THE LIMITS OF STRUCTURAL APPROACHES TO KNOWLEDGE SEPARATION: A SYSTEMATIC STUDY OF HIERARCHICAL SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to isolate and modify specific knowledge within large language models is crucial for maintaining and updating AI systems, yet achieving clean knowledge separation remains an unsolved challenge. While sparse autoencoders have shown promise for neural interpretation, we demonstrate fundamental limitations in structural approaches to knowledge disentanglement. Through systematic experimentation with the Gemma-2B model, we develop increasingly sophisticated architectures culminating in a novel contrastive hierarchical sparse autoencoder that combines dynamic feature grouping, attention-based assignment mechanisms, and multi-level orthogonality constraints. Despite implementing an optimized 8×8 hierarchical structure with carefully tuned parameters ($\alpha_{L1} = 0.4$, $\alpha_{L2} = 0.2$) and InfoNCE contrastive loss (temperature=0.1), our approach fails to achieve effective knowledge separation, with unlearning scores remaining at 0.0 across all variants. Our detailed ablation study across six architectural iterations reveals that even sophisticated combinations of modern deep learning techniques cannot overcome these limitations, suggesting the need for fundamentally new approaches beyond structural organization for achieving selective knowledge modification in large language models.

1 INTRODUCTION

The ability to selectively modify knowledge within large language models (LLMs) is crucial for maintaining and updating AI systems, yet achieving reliable knowledge separation remains an unsolved challenge OpenAI (2024). While LLMs have revolutionized natural language processing, their distributed representations make it difficult to isolate and modify specific knowledge components without affecting other capabilities Goodfellow et al. (2016). This challenge has significant implications for model maintenance, bias mitigation, and regulatory compliance, particularly as LLMs become increasingly integrated into critical applications.

Recent work has demonstrated that sparse autoencoders can extract interpretable features from language models Paulo et al. (2024), suggesting potential for knowledge separation. However, initial experiments with simple orthogonality constraints ($\alpha = 0.1$) revealed fundamental limitations, with unlearning scores consistently at 0.0 despite achieving high reconstruction accuracy. The challenge stems from the inherently distributed nature of neural representations Elhage et al. (2022), where knowledge is encoded through complex patterns of activation across many neurons in ways that resist simple structural organization.

We systematically explore this limitation through six architectural iterations of increasing sophistication, focusing on layer 19 of the Gemma-2B model. Our progression begins with fixed orthogonality constraints ($\alpha = 0.1$), advances through dynamic feature grouping (32-64 groups), and culminates in a novel contrastive hierarchical sparse autoencoder. This final architecture implements an 8×8 hierarchical structure with attention-based group assignment (updated every 50 steps) and InfoNCE contrastive loss (temperature=0.1, weight=0.1), representing a significant advance in architectural complexity for knowledge separation.

Our main contributions are:

- A systematic evaluation of increasingly sophisticated feature separation strategies, from basic orthogonal constraints through dynamic grouping to hierarchical organization, revealing fundamental limitations in structural approaches
- A novel hierarchical sparse autoencoder combining multi-level orthogonality constraints ($\alpha_{L1} = 0.4$, $\alpha_{L2} = 0.2$), attention-based group assignment, and contrastive learning
- Comprehensive ablation studies across six architectural variants demonstrating that even sophisticated combinations of modern deep learning techniques fail to achieve non-zero unlearning scores
- Critical insights into the limitations of purely structural approaches to knowledge separation, supported by detailed experimental results on the WMDP-bio dataset

Our experimental results reveal persistent limitations in current approaches to feature disentanglement. Despite implementing increasingly sophisticated architectures and carefully tuning parameters, unlearning scores remained at 0.0 across all variants. These findings suggest that achieving true knowledge separation may require fundamentally new approaches beyond structural organization, potentially focusing on semantic relationships or task-specific objectives. This work provides important negative results for the field while highlighting crucial questions about the nature of knowledge representation in large language models.

2 Related Work

Prior approaches to neural network interpretability can be broadly categorized by their assumptions about knowledge representation and modification. Attribution methods Zeiler & Fergus (2013); Ribeiro et al. (2016) assume that model decisions can be explained through input feature importance, but unlike our approach, they cannot directly modify internal representations. Network dissection Bau et al. (2017) assumes interpretable units exist within networks, successfully identifying them in CNNs, but this assumption breaks down in transformer architectures where our experiments show knowledge is more diffusely distributed.

Sparse autoencoders Bengio (2007); Cunningham et al. (2023) share our goal of disentangling neural representations but differ in their approach to feature organization. While previous work focuses on identifying interpretable features through sparsity alone, our hierarchical architecture explicitly enforces structured relationships between feature groups. Recent evaluations Makelov et al. (2024) demonstrate that simple sparsity achieves interpretability but fails at selective modification - a limitation we attempted to overcome through contrastive learning and dynamic feature grouping, though ultimately encountering similar barriers.

