. 87% of the advantage comes in unique features.



Cosine features intervene more cleanly
Standard features become increasingly entangled as ablated while cosine features preserve intended effect.



Discussion

- Geometry is the lever: loss reweighting and selector ablations do not close the gap.
- FVE is insufficient: matched reconstruction can hide norm-conditioned dictionaries.
- Learned cosine scoring should be the default for normalized residual streams.

Limitations

- Not universal. Weaker on deep LayerNorm layers and sentiment.
- Cosine fixes selection, but reconstruction still depends on restored activation scale.
- JumpReLU/gated/Matryoshka/crosscoders remain future work.

Website with interactive cosine SAE comparisons including **links to paper, GitHub, and checkpoints:**

Spotlight at the ICML Mechanistic Interpretability Workshop 2026

silen@experientiallabs.ai