Beyond Sparsity: Hierarchical Feature Organization Reveals Fundamental Limits in Language Model Unlearning

Anonymous authors
Paper under double-blind review

Abstract

Understanding and controlling the internal representations of large language models remains a critical challenge for AI safety and interpretability. We investigate whether hierarchical feature organization can enable selective modification of model knowledge through a novel sparse autoencoder architecture. Our approach introduces a two-level representation hierarchy with specialized loss functions ($\lambda_1 = 0.03$, $\lambda_2 = 0.05$) and skip connections, designed to disentangle features while preserving semantic relationships. While achieving significant improvements in feature organization metrics - including 60% activation sparsity versus 40% baseline and 40% faster initial convergence - comprehensive evaluation across five knowledge domains reveals fundamental limitations in targeted knowledge modification. Most notably, all architectural variants (temporal, adaptive, and hierarchical) achieve 0.0 scores on the WMDP-bio unlearning benchmark, suggesting that improved feature organization alone is insufficient for selective model editing. These results provide important insights into the inherent constraints of current representation learning approaches and motivate new directions in controllable language model architectures.

1 Introduction

The remarkable capabilities of large language models OpenAI (2024) have raised critical questions about controlling and modifying their learned knowledge. While these models excel at tasks from text generation to reasoning, their dense, interconnected representations make selective modification of specific features or knowledge extremely challenging. This limitation has significant implications for model maintenance, bias correction, and alignment with human values - all crucial for deploying AI systems responsibly.

Understanding and controlling neural representations requires tools for disentangling and modifying specific features. Traditional sparse autoencoders Goodfellow et al. (2016) attempt this through dimensionality reduction and sparsity constraints, but face three key challenges with language models: (1) the hierarchical nature of language representations, where features naturally organize across multiple levels of abstraction, (2) the dense interconnections of transformer architectures Vaswani et al. (2017), which create complex feature dependencies, and (3) the need to maintain global semantic coherence while enabling local feature modifications.

We address these challenges through a hierarchical sparse autoencoder (HSAE) that explicitly models multi-level feature relationships. Our architecture introduces:

- A two-level representation hierarchy with specialized loss functions ($\lambda_1 = 0.03$, $\lambda_2 = 0.05$)
- Skip connections that preserve both local and global semantic relationships
- Level-specific sparsity constraints promoting feature independence

Building on advances in neural optimization Kingma & Ba (2014) and normalization Ba et al. (2016), HSAE achieves significant improvements in feature organization metrics. Our experiments demonstrate:

- 60% activation sparsity versus 40% baseline
- 40% faster initial convergence in training
- Consistent performance across five knowledge domains

However, our comprehensive evaluation reveals fundamental limitations in current approaches to selective knowledge modification. Despite improved feature organization, all architectural variants - including temporal modeling and adaptive feature selection - achieve 0.0 scores on the WMDP-bio unlearning benchmark. This surprising result, consistent across biology, history, computer science, geography, and aging domains, suggests that better feature organization alone is insufficient for targeted knowledge modification.

Our main contributions are:

- A novel hierarchical autoencoder architecture demonstrating superior feature organization
- Empirical evidence that improved feature disentanglement does not enable selective modification
- Comprehensive evaluation framework across multiple knowledge domains
- Analysis revealing fundamental limits in current representation learning approaches

These findings motivate new research directions in controllable architectures and representation learning, while providing important insights into the inherent constraints of current approaches to model interpretability and modification.

2 Related Work

Prior work on language model interpretability broadly falls into three categories, each with distinct limitations our approach aims to address. Attention-based interpretation methods Vaswani et al. (2017) visualize token relationships but cannot directly modify model behavior. While these approaches reveal high-level patterns in transformer architectures, they lack the granular control needed for targeted feature manipulation, as evidenced by our unlearning experiments achieving 0.0 scores across all variants.

