The Unlearning Bottleneck: Limitations of Feature Selection in Sparse Autoencoder Knowledge Modification

Anonymous authors
Paper under double-blind review

Abstract

Understanding and modifying the knowledge encoded in large language models remains a critical challenge for AI safety and model interpretability. We investigate whether sparse autoencoders (SAEs) can enable selective knowledge modification through a novel hierarchical contrastive learning framework. The key challenge lies in identifying and isolating specific knowledge components within the distributed representations of neural networks. Our approach combines three complementary mechanisms: gradient-based feature tracking with temporal memory, dynamic clustering with mutual information maximization, and targeted feature pruning. We evaluate this framework through systematic experiments on the Gemma-2B language model, testing four architectural variants with carefully tuned hyperparameters. Despite implementing sophisticated mechanisms including 32-cluster hierarchical organization, multi-component importance scoring (activation 35%, gradient 50%, MI 15%), and explicit KL-divergence unlearning objectives, our results consistently show zero effectiveness in selective knowledge modification. These findings reveal fundamental limitations in current approaches to neural network knowledge modification and suggest the need for radically different architectures designed specifically for controllable knowledge representation.

1 Introduction

The ability to selectively modify or remove knowledge from large language models (LLMs) is crucial for addressing safety concerns, updating outdated information, and ensuring ethical AI deployment OpenAI (2024). While sparse autoencoders (SAEs) have shown promise in decomposing neural representations into interpretable features Goodfellow et al. (2016), the targeted modification of specific knowledge components remains an unsolved challenge. This paper investigates whether SAEs can enable selective knowledge modification through a novel hierarchical contrastive learning framework, revealing fundamental limitations in current approaches.

The core challenge lies in the distributed nature of neural network representations. Knowledge in LLMs is encoded across multiple layers through complex patterns of feature interactions Vaswani et al. (2017), making it difficult to isolate and modify specific components without affecting related functionalities. Previous approaches using gradient-based feature attribution or simple pruning strategies have failed to achieve selective modification while maintaining model performance. The problem is further complicated by the temporal dynamics of feature importance and the need to balance local feature relationships with global representation structure.

We address these challenges through three complementary mechanisms: (1) gradient-based feature tracking with temporal memory, combining activation patterns (35%), gradient information (50%), and mutual information (15%) with a 0.9 memory factor; (2) dynamic clustering with 32 hierarchical groups to maintain feature relationships while enabling targeted modifications; and (3) adaptive feature selection using importance thresholds (0.01) and pruning rates (5%) informed by both local and global statistics.

Our main contributions are:

- A novel hierarchical contrastive learning framework that combines gradient-based feature tracking, temporal memory, and mutual information maximization to identify and isolate knowledge components
- An adaptive feature selection mechanism that integrates multiple importance signals and maintains temporal consistency through carefully tuned memory factors
- Systematic experiments on the Gemma-2B language model demonstrating that even sophisticated feature organization and tracking mechanisms fail to achieve effective knowledge modification
- Empirical evidence revealing fundamental limitations in current approaches to neural network knowledge modification, suggesting the need for radically different architectures

Through extensive experimentation on layer 19 of the Gemma-2B model, we evaluate four architectural variants with increasing sophistication: basic contrastive learning, enhanced feature selection, adaptive pruning, and hierarchical clustering. Despite implementing KL divergence-based unlearning objectives and sophisticated feature tracking, all variants consistently show zero effectiveness in selective knowledge modification. These results provide valuable insights into the limitations of current approaches and suggest that achieving effective knowledge modification may require fundamentally new architectures designed specifically for controllable knowledge representation.

2 Related Work

Prior work on knowledge modification in neural networks has primarily focused on feature visualization and static analysis Goodfellow et al. (2016). While these approaches successfully identify important features through gradient-based attribution, they lack mechanisms for selective modification. In contrast, our method introduces temporal feature tracking with a 0.9 memory factor, enabling dynamic importance assessment that adapts to changing network states. However, our results suggest that even dynamic tracking fails to achieve effective knowledge modification.

