# AbsTopK: Rethinking Sparse Autoencoders For Bidirectional Features

Xudong Zhu  Mohammad Mahdi Khalili  Zhihui Zhu

Department of Computer Science and Engineering
The Ohio State University

ICLR 2026
April 23–27, 2026

# Mechanistic Interpretability

1. Mechanistic interpretability aims to decompose hidden representations into meaningful features.
2. Sparse autoencoders (SAEs) aim to recover these features through a sparse latent representation.

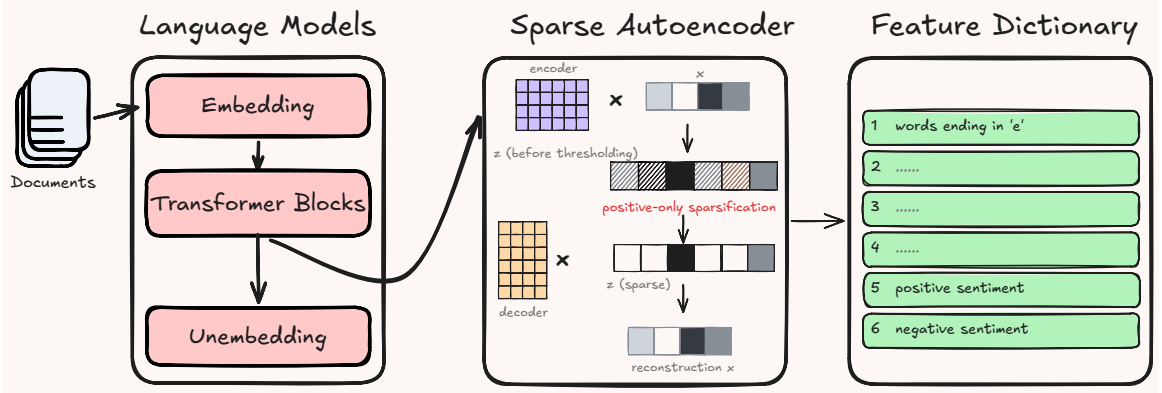


Figure: Overall pipeline of a sparse autoencoder.

# Sparse Autoencoders (SAEs)

Assume the hidden representation $\boldsymbol{x}$ can be decomposed as

$$\boldsymbol{x} = \sum_{p=1}^{P} \alpha_p \boldsymbol{h}_p + \text{residual}.$$

An SAE maps the input $\boldsymbol{x}$ to a sparse code $\boldsymbol{z}$, and then reconstructs the original representation:

$$\text{encoder: } \boldsymbol{z} = \sigma(\boldsymbol{W}^T \boldsymbol{x} + \boldsymbol{b}_e), \qquad \text{decoder: } \hat{\boldsymbol{x}} = \boldsymbol{D}\boldsymbol{z} + \boldsymbol{b}.$$

It is trained by minimizing the following objective:

$$\min_{\substack{\boldsymbol{D},\boldsymbol{W} \in \mathbb{R}^{d \times P} \\ \boldsymbol{b} \in \mathbb{R}^d, \boldsymbol{b}_e \in \mathbb{R}^P}} \mathbb{E}_{\boldsymbol{x}} \left[ \frac{1}{2} \|\boldsymbol{x} - (\boldsymbol{D}\boldsymbol{z} + \boldsymbol{b})\|_2^2 + \lambda R(\boldsymbol{z}) \right].$$

# Concepts

Each row of the decoder $\boldsymbol{D}$ can be interpreted as a concept in the learned dictionary, so the reconstructed hidden state is a linear combination of these concepts:

$$\hat{\boldsymbol{x}} = \sum_{p=1}^{P} z_p \boldsymbol{d}_p + \boldsymbol{b}.$$

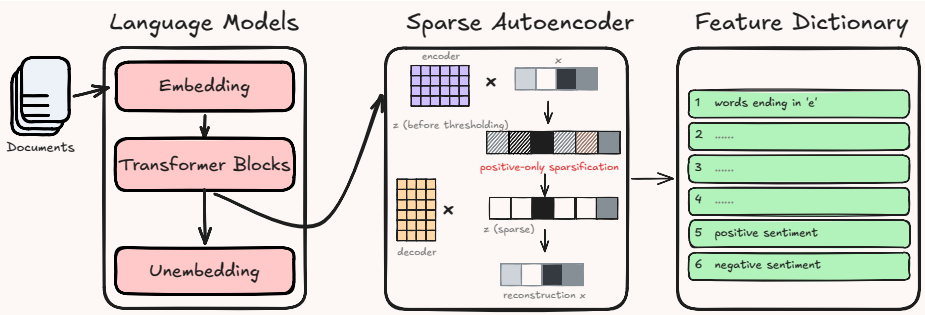


Figure: Each decoder vector can be viewed as a concept in the learned feature dictionary.

