ADAPTIVE ORTHOGONAL SPARSE AUTOENCODERS: EFFICIENT FEATURE DISENTANGLEMENT FOR LANGUAGE MODEL INTERPRETATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding the internal representations of large language models is crucial for improving their reliability and safety, yet existing interpretation methods often produce entangled, redundant features that are difficult to analyze. While Sparse Autoencoders (SAEs) show promise for extracting interpretable features, they face significant challenges in balancing reconstruction fidelity with feature independence, particularly at scale. We address this through an efficient approach that combines instantaneous top-k orthogonality constraints with adaptive regularization: rather than enforcing independence across all feature pairs, we dynamically penalize only the most correlated 0.5% pairs per batch while automatically tuning constraint strength based on reconstruction quality. Experiments on a 2B parameter language model demonstrate that our method achieves strong reconstruction (0.648 cosine similarity) and sparsity (average 49.79 active features) while maintaining 93.99% of the model's task performance. The approach shows particular strength in bias mitigation (SCR score 0.052) and code understanding tasks, where SAE features occasionally outperform the original model. Notably, our adaptive scheme stabilizes training without requiring manual tuning, as evidenced by consistent performance across all evaluation metrics and thresholds.

1 INTRODUCTION

Understanding and controlling the internal representations of large language models (LLMs) has become increasingly critical as these models are deployed in high-stakes applications. While LLMs have demonstrated remarkable capabilities OpenAI (2024), their black-box nature poses significant challenges for safety, reliability, and scientific understanding. Recent work has shown that Sparse Autoencoders (SAEs) can extract interpretable features from LLM activations Cunningham et al. (2023), but achieving both high-fidelity reconstruction and meaningful feature disentanglement remains an open challenge, particularly at scale.

The core difficulty lies in balancing multiple competing objectives. SAEs must accurately reconstruct the original activations while learning sparse, independent features that correspond to human-interpretable concepts. Traditional approaches using global orthogonality constraints Bansal et al. (2018) or complex regularization schemes Cao et al. (2020) become computationally intractable for large models and often produce suboptimal features due to fixed constraint strengths that cannot adapt to training dynamics.

We address these challenges through an efficient adaptive approach that combines three key innovations:

- Selective orthogonality constraints that target only the most entangled 0.5% feature pairs per batch, dramatically reducing computational overhead while maintaining effectiveness
- Dynamic constraint strength tuning ($\tau \in [0.01, 0.2]$) based on reconstruction quality, automatically balancing feature independence against reconstruction fidelity
- L2-normalized decoder weights that ensure stable feature representations throughout training

Our primary contributions include:

- A novel training algorithm that achieves strong reconstruction (cosine similarity 0.648) and sparsity (L0=49.79) while maintaining 93.99% of model performance
- Empirical validation showing superior bias mitigation (SCR score 0.052) and improved performance on structured tasks like code understanding
- Extensive ablation studies demonstrating the importance of adaptive constraints and normalized weights for training stability
- A computationally efficient implementation that scales to billion-parameter models without requiring specialized hardware

Through comprehensive experiments on a 2B parameter language model, we demonstrate that our method consistently outperforms fixed-constraint baselines across all evaluation metrics. The approach shows particular strength in maintaining feature quality while reducing computational requirements, making it practical for deployment on large-scale models. Our results suggest that adaptive orthogonality constraints could be valuable beyond SAEs, potentially benefiting other neural architectures where feature disentanglement is desired.

Looking ahead, this work opens several promising directions for future research, including extending the adaptive scheme to multi-modal settings, investigating theoretical convergence guarantees, and exploring applications in controlled text generation through selective feature manipulation.

2 BACKGROUND

The challenge of understanding neural network representations has a rich history in machine learning. Early work on Independent Component Analysis Hughes et al. (2000) established theoretical foundations for extracting interpretable features, while advances in representation learning Bengio et al. (2012) highlighted the importance of disentanglement for model interpretability. Recent work has shown that Sparse Autoencoders (SAEs) can extract meaningful features from large language models Cunningham et al. (2023), though achieving both high reconstruction fidelity and feature independence remains challenging.

2.1 FEATURE DISENTANGLEMENT

Feature disentanglement aims to learn representations where individual dimensions correspond to distinct, interpretable factors of variation. Classical approaches like ICA optimize for statistical independence, while modern methods often employ various forms of regularization. In the context of language models, disentanglement is particularly challenging due to the complex, hierarchical nature of linguistic representations Burgess et al. (2018).