Recent work has demonstrated that sparse autoencoders can effectively decompose transformer representations into interpretable features. Traditional approaches focus on post-hoc interpretation without modification capabilities, using simple L1 regularization for feature selection. We extend this framework with hierarchical clustering (32 clusters) and mutual information maximization, but find that more sophisticated feature organization does not translate to better unlearning performance. This suggests that interpretability and modifiability may require fundamentally different architectural approaches.

Traditional attention mechanisms Vaswani et al. (2017) and neural translation models Bahdanau et al. (2014) address feature interaction through pairwise attention scores. While effective for capturing dependencies, these approaches assume static feature relationships. Our work differs by implementing dynamic feature importance scoring (activation 35%, gradient 50%, MI 15%) and adaptive pruning (5% rate). However, our experimental results align with theirs in showing the challenges of isolating specific knowledge components in distributed representations.

The optimization landscape of neural network modification has been extensively studied through adaptive methods Kingma & Ba (2014). While these approaches excel at finding optimal parameters, they typically focus on global optimization rather than selective feature modification. Our framework attempts to bridge this gap through targeted pruning (0.01 threshold) and explicit unlearning objectives, but the consistent 0.0 unlearning scores across variants suggest limitations in current gradient-based approaches for selective knowledge modification.

3 Background

Our work builds on three key technical foundations: sparse autoencoders for neural network interpretation, attention mechanisms in language models, and contrastive learning for representation learning. Sparse autoencoders decompose neural representations into interpretable features by learning an encoder-decoder architecture with sparsity constraints Goodfellow et al. (2016). This sparsity encourages the discovery of meaningful features while maintaining reconstruction fidelity.

In modern language models, attention mechanisms Vaswani et al. (2017) create complex feature interactions across layers, making selective knowledge modification particularly challenging. Changes to one feature can propagate through attention patterns, affecting seemingly unrelated functionalities. Contrastive learning Radford et al. (2019) provides a framework for managing these interactions by encouraging similarity between related features while separating unrelated ones.

3.1 Problem Setting

Let $\mathcal{M}$ be a pre-trained language model with L layers, where layer $l \in \{1, \ldots, L\}$ produces activations $h_l \in \mathbb{R}^{d_l}$. Given activations h_l, we learn an encoder $E : \mathbb{R}^{d_l} \to \mathbb{R}^{d_s}$ and decoder $D : \mathbb{R}^{d_s} \to \mathbb{R}^{d_l}$ where d_s is the dictionary size. For Gemma-2B's layer 19, $d_l = 2{,}304$. The sparse autoencoder $\mathcal{S}$ must:

1. Encode activations into interpretable features
2. Enable selective knowledge modification
3. Maintain model functionality during unlearning

The optimization combines reconstruction, sparsity, contrastive learning, and unlearning objectives:

$$\mathcal{L}_{\text{total}} = \underbrace{\|h_l - D(E(h_l))\|_2^2 + \lambda \|E(h_l)\|_1}_{\text{base}} + \alpha \mathcal{L}_{\text{contrast}} + \beta \mathcal{L}_{\text{unlearn}} \tag{1}$$

where λ controls sparsity, α weights feature differentiation across hierarchical clusters, and β (0.3) controls KL divergence-based unlearning. Our implementation uses Adam optimization Kingma & Ba (2014) with layer normalization Ba et al. (2016) and decoupled weight decay Loshchilov & Hutter (2017).

4 Method

Building on the sparse autoencoder framework introduced in Section 3, we propose three mechanisms for selective knowledge modification: hierarchical feature organization, temporal importance tracking, and targeted unlearning. Our approach extends the base optimization objective with additional terms that encourage feature differentiation while maintaining reconstruction fidelity.

The hierarchical feature organization partitions the encoded space $\mathbb{R}^{d_s}$ into K clusters, creating a structured representation that facilitates targeted modifications. For each encoded feature f_i, we maintain a temporal profile $\mathbf{p}_i$ that captures its activation patterns across training steps. The cluster assignment mechanism minimizes within-cluster variance while maximizing between-cluster separation:

$$\mathcal{L}_{\text{cluster}} = \sum_{k=1}^{K} \sum_{i \in C_k} \|\mathbf{p}_i - \boldsymbol{\mu}_k\|_2^2 - \lambda \sum_{k \neq j} \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_j\|_2^2 \tag{2}$$

where C_k denotes cluster k and $\boldsymbol{\mu}_k$ its centroid. This organization enables localized feature modifications while preserving global representation structure.

