

Superposition, but at a Cross-MLP Layers view? *

Ng Woon Yee
Nanyang Technological University
Singapore

With
In collaboration with Apart Research and Goodfire

Abstract

To understand feature causation across multilayer perceptron (MLP) layers, this study introduces the Coordinated Sparse Autoencoder Network (CoSAEN) by integrating sparse autoencoder and causal discovery algorithm: PC algorithm. Using contrasting educational and adult-oriented datasets, the research applies CoSAEN to a GPT-2 Medium model, revealing how early-layer feature can simultaneously encode an abstract representation that influence multiple specialized features in later layers. The research demonstrates feature-level superposition, notably identifying features like L15_F49 that causally connected to both educational and adult-oriented features in another layers (L22). By providing insight into how information flows across neural network layers, the study also opens opportunities on exploring path-based activation steering as a potential method for more precise model interpretability and intervention.

Keywords:

Mechanistic Interpretability Track: sparse autoencoders, feature visualization, model steering, interpretability tools, neural features, activation patterns, model internals, feature attribution

1. Introduction

1.1. Problem Statement

In recent years, researchers have been using Dictionary Learning to address the polysemantic behavior of Multilayer Perceptron Neurons. Methods such as Sparse Autoencoder (SAE) are commonly used to extract interpretable features in Large Language Models (LLMs). (Elhage et al., 2022) While most researchers apply SAE on a single layer of MLP to eventually steer the model output with granular control (Huben et al., 2023), some groups have attempted to train SAE across all layers in hopes of understanding the hierarchical residual stream (Lawson et al., 2024). However, neither a single SAE across all layers nor a single SAE on one layer helps in understanding the causality in cross-layer residual streams. This is because residual streams are hierarchical; early layers might learn abstractions of features, and the later layers will learn more granular features. In this project, we aim to address this issue by using Causal Intervention to capture the path-based inter-layer relationships using multiple SAEs.

*Research conducted at the Women in AI Safety Hackathon, 2025

2. Methods

2.1. Approach

We propose a Coordinated Sparse Autoencoder Network (CoSAEN), a framework that consists of layer-specific SAEs connected through a causal intervention framework. We implemented our CoSAEN framework using Pytorch and the TransformerLens library. In this project, we will be focusing on 2 layers of MLP and SAEs trained specifically in each layers. Then, we deployed causal discovery method: PC algorithm, a constraint based causal discovery method, to output a directed acyclic graph (DAG) that represents the causal relationship between the extracted features across layers.

The aim of this approach is to show that superposition occurs not only at neuron level, it also happens at feature levels where feature extracted from earlier layers will serve as an abstraction for features extracted in later layers. By analyzing these cross-layer causal connections, we can identify how information flows through the network and how higher-level concepts are built upon earlier representations in the model's processing hierarchy.

2.2. Implementation

In our implementation, we used two contrasting and yet similar dataset. First dataset consists of educational texts with the word "teacher" always mentioned. The second dataset contained adult-oriented contexts where the same term "teacher" was employed in role-playing scenarios. This controlled contrast allow us to examine how the word "teacher" with different semantic meanings are encoded across neural network. In our implementation, we also expect to find a feature in the earlier layers that is serving as an abstraction on the word "teacher", causally influence two features extracted in the later layers, each of which represent the semantic "educational" and "adult-oriented"