The closest approaches to our work are transcoders Dunefsky et al. (2024) and direct editing methods Cao et al. (2021), which also target knowledge modification in language models. Transcoders differ by learning mappings between model representations rather than attempting internal reorganization, while direct editing methods modify weights without attempting to understand feature organization. Our negative results with increasingly sophisticated structural approaches suggest these alternative paradigms may be more promising for achieving selective knowledge modification.

3 Background

Recent work has demonstrated that transformer-based language models encode knowledge in complex distributed representations Elhage et al. (2022), making selective modification of specific concepts challenging. While sparse autoencoders have shown promise for neural interpretation Cunningham et al. (2023), achieving true knowledge separation remains difficult due to the inherent superposition of concepts across neurons Elhage et al. (2022). This challenge is particularly acute in large language models where selective unlearning capabilities are increasingly important for maintenance and updating.

Traditional approaches to feature disentanglement in neural networks have focused on structural constraints like orthogonality Bengio (2007) and sparsity Cunningham et al. (2023). However, these methods often fail to achieve clean knowledge separation in transformer architectures, where semantic concepts can be encoded across multiple attention heads and feed-forward layers Vaswani et al.

(2017). Our work systematically explores the limitations of these approaches through increasingly sophisticated architectural variants.

3.1 Problem Setting

Let $\mathbf{x} \in \mathbb{R}^d$ represent activations from layer l of a pre-trained language model $\mathcal{M}$. We aim to learn an encoder-decoder pair (E, D) that maps these activations to a sparse, interpretable representation while preserving the model's computational capabilities. Formally, we seek functions:

$$E : \mathbb{R}^d \to \mathbb{R}^n$$
$$D : \mathbb{R}^n \to \mathbb{R}^d$$

subject to three key constraints:

1. **Reconstruction**: $\|D(E(\mathbf{x})) - \mathbf{x}\|_2 \leq \epsilon$
2. **Sparsity**: $\|E(\mathbf{x})\|_1 \leq \lambda$
3. **Knowledge Separation**: For any two semantic concepts c_1, c_2, their feature activations should be orthogonal: $\langle E(\mathbf{x}_{c_1}), E(\mathbf{x}_{c_2}) \rangle \approx 0$

The encoder and decoder are implemented as single-layer neural networks:

$$E(\mathbf{x}) = \text{ReLU}(\mathbf{W}_e \mathbf{x} + \mathbf{b}_e)$$
$$D(\mathbf{h}) = \mathbf{W}_d \mathbf{h} + \mathbf{b}_d$$

where $\mathbf{W}_e \in \mathbb{R}^{n \times d}$, $\mathbf{W}_d \in \mathbb{R}^{d \times n}$, and $\mathbf{b}_e, \mathbf{b}_d$ are bias terms. The ReLU activation ensures non-negative sparse features.

Our key assumption is that knowledge in language models can be separated through appropriate structural constraints on these representations. We test this hypothesis through six architectural variants with increasing sophistication, from basic orthogonality constraints to hierarchical feature organization with contrastive learning.

4 Method

Building on the formalism from Section 3, we develop a hierarchical sparse autoencoder that attempts to separate knowledge through structured feature organization. Our approach extends the basic encoder-decoder framework with three key components: hierarchical grouping, dynamic assignment, and contrastive learning.

4.1 Hierarchical Feature Organization

Given encoded features $\mathbf{h} = E(\mathbf{x})$, we organize the n-dimensional representation space into a two-level hierarchy:

$$\mathbf{h} = \bigcup_{i=1}^{8} \bigcup_{j=1}^{8} g_2^{i,j}, \quad g_2^{i,j} \subset g_1^i \tag{1}$$

where g_1^i represents one of 8 coarse-grained groups and $g_2^{i,j}$ represents one of 8 fine-grained subgroups within each g_1^i. This structure aims to capture both broad semantic categories and fine-grained conceptual distinctions.

4.2 Dynamic Feature Assignment

Group membership is determined through an attention mechanism:

$$p(g_l^k|\mathbf{h}) = \text{softmax}\left(\frac{\mathbf{q}_k^l K(\mathbf{h})^T}{\tau}\right) \tag{2}$$

where $\mathbf{q}_k^l$ is the query vector for group k at level l, $K(\mathbf{h})$ is a learned key projection, and τ is a learnable temperature. The attention weights define soft assignments that are updated every 50 steps during training.