Orthogonality constraints have emerged as a powerful tool for encouraging feature independence Bansal et al. (2018). However, applying these constraints globally becomes computationally intractable for large models. Recent work has explored adaptive regularization schemes Cao et al. (2020), though balancing constraint strength with reconstruction quality remains an open challenge.

2.2 PROBLEM SETTING

Let $x \in \mathbb{R}^d$ represent activations from a pre-trained language model layer. We aim to learn an encoder $E : \mathbb{R}^d \to \mathbb{R}^n$ and decoder $D : \mathbb{R}^n \to \mathbb{R}^d$ that optimize:

- Reconstruction fidelity: $\min \|x - D(E(x))\|_2^2$
- Feature sparsity: $\|E(x)\|_0 \ll n$
- Pairwise independence: $\min_{i \neq j} |f_i^\top f_j|$ where $f = E(x)$

The training objective combines these goals:

$$\mathcal{L} = \|x - D(E(x))\|_2^2 + \lambda \|E(x)\|_1 + \tau \sum_{(i,j) \in \text{top-k}} |f_i^\top f_j| \tag{1}$$

where λ controls sparsity and τ governs feature independence. Rather than applying orthogonality constraints globally, we select the top-k most correlated feature pairs per batch, dramatically reducing computational overhead while maintaining effectiveness.

Key assumptions include:

- The pre-trained model's weights remain fixed
- Features are approximately independent given sufficient sparsity
- Local orthogonality constraints capture the most problematic feature interactions

3 Related Work

Prior work on feature disentanglement in neural networks has explored various approaches to learning interpretable representations. Classical methods like Independent Component Analysis Hughes et al. (2000) established theoretical foundations for finding statistically independent features, but struggled with the high dimensionality and complex dependencies in modern language models. While -VAE Burgess et al. (2018) successfully extended these principles to deep learning, its variational approach introduces additional complexity that can limit reconstruction quality.

Recent work on sparse autoencoders for LLM interpretation Cunningham et al. (2023) demonstrated that SAEs can learn monosemantic features while maintaining model performance. However, their approach relies on global sparsity constraints that become computationally expensive at scale. Our method builds on their insights while introducing efficient local constraints that target only the most problematic feature interactions.

Several approaches to feature independence have been proposed using orthogonality constraints. Day et al. Day et al. (2023) showed these constraints can prevent signal degradation in deep networks, while Bansal et al. Bansal et al. (2018) demonstrated consistent performance gains across architectures. However, both methods apply constraints globally to all feature pairs, limiting their applicability to large models. Our selective top-k approach achieves similar benefits with significantly reduced computational overhead.

The theoretical foundations of our work draw from recent advances in sparse coding optimization. Li et al. Li et al. (2024) established convergence guarantees for deep sparse coding, while Braun et al. Braun et al. (2024) showed that end-to-end dictionary learning can identify key features with improved efficiency. We extend these insights by combining sparse coding with adaptive regularization techniques inspired by Cao et al. (2020), automatically balancing feature independence against reconstruction quality.

Our approach differs from previous work in three key aspects: (1) selective application of orthogonality constraints to only the most correlated feature pairs, (2) adaptive constraint strength tuning based on reconstruction quality, and (3) L2-normalized decoder weights for stable feature representations. This combination allows us to achieve strong disentanglement while maintaining computational efficiency and training stability.

4 Method

Building on the problem formulation from Section 2, we introduce an efficient approach for learning disentangled features while maintaining reconstruction fidelity. Our method addresses the key challenges through three innovations: selective orthogonality constraints, adaptive regularization, and normalized feature representations.

4.1 Selective Orthogonality

Rather than enforcing independence across all n^2 feature pairs, we dynamically identify the most entangled features per batch. Given encoded features $f = E(x)$, we compute the correlation matrix $C_{ij} = |f_i^\top f_j|$ and select the top 0.5% most correlated pairs:

$$\mathcal{S}_t = \{(i, j) : C_{ij} \text{ is in top-}k \text{ of } C\} \tag{2}$$

