

NeurIPS 2026

Rethinking SAE Evaluation: An Atomic Interpretable Unit Framework Reveals Hidden Polysemanticity and Redundancy

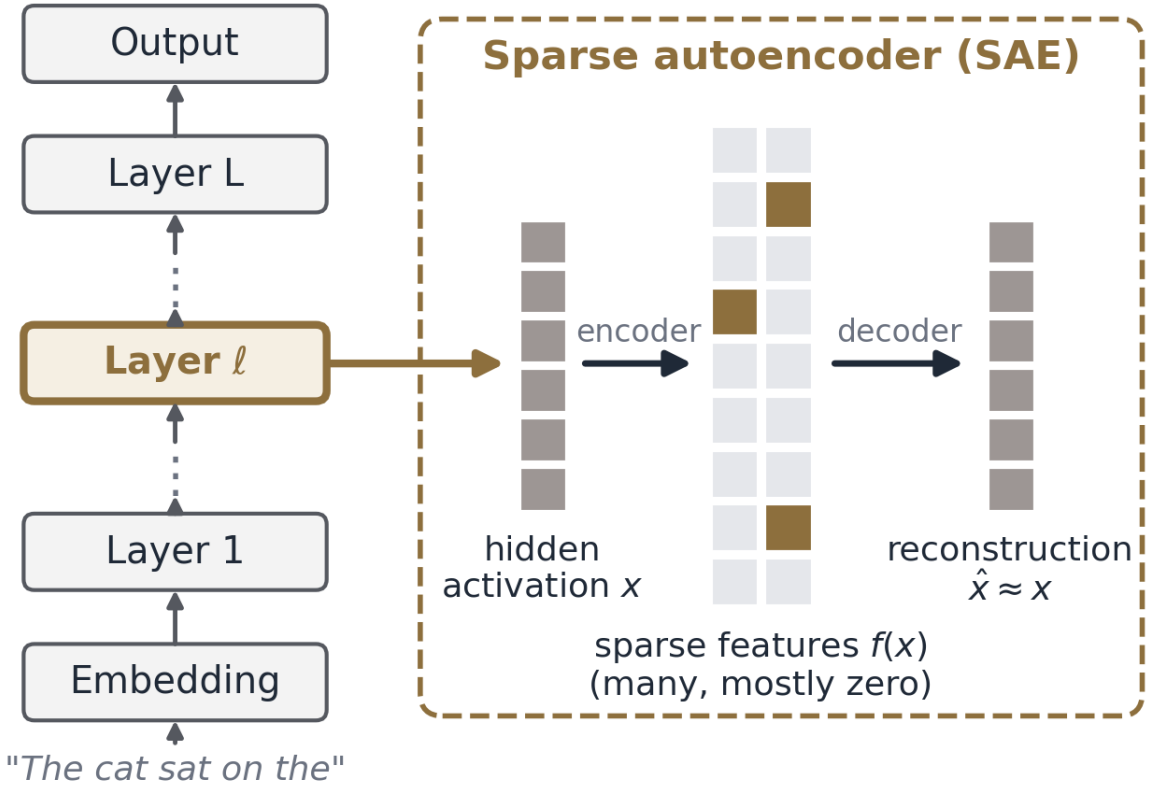
Riteng Zhang · Xiaoqian Wang



Background: sparse autoencoders (SAEs)

Decompose a model's hidden activations into many sparse features

Transformer LM



Encoder $f(x) = \mathrm{ReLU}(W_{\mathrm{enc}}\, x + b_{\mathrm{enc}})$

Decoder $\hat{x} = W_{\mathrm{dec}}\, f(x) + b_{\mathrm{dec}}$

Training objective

$$\min \ \|x - \hat{x}\|_2^2 \ + \ \lambda \, \|f(x)\|_1$$

reconstruction sparsity

Hope: each feature = one concept (monosemantic)

Global vs. local metrics

Global metrics are measured on the whole SAE; local metrics on one feature or a pair

Global metrics

measured directly on the whole SAE

- Reconstruction fidelity
- Sparsity (mean L0)
- Dead features (firing frequency)

