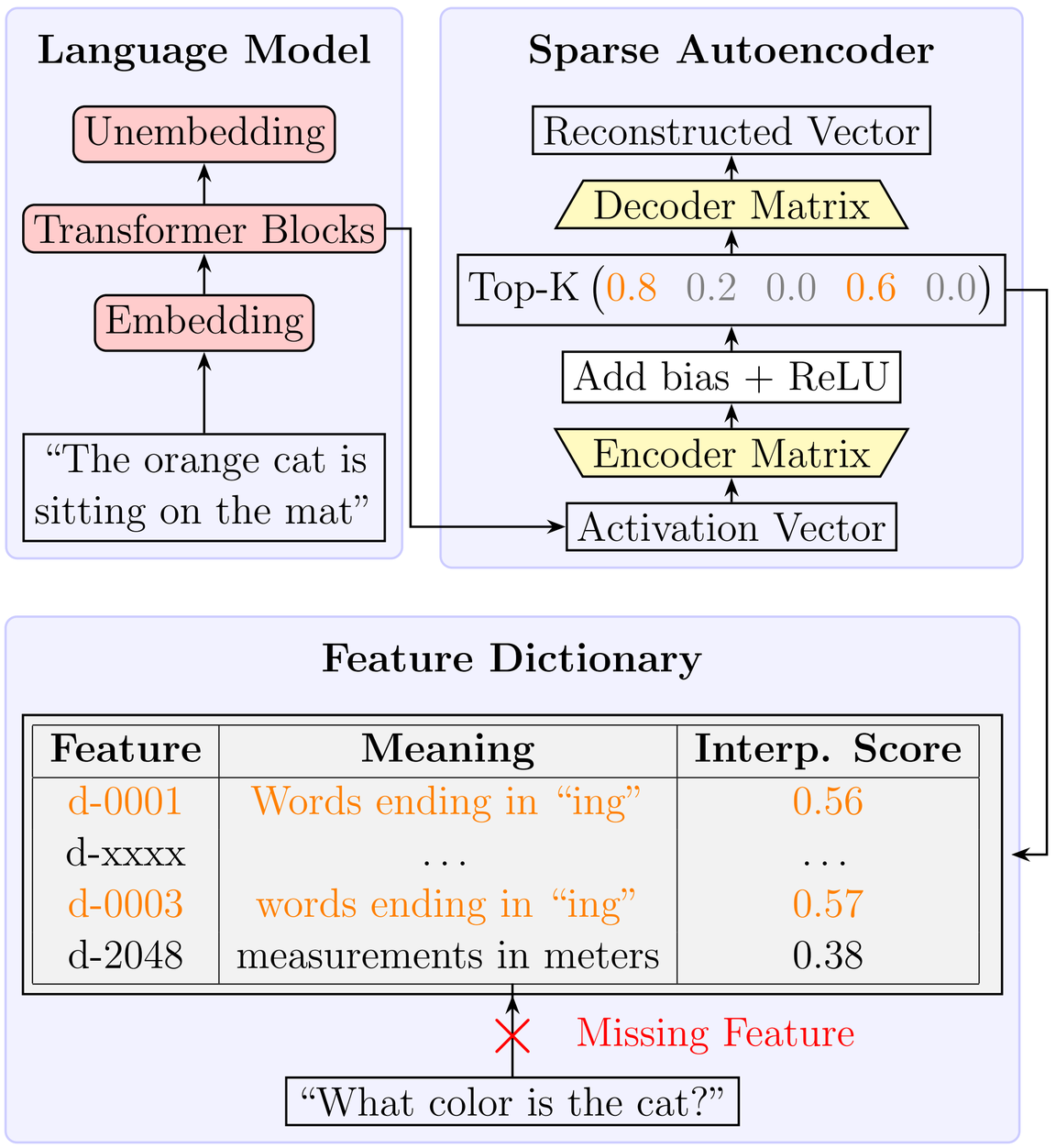
Beyond Sparsity: Improving Diversity in Sparse Autoencoders via Denoising Training