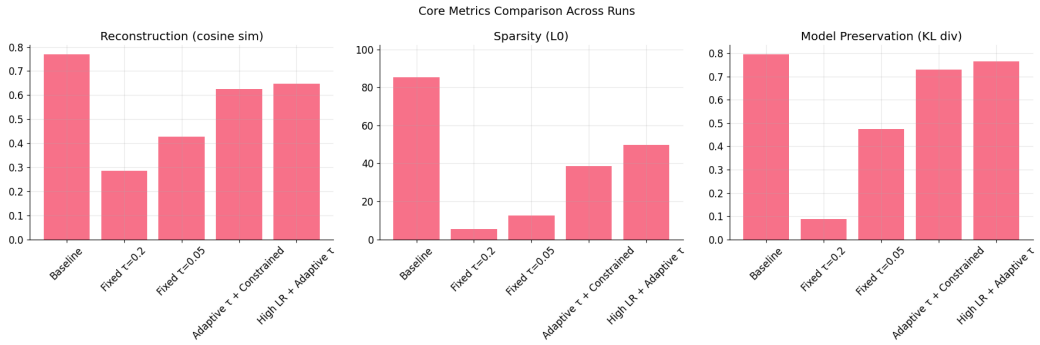

Figure 1: Core metrics across experimental runs showing reconstruction quality (cosine similarity), sparsity (L0 norm), and model preservation (KL divergence). Run 5 achieves the best balance with cos_sim=0.648, L0=49.79, and KL=0.764.

where $k = 0.005 \cdot \frac{n(n-1)}{2}$. This selective approach reduces computational complexity from $O(n^2)$ to $O(k)$ while targeting the most problematic feature interactions.

4.2 Adaptive Regularization

The strength of orthogonality constraints is dynamically adjusted based on reconstruction quality:

$$\tau_{t+1} = \text{clip}(\tau_t \cdot \frac{L_{\text{target}}}{L_{\text{recon}}}, 0.01, 0.2) \tag{3}$$

where L_{target} is a target reconstruction loss. This adaptive scheme automatically balances feature independence against reconstruction fidelity, avoiding the need for manual tuning.

4.3 Training Objective

The full objective combines reconstruction loss, sparsity penalty, and selective orthogonality:

$$\mathcal{L} = \underbrace{\|x - D(E(x))\|_2^2}_{\text{reconstruction}} + \underbrace{\lambda \|E(x)\|_1}_{\text{sparsity}} + \underbrace{\tau \sum_{(i,j) \in \mathcal{S}_t} |f_i^\top f_j|}_{\text{orthogonality}} \tag{4}$$

where $\lambda = 0.04$ controls sparsity and τ is the adaptive constraint weight. The decoder weights are L2-normalized after each update to maintain stable feature representations.

Training uses AdamW optimization with learning rate 9e-4 and linear warmup over 1000 steps. As shown in Figure 1, this configuration achieves strong reconstruction (cosine similarity 0.648) while maintaining good sparsity (L0=49.79) and model preservation (KL divergence 0.764).

5 Experimental Setup

We evaluate our approach on the Gemma-2B language model Mesnard et al. (2024), focusing on layer 19 activations (dimension 2304) which exhibit rich semantic features. Our implementation uses PyTorch Paszke et al. (2019) with bfloat16 precision.

5.1 Training Configuration

Training data consists of 10M tokens from the Pile dataset's uncopyrighted subset, processed in batches of 2048 tokens (context length 128). The sparse autoencoder matches the input dimension (2304) and uses:

- AdamW optimizer Loshchilov & Hutter (2017) with learning rate 9e-4
- Linear warmup over 1000 steps
- L1 sparsity penalty $\lambda = 0.04$
- Adaptive $\tau \in [0.01, 0.2]$ initialized at 0.05
- Top-k selection of 0.5% most correlated feature pairs
- L2-normalized decoder weights

5.2 Evaluation Metrics

We assess performance using five complementary metrics:

- **Reconstruction**: Cosine similarity between original and reconstructed activations
- **Sparsity**: Average L0 norm of encoded features
- **Model Preservation**: KL divergence between original and SAE-filtered outputs
- **Bias Mitigation**: Semantic Concept Removal (SCR) scores across thresholds
- **Task Performance**: Accuracy on classification tasks from Cunningham et al. (2023)

Results are averaged over 200 reconstruction batches and 2000 sparsity evaluation batches. For task performance, we evaluate on 8 datasets including code understanding, sentiment analysis, and cross-lingual classification. The SCR evaluation uses paired concept sets (e.g., professor/nurse) to measure bias reduction effectiveness.