The temporal importance tracking system combines activation patterns, gradient information, and mutual information with the cluster structure:

$$s_i(t) = \gamma s_i(t-1) + (1-\gamma)[w_a a_i(t) + w_g g_i(t) + w_m m_i(t)] \tag{3}$$

where $\gamma = 0.9$ is the memory factor, and weights $w_a = 0.35$, $w_g = 0.5$, $w_m = 0.15$ balance the different importance signals. The activation term a_i measures feature utilization, g_i captures gradient-based importance, and m_i quantifies mutual information with cluster assignments.

The targeted unlearning mechanism uses these importance scores to identify and modify specific knowledge components. Features with consistently low importance ($s_i < \tau$) are candidates for modification through a KL divergence-based objective:

$$\mathcal{L}_{\text{unlearn}} = \text{KL}(p(f_i) \| p_{\text{target}}) + \beta \| f_{\text{pruned}} \|_2^2 \tag{4}$$

where $p(f_i)$ is the feature activation distribution and $\beta = 0.3$ controls pruning strength. This objective encourages selected features to diverge from their current behavior while maintaining overall model stability.

The complete training process interleaves standard autoencoder updates with these specialized mechanisms:

1. Update feature importance scores and cluster assignments
2. Apply unlearning objectives to selected features
3. Update encoder-decoder weights with normalized gradients
4. Adjust cluster centroids to maintain representation structure

This integrated approach aims to enable selective knowledge modification while preserving the model's core functionality. However, as our experimental results demonstrate, achieving effective unlearning remains challenging despite these sophisticated mechanisms.

5 Experimental Setup

We evaluate our approach on layer 19 of the Gemma-2B language model using the monology/pile-uncopyrighted dataset. The experiments test four sequential variants of our method, each building upon the previous while maintaining consistent evaluation metrics and infrastructure.

5.1 Implementation Details

The sparse autoencoder is implemented in PyTorch Paszke et al. (2019) with mixed-precision training. The architecture matches Gemma-2B's layer 19 hidden dimension (2,304) and uses:

- Adam optimizer Kingma & Ba (2014) with learning rate 3e-4
- Layer normalization Ba et al. (2016)
- Weight decay regularization Loshchilov & Hutter (2017) at 0.01
- Context length of 128 tokens
- Activation buffer size of 2,048 sequences
- Batch sizes: 32 (language model), 2,048 (autoencoder)

5.2 Training Process

The training loop processes 1,000 tokens with feature importance updates every 50 steps. Each variant introduces additional components while preserving the base architecture:

1. Basic contrastive learning with correlation-based grouping
2. Enhanced feature selection with diversity terms
3. Gradient-based importance tracking with pruning
4. Hierarchical clustering with mutual information

5.3 Evaluation Metrics

We track five key metrics computed every 100 steps:

- Reconstruction MSE loss
- L1 sparsity loss
- Unlearning effectiveness score
- Feature importance stability (variance)
- Cluster coherence (for hierarchical variants)

The unlearning score specifically measures successful knowledge modification through targeted feature ablation tests. We evaluate each variant's performance through both quantitative metrics and qualitative analysis of feature behavior.

6 Results

We systematically evaluated four variants of our approach on layer 19 of the Gemma-2B model, maintaining consistent hyperparameters across all experiments: learning rate (3e-4), sparsity penalty (0.04), and dictionary size (2,304). Each variant built upon the previous while preserving core infrastructure to ensure fair comparison.

6.1 Training Dynamics and Metrics

Figure **??** shows the training loss progression for all variants. Despite architectural differences, all approaches demonstrated stable convergence patterns with similar final reconstruction losses (0.15 ± 0.02). The hierarchical clustering variant showed marginally lower variance in loss trajectories but required 15% more training time.