✓ **one direct dictionary-level number**

Local metrics

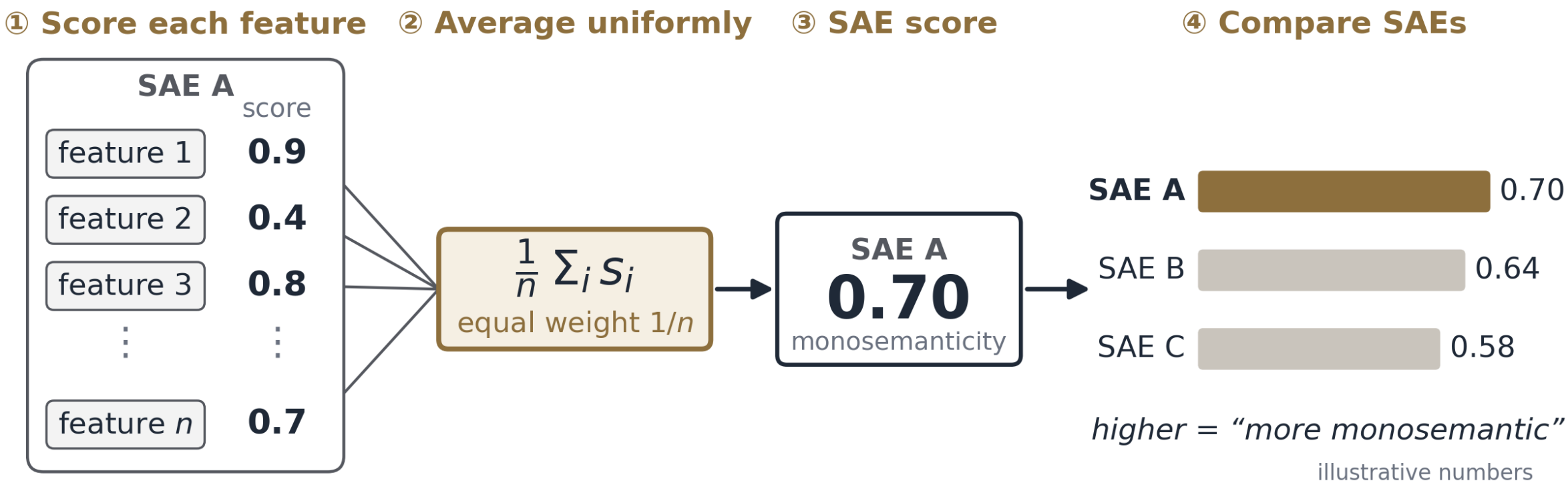
defined per feature or per feature pair

- Monosemanticity: one feature
- Redundancy: a pair of features

→ **need a rule to aggregate them: which features, what weights?**

From local scores to an SAE score

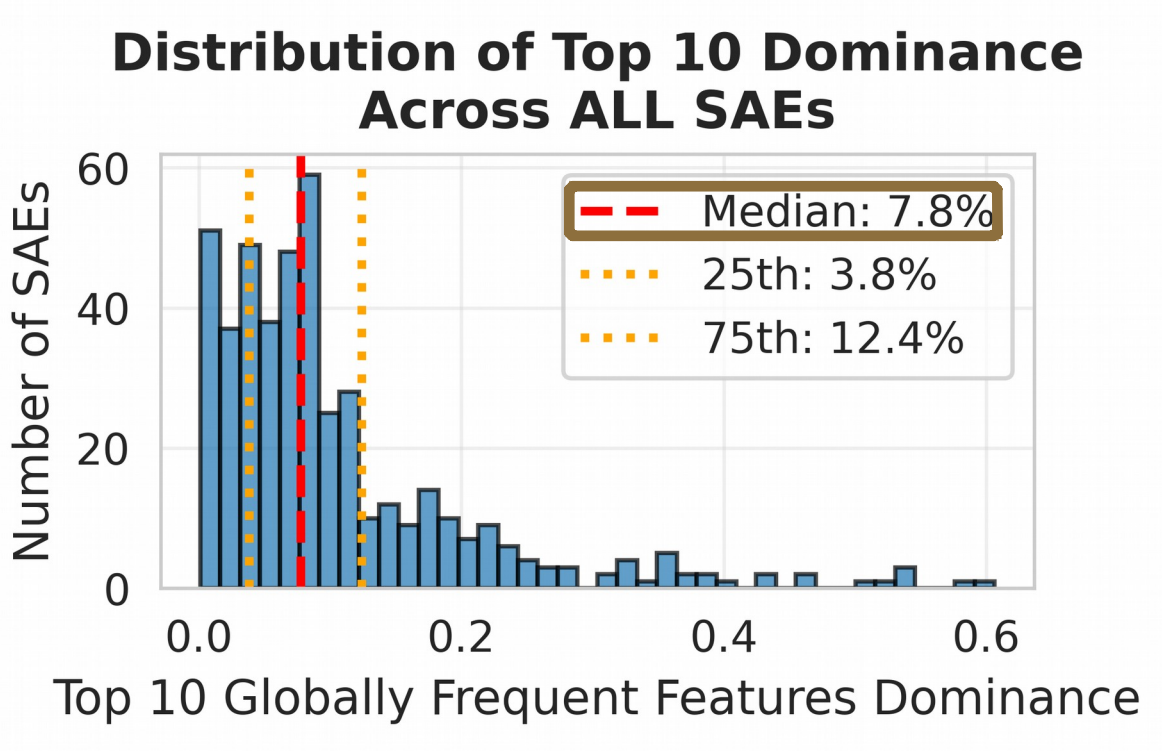
Common practice: score each feature, average uniformly, then compare SAEs



Every feature counts equally. But is every feature equally important to the model?

Feature dominance

A few features carry much of each input's activation; uniform averaging hides them



Dominance = a feature's share of the total activation on an input, averaged over all inputs in a dataset

Top-10 most frequent features: **median 7.8%** of each input's activation (SAEs with 4k–128k features)

Dominant features carry a disproportionate share of attribution → **their quality should count more**

We split features into four quartiles by mean dominance: **Q1 = most dominant 25%**, Q4 = least

Finding 1: hidden polysemanticity

Dominant features are less monosemantic · 600+ SAEs, 7 SAE families, 28,000+ features