# The Proximal Perspective on SAEs

**What distinguishes different SAE variants?**

Our key observation is that several popular SAEs admit a unified view through proximal optimization, rooted in classical sparse coding and dictionary learning.

$$\min_{\substack{\boldsymbol{D},\boldsymbol{W}\in\mathbb{R}^{d\times P}\\ \boldsymbol{b}\in\mathbb{R}^{d},\boldsymbol{b}_e\in\mathbb{R}^{P}}} \mathbb{E}_{\boldsymbol{x}}\left[\frac{1}{2}\left\|\boldsymbol{x}-(\boldsymbol{D}\boldsymbol{z}+\boldsymbol{b})\right\|_2^2+\lambda R(\boldsymbol{z})\right], \qquad \text{where } \boldsymbol{z}=\operatorname{prox}_{\lambda R}\left(\boldsymbol{W}^\top\boldsymbol{x}+\boldsymbol{b}_e\right).$$

- ReLU SAE:

$$R(\boldsymbol{z})=\|\boldsymbol{z}\|_1+\iota_{\{\boldsymbol{z}\geq 0\}}(\boldsymbol{z}) \quad\Longrightarrow\quad \operatorname{prox}_{\lambda R}=\mathrm{ReLU}_\lambda$$

- JumpReLU SAE:

$$R(\boldsymbol{z})=\|\boldsymbol{z}\|_0+\iota_{\{\boldsymbol{z}\geq 0\}}(\boldsymbol{z}) \quad\Longrightarrow\quad \operatorname{prox}_{\lambda R}=\mathrm{JumpReLU}_{\sqrt{2\lambda}}$$

- TopK SAE:

$$R(\boldsymbol{z})=\iota_{\{\|\boldsymbol{z}\|_0\leq k,\ \boldsymbol{z}\geq 0\}}(\boldsymbol{z}) \quad\Longrightarrow\quad \operatorname{prox}_{\lambda R}=\mathrm{TopK}_k$$

# The Non-Negativity Constraint

Popular SAE variants all enforce a non-negativity constraint:

$$\begin{aligned} \text{ReLU SAE:} \quad & R(\boldsymbol{z}) = \|\boldsymbol{z}\|_1 + \iota_{\{\boldsymbol{z} \geq 0\}}(\boldsymbol{z}) \\ \text{JumpReLU SAE:} \quad & R(\boldsymbol{z}) = \|\boldsymbol{z}\|_0 + \iota_{\{\boldsymbol{z} \geq 0\}}(\boldsymbol{z}) \\ \text{TopK SAE:} \quad & R(\boldsymbol{z}) = \iota_{\{\|\boldsymbol{z}\|_0 \leq k,\ \boldsymbol{z} \geq 0\}}(\boldsymbol{z}) \end{aligned}$$

- A latent can only activate in one direction.
- Opposite concepts are split into different features.
- AbsTopK removes this constraint and allows bidirectional features.

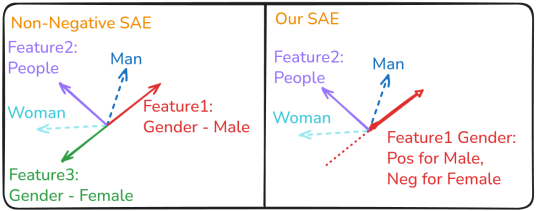


Figure: Non-negativity fragments a semantic axis into separate features, while AbsTopK captures both directions in one latent.

# Beyond Non-Negativity: AbsTopK

Specifically, we use the regularizer $R(\boldsymbol{z}) = \iota_{\{\|\boldsymbol{z}\|_0 \le k\}}$ which removes the non-negativity constraint present in the TopK-inducing regularizer. The corresponding proximal operator is

$$\text{prox}_{\lambda R}(\boldsymbol{u}) = \arg\min_{\boldsymbol{z}\in\mathbb{R}^d} \ \tfrac{1}{2}\|\boldsymbol{u}-\boldsymbol{z}\|_2^2 \quad \text{s.t.} \quad \|\boldsymbol{z}\|_0 \le k, \tag{1}$$

whose closed-form solution is further given by

$$\big(\text{prox}_R(\boldsymbol{u})\big)_i = (\text{AbsTopK}_k(\boldsymbol{u}))_i = \begin{cases} u_i, & i \in \mathcal{H}_k(\boldsymbol{u}), \\ 0, & i \notin \mathcal{H}_k(\boldsymbol{u}), \end{cases} \tag{2}$$

where $\mathcal{H}_k$ denotes the indices of the $k$ largest (in modulus) components.