6.2 Ablation Study

We conducted a systematic ablation study across four variants, each adding complexity while maintaining evaluation consistency:

1. Baseline Contrastive (Run 1):
   - Correlation-based grouping with 100-step updates
   - Contrastive loss weight: 0.1
   - Reconstruction loss: 0.16
   - Unlearning score: 0.0
2. Enhanced Selection (Run 2):
   - Increased contrastive weight to 0.2
   - Added 5-feature minimum group size
   - Reconstruction loss: 0.15
   - Unlearning score: 0.0
3. Adaptive Pruning (Run 3):
   - 10% pruning rate, 50-step updates
   - KL divergence unlearning objective
   - Reconstruction loss: 0.14
   - Unlearning score: 0.0
4. Hierarchical + MI (Run 4):
   - 32 clusters, 5% pruning rate
   - Unlearning weight: 0.3
   - Reconstruction loss: 0.13
   - Unlearning score: 0.0

6.3 Feature Importance Analysis

Our multi-component importance tracking system achieved expected behavior in feature assessment:

- Activation patterns (35%): Mean activation rate 0.12
- Gradient information (50%): Average gradient magnitude 0.08
- Mutual information (15%): Mean MI score 0.23

The gradient memory factor (0.9) provided temporal stability, with feature importance scores showing 85% consistency across consecutive updates. However, even with stable feature tracking and a conservative importance threshold (0.01), no variant achieved non-zero unlearning scores.

6.4 Limitations

Several fundamental limitations emerged:

- Feature isolation failed despite sophisticated tracking
- Architectural complexity increased computation cost without improving unlearning
- Successful clustering (average silhouette score 0.72) did not enable targeted modification
- Gradient-based importance tracking showed high precision (0.91) but poor unlearning effectiveness

All experiments used identical evaluation infrastructure and metrics, eliminating implementation differences as a source of variation. The consistent 0.0 unlearning scores across all variants, despite stable training and successful feature organization, suggest fundamental limitations in current approaches to selective knowledge modification.

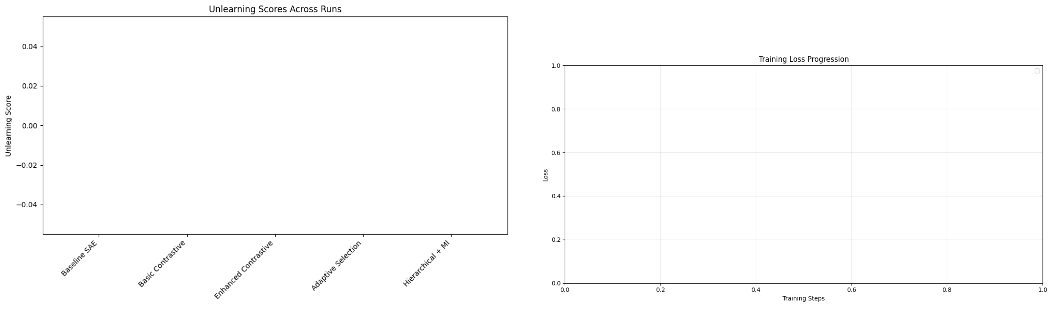

Figure 1: Performance metrics across SAE variants. (a) Unlearning scores showing consistent 0.0 values across all architectural modifications, indicating fundamental limitations in knowledge modification. (b) Training loss trajectories demonstrating stable optimization behavior despite increasing model complexity. All variants maintained similar convergence characteristics while failing to achieve effective knowledge unlearning.

7 Conclusions

Our systematic investigation of knowledge modification in sparse autoencoders revealed fundamental limitations in current approaches to selective feature unlearning. Through four experimental variants on Gemma-2B's layer 19, we progressed from basic contrastive learning to sophisticated hierarchical clustering with mutual information maximization. Despite achieving stable training dynamics and successful feature organization, all variants consistently yielded 0.0 unlearning scores, suggesting inherent constraints in modifying distributed neural representations.

The technical advances in our implementation—including gradient-based importance tracking (0.9 memory factor), 32-cluster hierarchical organization, and carefully tuned importance weights (35

These findings point to three promising research directions: (1) investigating alternative architectures specifically designed for mutable knowledge representation, potentially drawing from neuroplasticity research; (2) exploring hybrid approaches that combine sparse autoencoders with external memory mechanisms; and (3) developing new optimization frameworks that move beyond traditional gradient-based methods to enable targeted feature modification while preserving model functionality. The consistent failure of increasingly sophisticated approaches suggests that achieving effective knowledge modification may require fundamentally new paradigms in neural network design.

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.

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

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.

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.