Xiang Pan[1] **Yifei Wang**[2] **Qi Lei**[1]

[1]New York University [2]Massachusetts Institute of Technology

xiangpan@nyu.edu yifei_w@mit.edu ql518@nyu.edu

Problem: Feature Redundancy in SAEs



- **Current SAEs suffer from high feature redundancy**
- **Only 85% of features are unique** in standard Top-K SAEs (1)
- **Redundant features** take up feature space capacity, preventing capture of important semantic directions
- **High interpretation scores** on redundant features can bias evaluation metrics

Example of a redundant feature: "A neuron about the [Beginning Of the Sentence]" — this feature reflects a preprocessing artifact, not a meaningful or interpretable concept.

Evaluation Metrics

1. Feature Coverage Score

$$\frac{1}{d'}\sum_{i=1}^{d'}\mathbb{1}\left(\max_j \frac{(\mathbf{W}_2'^T\mathbf{W}_2)_{ij}}{\|\mathbf{W}_2'[:,i]\|_2\|\mathbf{W}_2[:,j]\|_2} > \tau\right) \tag{1}$$

Measures fraction of feature bank directions covered by current SAE features.

2. Semantic Volume

$$\text{Semantic Volume} = \log\det(\tilde{\mathbf{V}}^T\tilde{\mathbf{V}} + \epsilon\mathbf{I}) \tag{2}$$

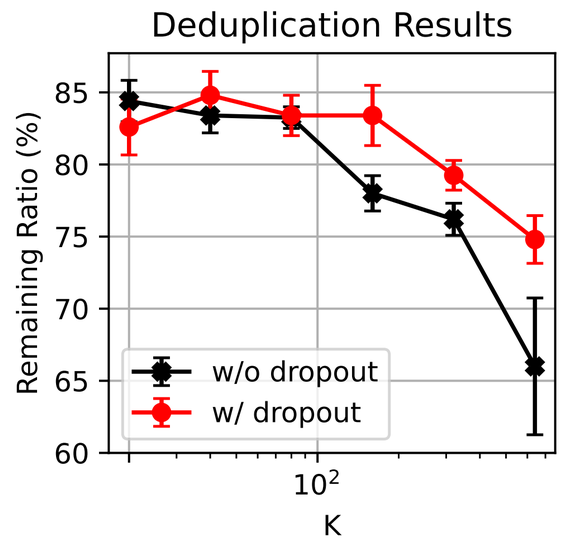
Quantifies diversity of feature explanation space using semantic embeddings.

3. Effective Rank

$$\text{Effective Rank} = \exp\left(-\sum_{i=1}^{r}\tilde{\sigma}_i \log\tilde{\sigma}_i\right) \tag{3}$$

Measures how evenly the feature space is distributed across semantic dimensions.

Deduplication Results



Methodology Overview

Standard SAE Training:

- Input: LLM layer activations $\mathbf{x}$
- Encoder: $\mathbf{z} = \text{TopK}(\text{ReLU}(\mathbf{W}_1\mathbf{x}))$
- Decoder: $\hat{\mathbf{x}} = \mathbf{W}_2\mathbf{z}$
- Loss: $\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \cdot \text{auxiliary}$