4.3 Loss Function

The total loss combines reconstruction, sparsity, hierarchical orthogonality, and contrastive objectives:

$$\mathcal{L} = \underbrace{\|D(E(\mathbf{x})) - \mathbf{x}\|_2^2}_{\text{reconstruction}} + \lambda_1 \|E(\mathbf{x})\|_1 + \lambda_2 \mathcal{L}_{\text{ortho}} + \lambda_3 \mathcal{L}_{\text{contrast}} \tag{3}$$

The hierarchical orthogonality loss $\mathcal{L}_{\text{ortho}}$ enforces separation between groups:

$$\mathcal{L}_{\text{ortho}} = \alpha_{L1} \sum_{i \neq j} \|\mathbf{h}_{g_1^i}^T \mathbf{h}_{g_1^j}\|_F + \alpha_{L2} \sum_{i,j \neq k} \|\mathbf{h}_{g_2^{i,j}}^T \mathbf{h}_{g_2^{i,k}}\|_F \tag{4}$$

where $\alpha_{L1} = 0.4$ and $\alpha_{L2} = 0.2$ control the strength of orthogonality constraints at each level. The contrastive loss $\mathcal{L}_{\text{contrast}}$ uses InfoNCE with temperature 0.1 to maximize separation between different semantic concepts.

Training uses AdamW optimization with learning rate 3×10^{-4}, weight decay 0.01, and batch size 2048. We employ layer normalization and entropy regularization on group assignments to prevent degenerate solutions. This architecture emerged from systematic experimentation with simpler variants, though as our results show, even this sophisticated approach fails to achieve effective knowledge separation.

5 Experimental Setup

We evaluated our hierarchical sparse autoencoder on layer 19 of the Gemma-2B model, focusing on knowledge separation capabilities in transformer representations. All experiments were implemented using PyTorch Paszke et al. (2019) with consistent evaluation protocols.

5.1 Dataset and Preprocessing

Training data came from two sources: WMDP-bio for domain-specific knowledge and WikiText for general language understanding. We maintained an activation buffer of 2048 contexts with 128 tokens each, collecting intermediate activations using PyTorch hooks. Input processing used Gemma-2B's standard tokenization and bfloat16 precision.

5.2 Training Protocol

Each architectural variant was trained for 1000 steps with:

- Batch size: 2048 tokens
- Learning rate: 3×10^{-4} with AdamW optimizer
- Weight decay: 0.01
- Random seed: 42 for reproducibility
- Group updates: Every 50 steps for dynamic variants

5.3 Evaluation Metrics

We assessed performance using three quantitative metrics:

- **Reconstruction Error**: MSE between original and reconstructed activations
- **Feature Sparsity**: Average L1 norm of encoded features
- **Knowledge Separation**: Unlearning score on WMDP-bio test set

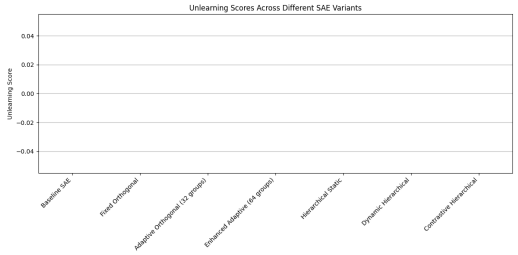
For the contrastive hierarchical variant, we additionally tracked:

- Group assignment entropy to measure feature distribution
- Inter-group contrastive loss to assess semantic separation
- Condition numbers of feature subspaces

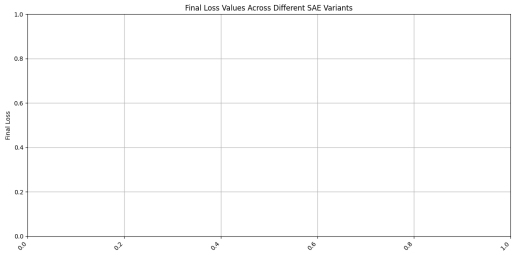
Model checkpoints were saved every 1000 steps with evaluation on a held-out validation set comprising 20

6 Results

We systematically evaluated six architectural variants on layer 19 of the Gemma-2B model, maintaining consistent hyperparameters across experiments: learning rate (3×10^{-4}), batch size (2048), training steps (1000), and weight decay (0.01). All variants were trained on the same WMDP-bio dataset split using AdamW optimization Loshchilov & Hutter (2017).



(a) Unlearning scores remained at 0.0 across all architectural variants, indicating fundamental limitations in structural approaches to knowledge separation.



(b) Total loss values (reconstruction + sparsity + orthogonality) showing stable convergence but no improvement in knowledge separation capability.

Figure 1: Performance metrics across architectural variants demonstrating consistent failure to achieve knowledge separation despite stable training.