65.2%

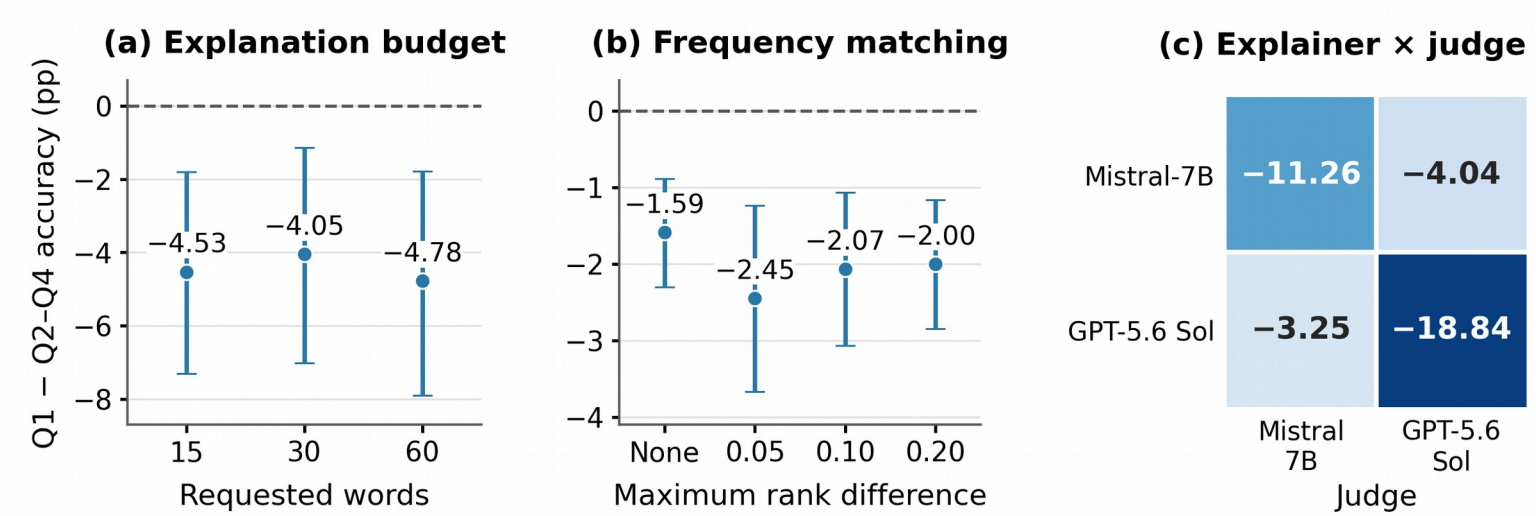
of SAEs: Q1 scores below Q2–Q4

$p = 6.76 \times 10^{-8}$

per-SAE Wilcoxon signed-rank test

Synthetic SAEs with known concepts: same deficit, no LLM judge

Robust to explanation length, frequency matching, and explainer × judge

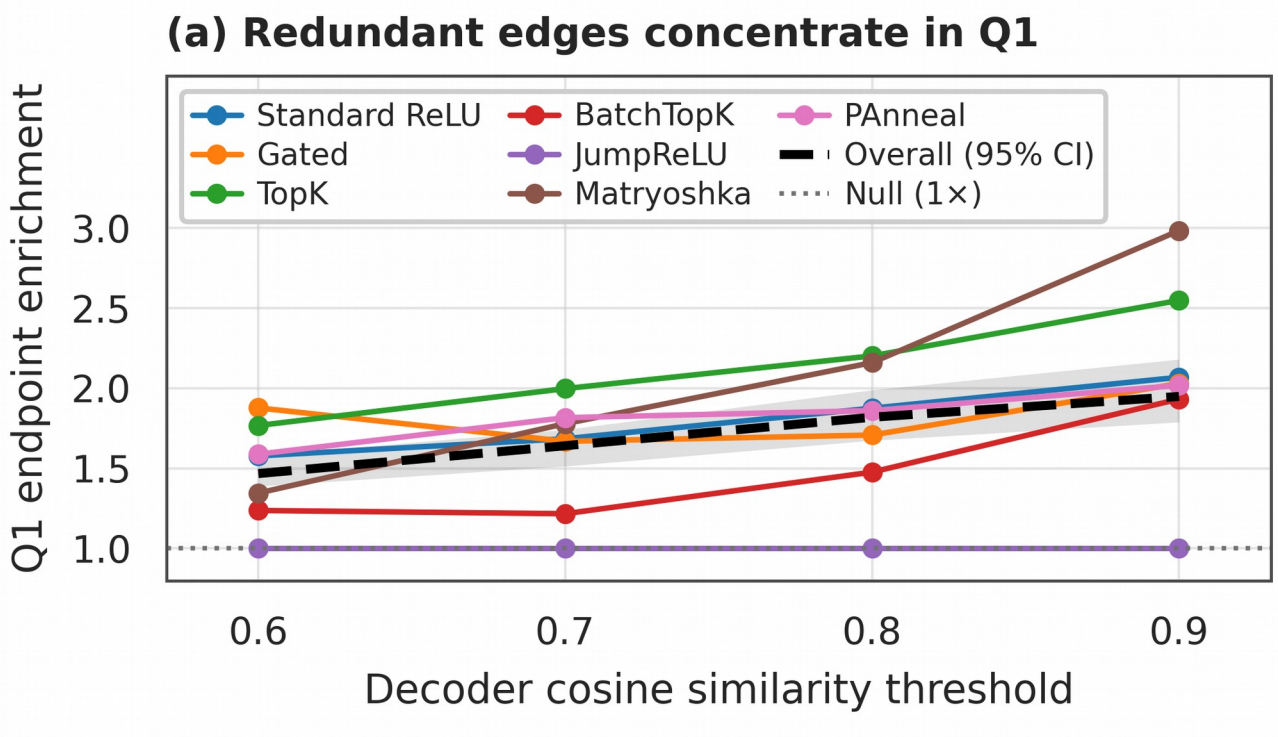


Q1 minus Q2–Q4 accuracy (percentage points); below 0 = Q1 worse

Measured with our DAME: an LLM judge ranks a feature's low / medium / high examples using only its explanation