# Beyond Non-Negativity: AbsTopK

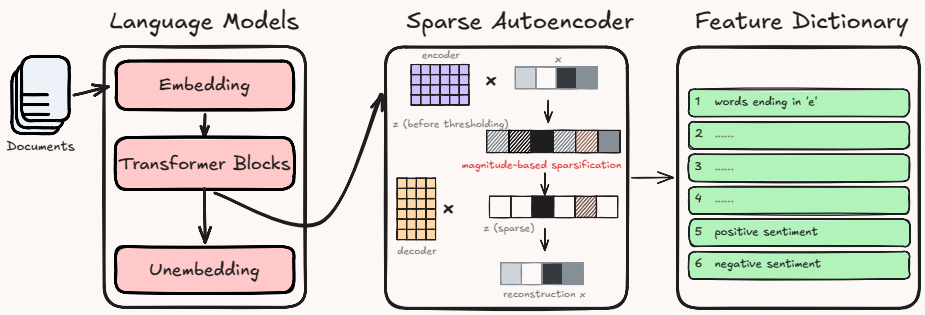


Figure: Instead of keeping only positive activations, AbsTopK selects coefficients by magnitude, preserving both directions of a semantic axis.

# AbsTopK Learns Bidirectional Features

**SAE feature 15741: Positive activation on man**

The man who had traveled across ...

Despite the storm outside , the man remained calm and focused on her work ...

In the bustling market , a man stood by the flower stall ...

**SAE feature 15741: Negative activation on woman**

The woman who had traveled across ...

Despite the storm outside , the woman remained calm and focused on her work ...

In the bustling market , a woman stood by the flower stall ...

Figure: A single AbsTopK feature captures both directions of a semantic axis: positive activation aligns with man, while negative activation aligns with woman.

# AbsTopK Learns Better SAEs

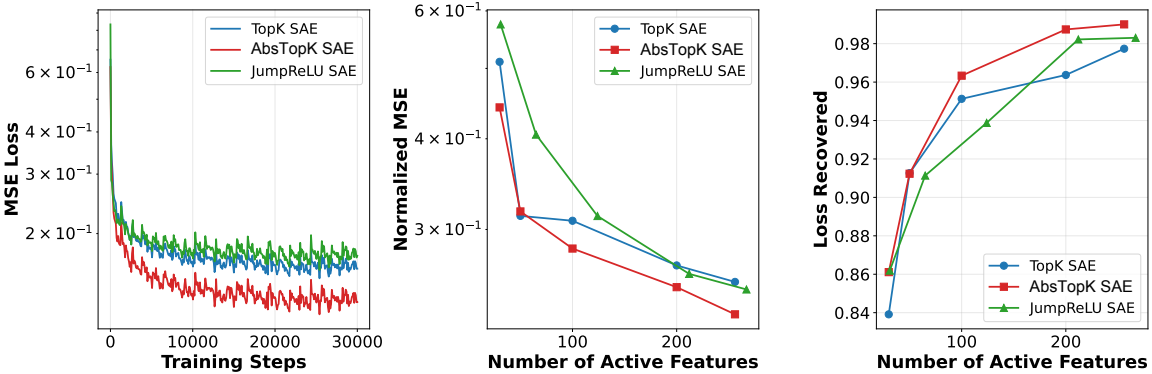


Figure: AbsTopK achieves lower reconstruction error and higher loss recovered than TopK and JumpReLU across different sparsity levels.

# AbsTopK Improves Steering Performance

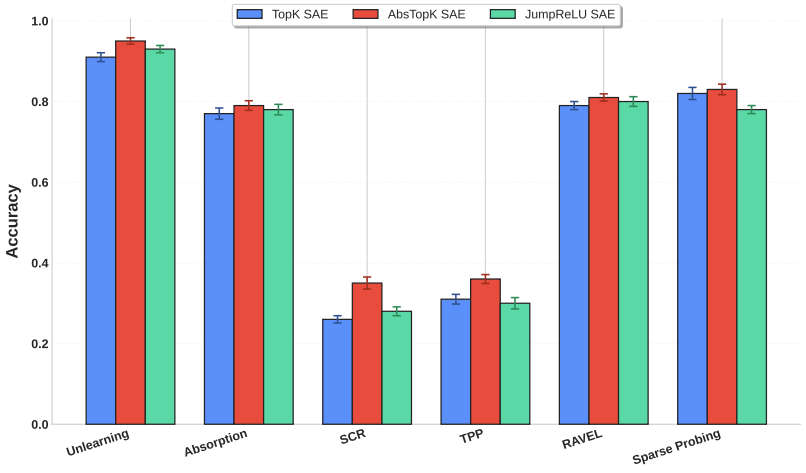


Figure: AbsTopK achieves the best or near-best accuracy across multiple downstream tasks.

# Conclusion

1. Popular SAEs can be understood through a unified proximal perspective, that their shared non-negativity constraint fragments semantic axes.
2. AbsTopK removes this limitation to learn bidirectional features.
3. Empirically, this leads to better interpretability, better reconstruction, and stronger downstream performance.