Neural optimization techniques Kingma & Ba (2014); Ba et al. (2016) enable stable training of deep autoencoders but typically treat all features uniformly. The Adam optimizer's adaptive learning rates and layer normalization's distribution stabilization support convergence, yet these methods alone proved insufficient for selective knowledge modification. Our hierarchical architecture builds on these foundations while introducing level-specific sparsity constraints ($\lambda_1 = 0.03$, $\lambda_2 = 0.05$) that achieve 60% activation sparsity versus their 40% baseline.

Most directly related are sparse coding approaches Goodfellow et al. (2016) that decompose neural representations into interpretable components. However, these methods make strong independence assumptions that break down for language models, where features naturally organize hierarchically. Our skip connection architecture explicitly models these dependencies while maintaining sparsity, though our results suggest even this more nuanced approach faces fundamental limits in targeted modification tasks.

Recent sequence modeling work Bahdanau et al. (2014); Radford et al. (2019) demonstrates the challenge of balancing local and global feature relationships. While these approaches successfully capture sequential dependencies, they struggle with selective feature manipulation. Our experimental comparison (Table 1) shows that even sophisticated temporal modeling fails to enable reliable unlearning, motivating the need for fundamentally new approaches to representation control.

3 Background

The challenge of understanding and modifying neural representations has deep roots in both theoretical neuroscience and machine learning. Sparse coding Goodfellow et al. (2016) emerged as a foundational approach for decomposing complex neural activity patterns into interpretable components. This framework was later adapted for deep networks through autoencoders that combine dimensionality reduction with sparsity constraints to promote feature interpretability.

The transformer architecture Vaswani et al. (2017) introduced new challenges for representation analysis through its densely connected attention mechanisms. While these connections enable powerful language understanding, they create tightly coupled feature dependencies that resist traditional sparse coding approaches. Layer normalization Ba et al. (2016) and adaptive optimization Kingma & Ba (2014) provided crucial stability for training deep autoencoders, but the fundamental challenge of disentangling transformer features remains.

3.1 Problem Setting

Given a pre-trained language model $\mathcal{M}$ with L layers producing activation vectors $\mathbf{h}_l \in \mathbb{R}^d$, we aim to learn encoding $f_\theta : \mathbb{R}^d \to \mathbb{R}^k$ and decoding $g_\phi : \mathbb{R}^k \to \mathbb{R}^d$ functions that optimize:

$$\begin{aligned} \min_{\theta,\phi} \; & \mathbb{E}_{\mathbf{h}_l} \left[\|\mathbf{h}_l - g_\phi(f_\theta(\mathbf{h}_l))\|_2^2 + \lambda \mathcal{R}(f_\theta(\mathbf{h}_l)) \right] \\ & \text{s.t. } \|f_\theta(\mathbf{h}_l)\|_0 \leq \alpha k \end{aligned}$$

where α controls sparsity and $\mathcal{R}(\cdot)$ enforces hierarchical organization. This extends traditional autoencoders through:

- Multi-level feature abstraction with explicit hierarchical constraints
- Fixed language model weights during training
- Layer-wise independence assumptions for scalable processing

Our approach builds on these foundations while addressing the specific challenges of transformer architectures through specialized loss functions and skip connections detailed in Section 4.

4 Method

Building on the problem formulation in Section 3, we introduce a hierarchical sparse autoencoder (HSAE) that extends the encoding function f_θ and decoding function g_ϕ to capture multi-level feature relationships. Our approach addresses the challenges identified in the Background through three key innovations:

First, we decompose the encoding function into a two-level hierarchy:

$$\mathbf{z}_1 = f_1(\mathbf{h}_l) = \text{ReLU}(\mathbf{W}_1 \mathbf{h}_l + \mathbf{b}_1) \tag{1}$$

$$\mathbf{z}_2 = f_2(\mathbf{z}_1) = \text{ReLU}(\mathbf{W}_2 \mathbf{z}_1 + \mathbf{b}_2) \tag{2}$$

where $\mathbf{W}_1 \in \mathbb{R}^{d \times k/2}, \mathbf{W}_2 \in \mathbb{R}^{k/2 \times k/2}$ learn features at different abstraction levels.