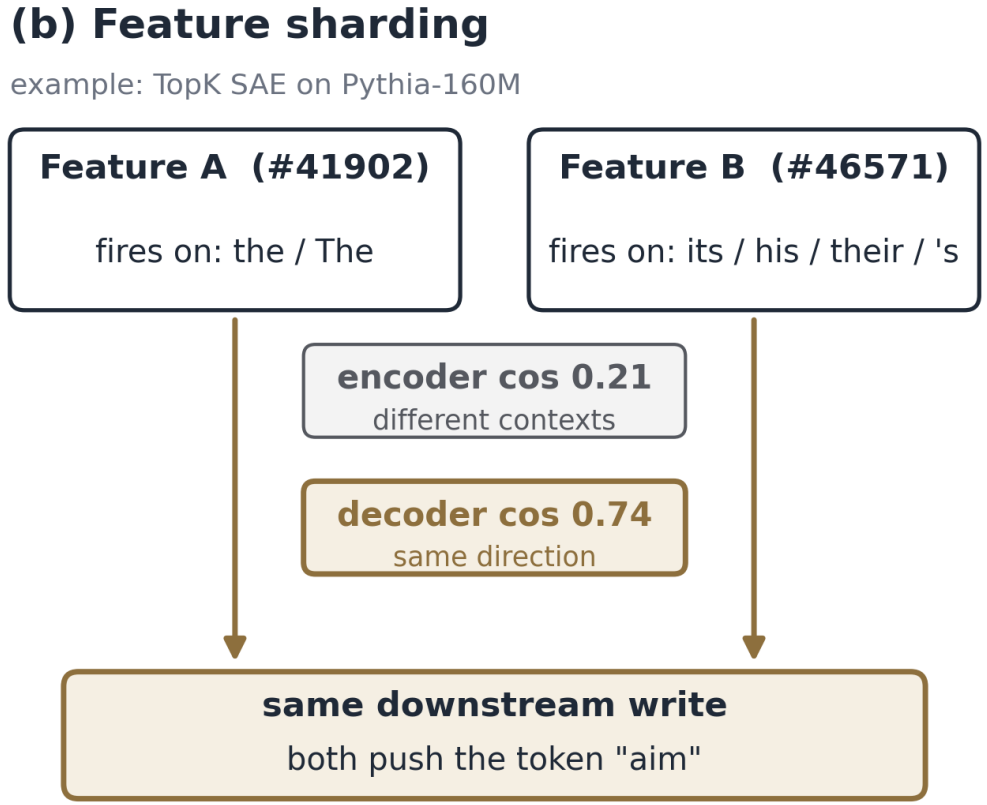
Finding 2: hidden redundancy

Dominant features are more redundant · feature sharding: same decoder direction, different contexts



Median Q1 enrichment 1.64× at decoder cos 0.7; holds for 0.6–0.9

Q1 redundant pairs have much less similar encoders



Standard metrics look healthy, but the effective dictionary is smaller.

Recommendations

Averaged scores hide these failures

For practitioners

Before attribution, circuit tracing or intervention, run dominance-stratified checks:

① monosemanticity across dominance quartiles

② decoder similarity among dominant features

For SAE researchers

Polysemantic and redundant dominant features are targets for better training objectives and architectures.

Thank you! Contact: **zhan5982@purdue.edu**

Paper: neurips.cc/virtual/2026/poster/139009 · arXiv 2610.30474

Code: github.com/ritengzhang/SAE-AIU · Checkpoints: huggingface.co/ritengzhang/SAE-AIU