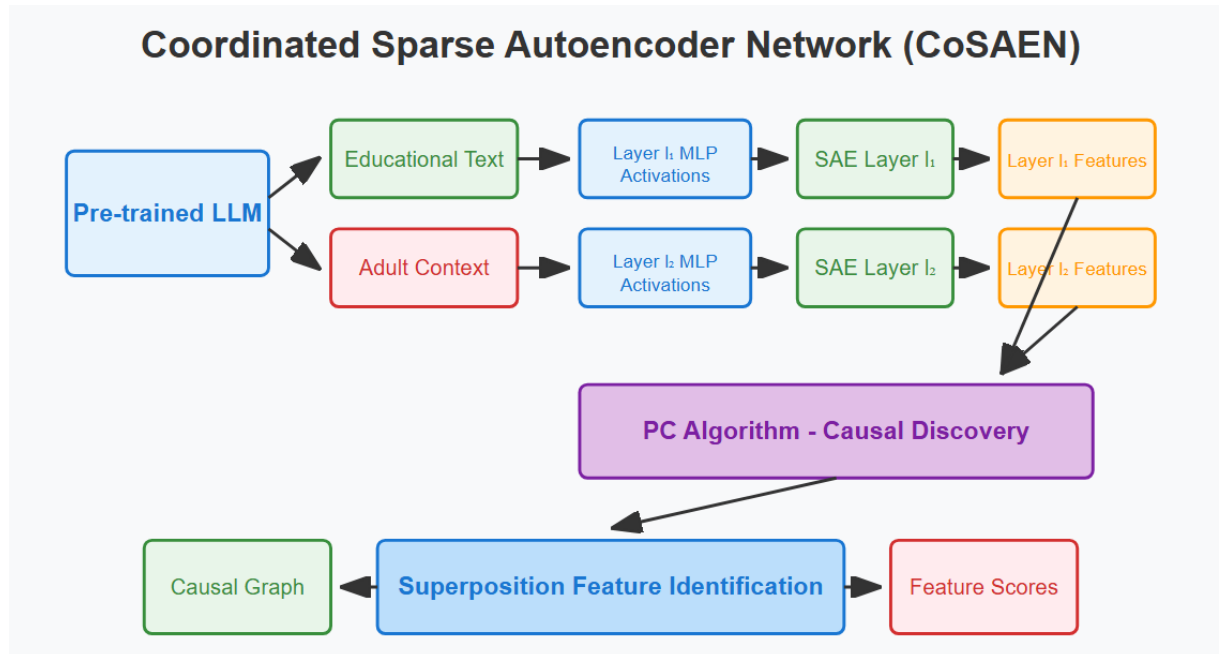


Figure 1: Starting with a pre-trained language model processing both educational and adult contexts, CoSEAN extracts MLP activations from two distinct layers. These activations are passed through layer-specific sparse autoencoders to obtain interpretable latent features. The PC Algorithm then discovers causal relationships between features across layers, identifying which early-layer features causally influence specialized features in later layers.

3. Results

3.1. Analysis and Findings

In our analysis, we indeed found strong evidence of feature-level superposition across layers in the GPT-2 Medium model. Particularly notable were features L15_F49 which demonstrated significant causal connections to both educational features (e.g., L22_F3) and adult-specific features (e.g., L22_F38). These connections, highlighted by asterisks in Figure 1, provide compelling evidence that these early-layer features encode polysemantic features representation of the word "teacher" and later represented by context-specific features in the later layer (eg: L22)

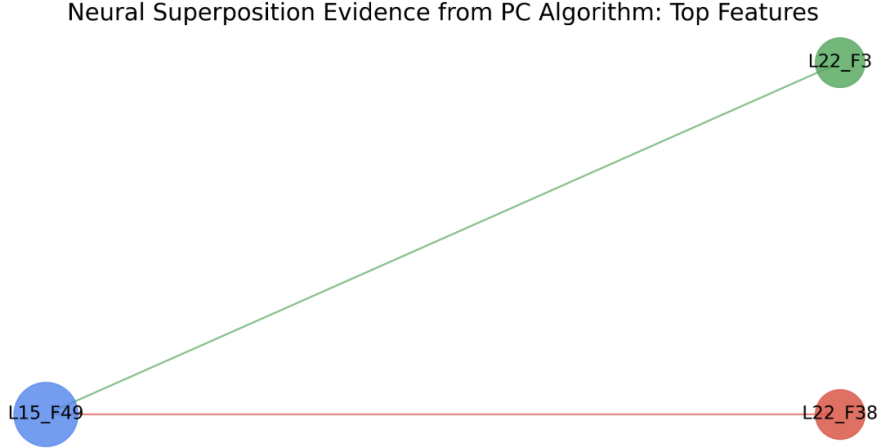


Figure 2: This network diagram illustrates neural superposition in a language model, showing a Layer 15 feature (L15_F49, blue) that simultaneously connects to both educational content (L22_F3, green) and adult content features (L22_F38, red) in Layer 22. The directed edges indicate causal relationships identified by the PC algorithm, demonstrating how a single feature in an earlier layer can encode and influence multiple types of specialized information in later layers.