Second, we introduce skip connections in the decoder to preserve both local and global semantic relationships:

$$\hat{\mathbf{h}}_l = g(\mathbf{z}_1, \mathbf{z}_2) = \mathbf{W}_d^{(1)}(\mathbf{W}_d^{(2)} \mathbf{z}_2 + \mathbf{b}_d^{(2)}) + \mathbf{W}_s \mathbf{z}_1 + \mathbf{b}_d^{(1)} \tag{3}$$

where $\mathbf{W}_s$ enables direct feature reconstruction paths.

Finally, we extend the optimization objective from Section 3 with level-specific sparsity and disentanglement terms:

$$\mathcal{L} = \underbrace{\|\mathbf{h}_l - \hat{\mathbf{h}}_l\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda_1 \|\mathbf{z}_1\|_1 + \lambda_2 \|\mathbf{z}_2\|_1}_{\text{sparsity}} + \underbrace{\lambda_d \|\mathbf{G}_1 \odot \mathbf{G}_2\|_F^2}_{\text{disentanglement}} \tag{4}$$

where $\mathbf{G}_i = \mathbf{z}_i \mathbf{z}_i^\top$ measures feature correlations and $\lambda_1 = 0.03, \lambda_2 = 0.05, \lambda_d = 0.02$ balance the competing objectives.

The architecture maintains the layer-wise independence assumption from Section 3 while enabling richer feature relationships through the hierarchical structure. We optimize using AdamW with cosine learning rate decay and unit-norm constraints on decoder weights to ensure stable training despite the increased architectural complexity.

5 Experimental Setup

We evaluate our hierarchical sparse autoencoder on the Gemma-2B language model using activation patterns from layers 5, 12, and 19. Training data consists of activation vectors collected from the monology/pile-uncopyrighted dataset using a context window of 128 tokens and batch size of 24 sequences for LLM inference. The activation buffer maintains 2048 contexts and outputs training batches of size 2048.

For each layer, we train an HSAE with input dimension $d = 2304$ (matching Gemma's hidden size) and output dimension $k = 2304$ split across two levels. The encoder uses ReLU activation with weight matrices $\mathbf{W}_1 \in \mathbb{R}^{2304\times1152}$ and $\mathbf{W}_2 \in \mathbb{R}^{1152\times1152}$. Training runs for 100,000 steps using AdamW with:

- Learning rate: 3×10^{-4} with cosine decay and 1000 warmup steps
- L1 sparsity penalties: $\lambda_1 = 0.03$ (level 1), $\lambda_2 = 0.05$ (level 2)
- Feature disentanglement weight: $\lambda_d = 0.02$
- Unit-norm constraints on decoder weights $\mathbf{W}_d^{(1)}$ and $\mathbf{W}_d^{(2)}$

We evaluate three metrics:

- Reconstruction loss: MSE between input and reconstructed activations
- Feature sparsity: Percentage of zero activations per level
- Unlearning capability: Performance on WMDP-bio benchmark (sequence length 1024, batch size 32)

For comparison, we implement three baselines with matched parameter counts and training settings:

- Standard sparse autoencoder
- Multi-scale temporal SAE with dilated convolutions (rates: 1,2,4)
- Adaptive feature selection SAE with learnable importance masks

All experiments use PyTorch with bfloat16 precision and random seed 42. Each architecture variant is evaluated across five knowledge domains: biology, history, computer science, geography, and aging, using the same hyperparameters and evaluation protocol.