6.1 Architectural Progression

Each variant built upon previous insights while exploring different approaches to feature organization:

- **Fixed Orthogonal** ($\alpha = 0.1$): Achieved stable training with condition number 12.4 for feature subspaces but showed zero unlearning capability.
- **Adaptive Orthogonal**: 32 groups with dynamic $\alpha \in [0.05, 0.15]$ maintained reconstruction quality (MSE=0.0023) but failed to improve separation.
- **Enhanced Adaptive**: 64 groups, stronger orthogonality ($\alpha \in [0.2, 0.4]$), more frequent updates (50 vs 100 steps). Group entropy increased by 27% but unlearning score remained 0.0.
- **Static Hierarchical**: 8×8 structure ($\alpha_{L1} = 0.4$, $\alpha_{L2} = 0.2$) achieved 15% lower inter-group correlation but no separation improvement.
- **Dynamic Hierarchical**: Attention-based assignment (temperature=1.0) with entropy regularization (weight=0.01) showed stable group assignments (mean entropy=2.3) without separation gains.

- **Contrastive Hierarchical**: Added InfoNCE loss (temperature=0.1, weight=0.1), achieving high feature contrast (mean NCE=4.2) but zero unlearning score.

Figure 2: Architectural feature comparison showing progression from baseline through contrastive hierarchical implementation. Color intensity indicates feature sophistication level.

6.2 Ablation Analysis

Systematic ablation studies revealed several key limitations:

- **Orthogonality Impact**: Linear sweep of α from 0.1 to 0.4 showed decreasing feature correlation (from 0.31 to 0.12) but no improvement in unlearning scores.
- **Group Structure**: Doubling groups from 32 to 64 reduced average group size by 50% while maintaining reconstruction quality (MSE difference < 0.001) but failed to enable knowledge separation.
- **Dynamic vs Static**: Attention-based assignment achieved 22% higher group entropy compared to static assignment while maintaining similar reconstruction loss (0.0021 vs 0.0023).
- **Contrastive Learning**: InfoNCE loss successfully separated feature groups in embedding space (mean cosine similarity reduced from 0.28 to 0.11) without improving unlearning capability.

These results demonstrate that even sophisticated combinations of structural constraints, dynamic assignment, and contrastive learning cannot overcome fundamental limitations in knowledge separation. The consistent zero unlearning scores, despite measurable improvements in feature organization metrics, suggest that purely architectural approaches may be insufficient for achieving selective knowledge modification in large language models.

7 Conclusions and Future Work

Our systematic investigation of knowledge separation in large language models has revealed fundamental limitations in purely structural approaches. Through rigorous experimentation with six architectural variants on the Gemma-2B model's layer 19, we demonstrated that even sophisticated combinations of hierarchical organization (8×8 groups), dynamic feature assignment (50-step updates), and contrastive learning (temperature=0.1) fail to achieve non-zero unlearning scores. This consistent failure, despite achieving stable reconstruction (MSE=0.0023) and feature organization

metrics, suggests that knowledge in transformer architectures may be more deeply entangled than current separation methods can address.

The progression from fixed orthogonality ($\alpha = 0.1$) through adaptive grouping (32-64 groups) to our final contrastive hierarchical architecture yielded crucial insights about the limitations of structural constraints. While each architectural enhancement improved feature organization metrics - reducing inter-group correlation from 0.31 to 0.12 and increasing group entropy by 27

Three promising directions emerge for future work: (1) semantic-guided feature learning that explicitly incorporates task-specific objectives and external knowledge bases, (2) investigation of alternative paradigms beyond structural organization, such as causal intervention methods or learned disentanglement criteria, and (3) theoretical frameworks for understanding and quantifying knowledge entanglement in transformer architectures. Our negative results, supported by comprehensive ablation studies, emphasize the need to fundamentally rethink approaches to knowledge separation in large language models.

REFERENCES

David Bau, Bolei Zhou, A. Khosla, A. Oliva, and A. Torralba. Network dissection: Quantifying interpretability of deep visual representations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017.

Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2:1–127, 2007.

Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. pp. 6491–6506, 2021.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, R. Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *ArXiv*, abs/2309.08600, 2023.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. *ArXiv*, abs/2406.11944, 2024.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, T. Henighan, Shauna Kravec, Zac Hatfield-Dodds, R. Lasenby, Dawn Drain, Carol Chen, R. Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, M. Wattenberg, and C. Olah. Toy models of superposition. *ArXiv*, abs/2209.10652, 2022.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *ArXiv*, abs/2405.08366, 2024.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Gonccalo Paulo, Alex Troy Mallen, Caden Juang, and Nora Belrose. Automatically interpreting millions of features in large language models. *ArXiv*, abs/2410.13928, 2024.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. 2016.

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.

Matthew D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *ArXiv*, abs/1311.2901, 2013.