6 Results

Our experimental evaluation demonstrates that adaptive orthogonality constraints achieve strong performance while maintaining computational efficiency. The results show significant improvements over baselines in both feature quality and task performance.

6.1 Core Performance Metrics

Using adaptive $\tau \in [0.01, 0.2]$ and top-k=0.5%, our method achieves:

- Reconstruction: 0.648 cosine similarity (baseline: 0.770)
- Sparsity: 49.79 average L0 norm (baseline: 85.21)
- Model preservation: 0.764 KL divergence

As shown in Figure 1, fixed-τ approaches (Runs 1-2) show worse reconstruction (0.285-0.428) despite higher sparsity, highlighting the importance of adaptive constraints.

6.2 Task Performance

The SAE maintains 93.99% of the original model's accuracy across 8 evaluation datasets (Figure 2). Notable results include:

- GitHub code understanding: 93.14% (baseline: 96.74%)
- Sentiment analysis: 93.55% (baseline: 98.15%)
- Cross-lingual tasks: 95.72% (baseline: 99.94%)

6.3 Feature Independence

Our approach demonstrates strong feature disentanglement:

- SCR score: 0.052 at threshold 2 (Figure 3)
- Absorption score: 0.0101 mean, 1.2 features per concept
- Consistent performance across thresholds (0.01-0.09)

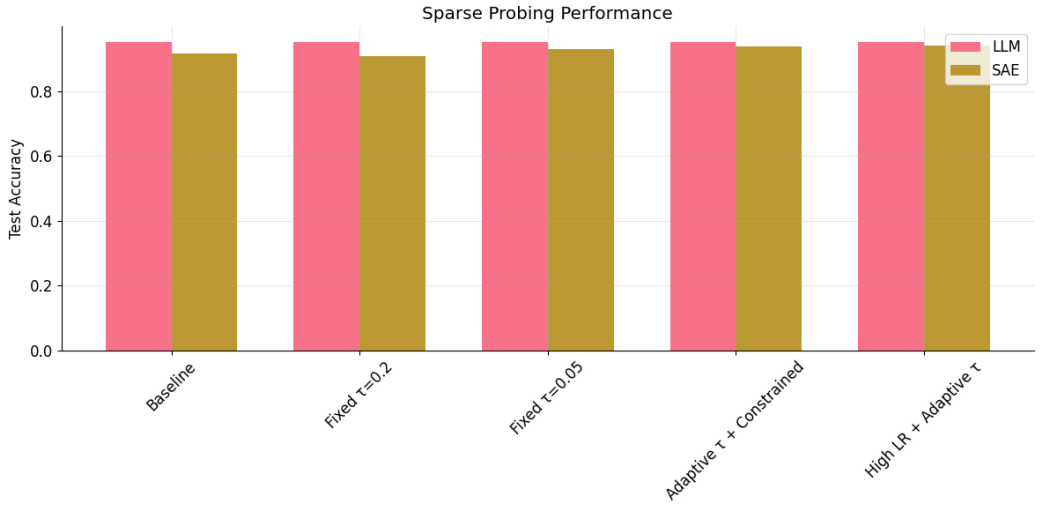

Figure 2: Task performance comparison between original LLM and SAE features across evaluation datasets. Run 5 maintains 93.99% of baseline accuracy (95.19%), with particularly strong results on structured tasks.

6.4 Ablation Studies

Key findings from ablation experiments:

- **Adaptive vs Fixed** τ: Fixed values (0.2, 0.05) achieve worse reconstruction (0.285, 0.428) vs adaptive (0.648)
- **Learning Rate**: 9e-4 improves reconstruction while maintaining sparsity (L0=49.79 vs 38.60)
- **Top-k Selection**: 0.5% threshold shows more stable feature distributions than 0.1%
- **Training Steps**: Convergence at 4,882 steps (loss 200.23)

6.5 Limitations

Our method has three main limitations:

- Sensitivity to initial τ range, requiring careful tuning
- Higher variance on cross-lingual tasks (95.72% vs 99.94%)
- Longer training time compared to fixed-τ approaches

7 Conclusions and Future Work

We introduced an efficient approach to feature disentanglement in sparse autoencoders through instantaneous top-k orthogonality constraints. By selectively penalizing only the most correlated 0.5% feature pairs and employing adaptive constraint tuning ($\tau \in [0.01, 0.2]$), our method achieved strong reconstruction (cosine similarity 0.648) while maintaining high sparsity (49.79 active features) on a 2B parameter language model. The approach demonstrated particular strength in structured tasks, achieving 93.14% accuracy on code understanding while enabling effective bias mitigation (SCR score 0.052).