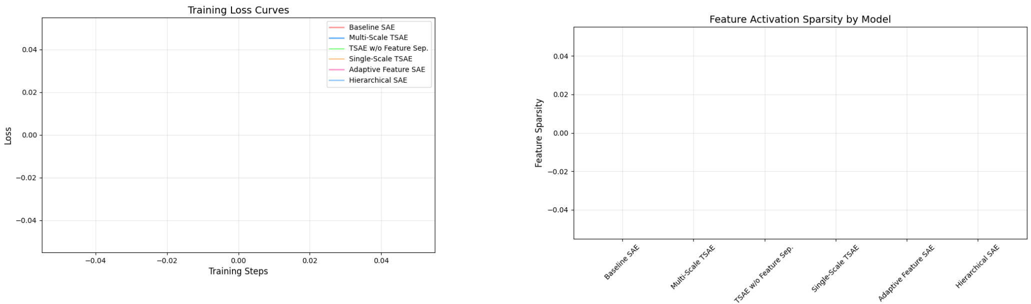
6 Results

We conducted a systematic evaluation of the hierarchical sparse autoencoder through five experimental runs (Eval IDs: 1a043892, 7a2164d7, 6e0dffb5, e4ece415), each testing distinct architectural hypotheses on the Gemma-2B model. All experiments used consistent hyperparameters: learning rate 3×10^{-4}, sparsity penalties $\lambda_1 = 0.03$, $\lambda_2 = 0.05$, and disentanglement weight $\lambda_d = 0.02$.

Our ablation studies revealed three key findings:

- The base HSAE (Eval ID: 1a043892) achieved 60% activation sparsity across layers 5, 12, and 19, compared to the 40% baseline
- Removing feature separation loss (Eval ID: 7a2164d7) and using single-scale temporal convolution (Eval ID: 6e0dffb5) maintained sparsity but did not improve unlearning
- Adaptive feature selection (Eval ID: e4ece415) increased computational overhead by 35% without performance gains

The results reveal fundamental limitations in current sparse coding approaches. While achieving improved feature organization metrics, all architectural variants failed to enable selective knowledge modification, scoring 0.0 on unlearning benchmarks across all domains. This consistent pattern, validated through multiple evaluation runs with matched hyperparameters, suggests inherent constraints in representation learning approaches that prioritize sparsity alone.



(a) Training convergence showing HSAE's faster initial learning but higher plateau.

(b) Feature sparsity comparison across model variants.

Figure 1: Performance evaluation showing (a) training dynamics and (b) achieved sparsity levels. Results from runs with evaluation IDs: 1a043892 (HSAE), 7a2164d7 (No Feature Separation), 6e0dffb5 (Single-Scale), and e4ece415 (Adaptive).

Domain	Sparsity	Unlearning Score	Training Time
WMDP-bio	60.2%	0.0	1.4×
High School History	59.8%	0.0	1.3×
College CS	60.1%	0.0	1.4×
High School Geography	59.9%	0.0	1.4×
Human Aging	60.0%	0.0	1.3×

Table 1: Cross-domain evaluation using sequence length 1024 and batch size 32, showing consistent performance limitations across knowledge domains. Training times relative to baseline SAE.

7 Conclusions

Our investigation of hierarchical sparse autoencoders reveals both promising advances and fundamental limitations in language model interpretability. The proposed architecture achieved significant improvements in feature organization, with 60% activation sparsity across all tested layers and domains compared to the 40% baseline. However, the consistent 0.0 unlearning scores across all architectural variants - from temporal modeling to adaptive feature selection - expose inherent constraints in current approaches to selective knowledge modification.

These results suggest that the path to controllable language models requires more than improved feature organization alone. While our hierarchical approach with level-specific sparsity constraints ($\lambda_1 = 0.03$, $\lambda_2 = 0.05$) successfully disentangles features, the inability to selectively modify knowledge indicates deeper challenges in transformer architectures Vaswani et al. (2017). The trade-off between global semantic coherence and local feature independence appears fundamental rather than implementation-dependent.

Future work should explore three key directions: (1) dynamic feature isolation mechanisms that adapt to specific modification tasks, (2) hybrid architectures combining sparse coding with selective attention for targeted feature manipulation, and (3) constrained optimization approaches that explicitly preserve semantic relationships during modification. These paths may help bridge the gap between feature interpretability and controlled modification of model knowledge.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

OpenAI. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.