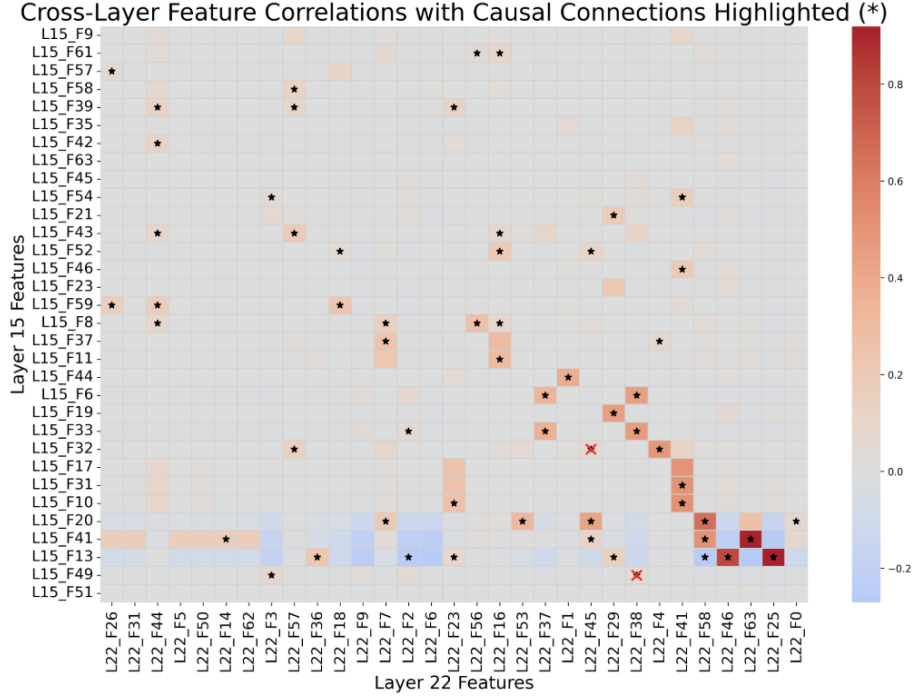


Figure 3: This heatmap visualizes correlation coefficients between neural network features in Layer 15 (y-axis) and Layer 22 (x-axis), with colors ranging from blue (negative) to red (positive). Black asterisks (*) mark statistically significant causal connections identified by the PC algorithm, while red "x" markers highlight superposition—features that simultaneously influence both educational and adult-specific features. Notable features like L15_F13 and L15_F41 show strong correlations with multiple Layer 22 features, revealing how information transforms as it flows through the neural network, with earlier layers causally influencing the emergence of more specialized features in later layers.

3.2. Impact Assessment

Our findings have positive implications for AI Safety and interpretability efforts. By identifying cross-layer feature superposition and causal relationship, we provide a framework to understanding how language models represent and process content. With path-based SAE exploration, we can also touch on path based activation steering.

4. Discussion and Conclusion

Our research demonstrates feature-level superposition across neural network layers, revealing how abstract representations in earlier layers causally influence specialized features in later layers (Olah et al., 2020; Goh et al., 2021). The CoSAEN framework provides both an auditing methodology for sensitive content representation and potential intervention points for targeted safety controls, enabling path-based feature steering.

5. References

- Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3), 10.23915/distill.00030. <https://doi.org/10.23915/distill.00030>
- Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, 5(3), 10.23915/distill.00024.001. <https://doi.org/10.23915/distill.00024.001>

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., et al. (2022). Toy models of superposition. arXiv preprint arXiv:2209.10652.

Huben, R., Cunningham, H., Smith, L. R., Ewart, A., & Sharkey, L. (2023). Sparse autoencoders find highly interpretable features in language models. In The Twelfth International Conference on Learning Representations.

Lawson, T., Farnik, L., Houghton, C., & Aitchison, L. (2024). Residual Stream Analysis with Multi-Layer SAEs. arXiv preprint arXiv:2409.04185.