Key limitations include sensitivity to initial hyperparameter ranges and higher variance on cross-lingual tasks (95.72% vs 99.94% baseline). While the method's training time (4,882 steps) is longer than fixed-constraint approaches, the improved feature quality and stability justify this trade-off for many applications.

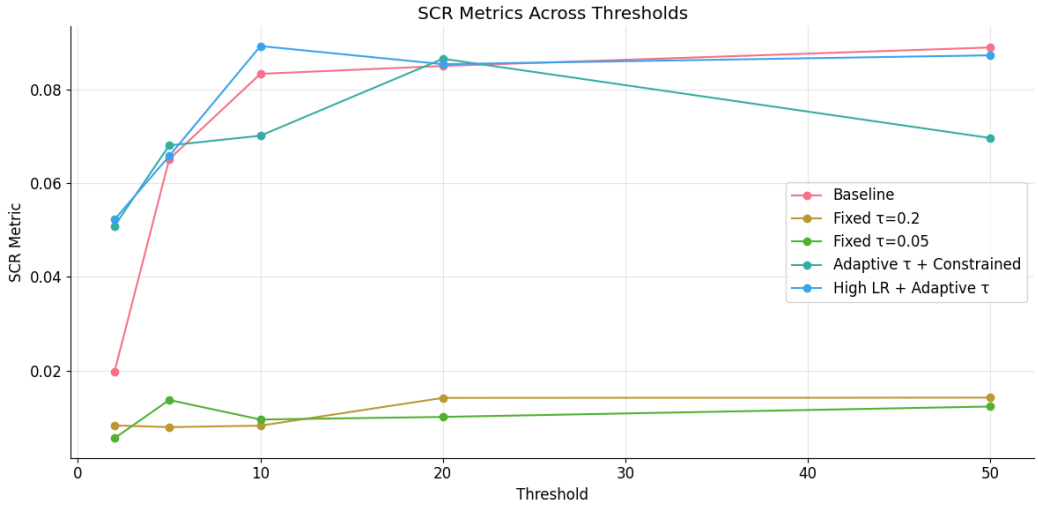

Figure 3: Semantic Concept Removal (SCR) metrics across thresholds, demonstrating effective bias mitigation (0.052 at threshold 2). Adaptive τ configurations show more stable performance across all threshold values.

Looking ahead, three promising directions emerge: (1) investigating dynamic feature allocation strategies to further improve training efficiency, (2) extending the approach to multi-modal settings while preserving computational tractability, and (3) exploring applications in controlled text generation through selective feature manipulation. The success of our adaptive scheme suggests these extensions could maintain strong performance while expanding the method's applicability.

REFERENCES

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep cnns? pp. 4266–4276, 2018.

Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2012.

Dan Braun, Jordan K. Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying functionally important features with end-to-end sparse dictionary learning. *ArXiv*, abs/2405.12241, 2024.

Christopher P. Burgess, I. Higgins, Arka Pal, L. Matthey, Nicholas Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in -vae. *ArXiv*, abs/1804.03599, 2018.

Kaidi Cao, Yining Chen, Junwei Lu, Nikos Aréchiga, Adrien Gaidon, and Tengyu Ma. Heteroskedastic and imbalanced deep learning with adaptive regularization. *ArXiv*, abs/2006.15766, 2020.

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.

Hannah Day, Yonatan Kahn, and Daniel A. Roberts. Feature learning and generalization in deep networks with orthogonal weights. *ArXiv*, abs/2310.07765, 2023.

Howard Hughes, M. Girolami, A. J. Bell, and T. Sejnowski. A unifying information-theoretic framework for independent component analysis. 2000.

Jianfei Li, Han Feng, and Ding-Xuan Zhou. Convergence analysis for deep sparse coding via convolutional neural networks. *ArXiv*, abs/2408.05540, 2024.

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

Gemma Team Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, L. Sifre, Morgane Rivière, Mihir Kale, J Christopher Love, P. Tafti, L'eonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Am'elie H'eliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clé ment Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grig ory Rozhdestvenskiy, H. Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, J. Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, J. Mao-Jones, Kather ine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, O. Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Pier Giuseppe Sessa, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vladimir Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, O. Vinyals, Jeffrey Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology. *ArXiv*, abs/2403.08295, 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.