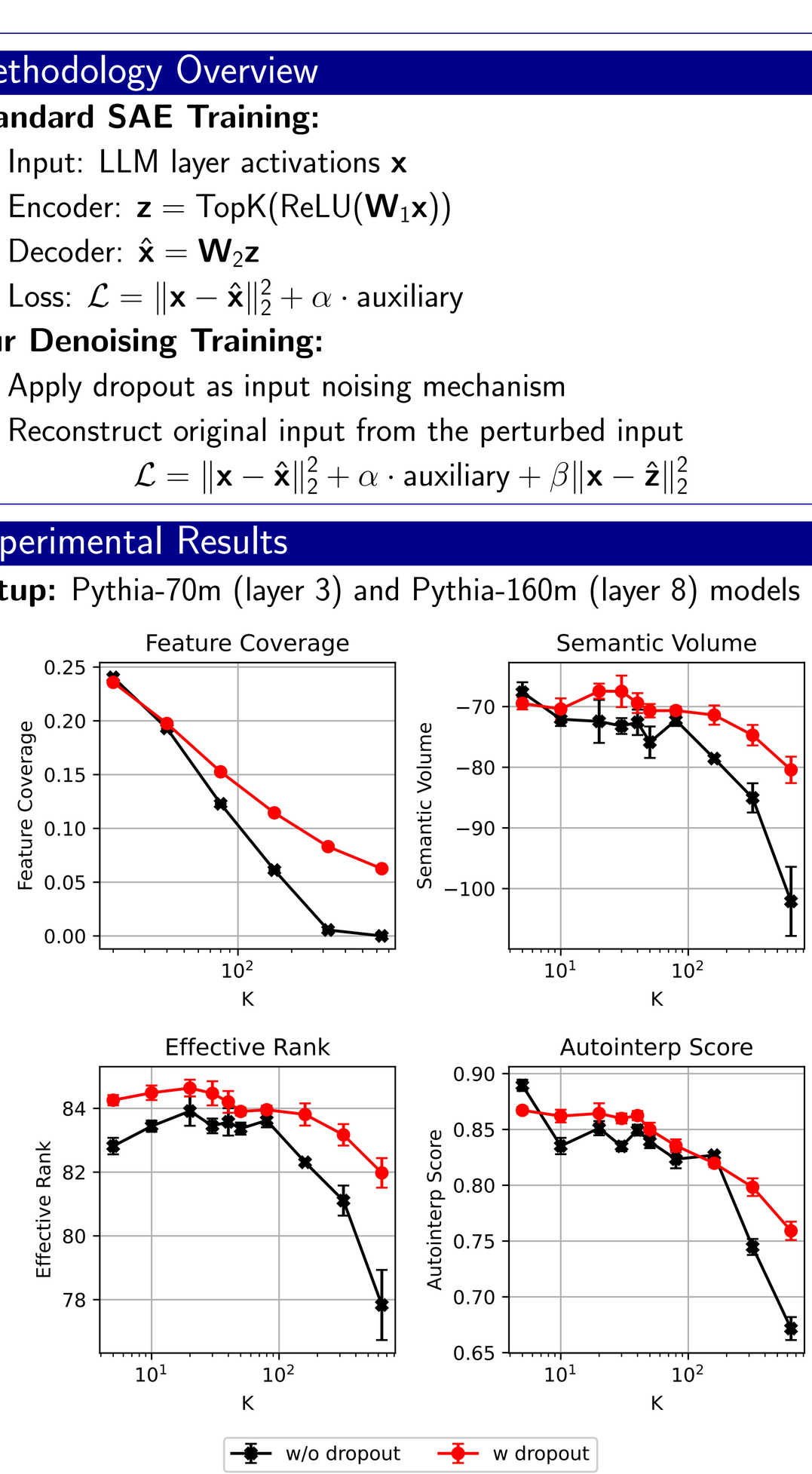
Our Denoising Training:

1. Apply dropout as input noising mechanism
2. Reconstruct original input from the perturbed input

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \cdot \text{auxiliary} + \beta\|\mathbf{x} - \hat{\mathbf{z}}\|_2^2 \tag{4}$$

Experimental Results

Setup: Pythia-70m (layer 3) and Pythia-160m (layer 8) models



Key Findings:

- **Improved Feature Diversity:** Higher feature coverage, semantic volume, and effective rank.
- **Maintained Interpretation Quality:** No degradation in interpretation scores.
- **Better Reconstruction Quality**