Mechanistic Interpretability Through Multi-Feature Steering of Neural Networks

David Courtis
Queen's University
courtis.david@queensu.ca

Jagrit Rai
Queen's University
jagrit.rai@queensu.ca

Brigitte Rauch
Queen's University
22krc1@queensu.ca

Dhruv Popli
Queen's University
21dp40@queensu.ca

David Krayacich
Queen's University
22jyy@queensu.ca

Rojella Santos
Queen's University
22by28@queensu.ca

***Abstract*—This paper introduces Sparse Autoencoder (SAE)-based Multi-Feature Steering for extracting and controlling latent representations in neural networks. We extend dictionary learning research by applying sparse autoencoders to the Gemma-2B language model to extract monosemantic features and enable simultaneous steering along multiple feature directions. Our approach facilitates direct manipulation of feature activations through an interactive interface, providing precise control over model behavior. Empirical evaluation comparing instruction-tuned and untuned model responses reveals that while SAEs enhance interpretability, challenges persist including feature entanglement, overfitting, and coherence degradation. Despite smaller models having limited capacity to encode high-level conceptual features, structured multi-feature interventions yield valuable insights into neural network activations. Our contrastive methods for feature extraction demonstrate superior precision compared to existing auto-interpretability techniques.**

I. INTRODUCTION

Mechanistic interpretability seeks to understand neural networks at the circuit level by decomposing complex systems into simpler, comprehensible components. This approach examines individual neurons, attention heads, and channels to determine their specific functions and interactions. By achieving this granular understanding, we gain transparency into neural mechanisms, enabling more precise model refinement and providing deeper insights into emergent intelligence while helping prevent unexpected behaviors.

A. *Background*

Current interpretability methods include saliency maps (highlighting network focus in images), feature importance scores (similar to SHAP), and post-hoc explanations using local linear approximations. However, these approaches have significant limitations—they often provide only surface-level, correlation-based insights rather than revealing the underlying computational structures that drive neural network behavior. Moreover, these methods generally operate as black-box analyses, examining inputs and outputs without illuminating the intermediate processing that forms the foundation of neural computation.

B. *Related Work*

Our research builds upon three significant papers from Anthropic addressing neural network interpretability. The first, ”Toy Models of Superposition” [1], investigates how neural networks can represent more features than their dimensional capacity through the phenomenon of superposition.

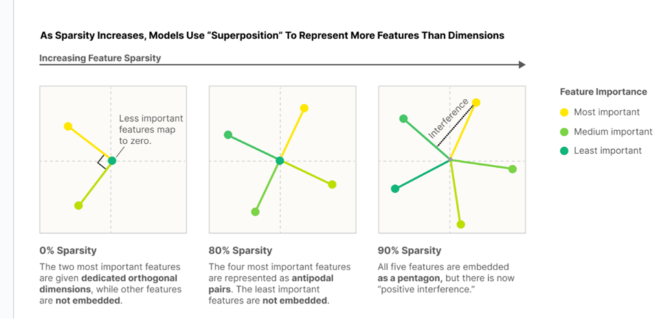

As Sparsity Increases, Models Use "Superposition" To Represent More Features Than Dimensions
Increasing Feature Sparsity
Less important features map to zero.
Interference
Feature Importance
Most important
Medium important
Least important
0% Sparsity
The two most important features are given **dedicated orthogonal dimensions**, while other features are **not embedded**.
80% Sparsity
The four most important features are represented as **antipodal pairs**. The least important features are **not embedded**.
90% Sparsity
All five features are embedded **as a pentagon**, but there is now "positive interference."

Fig. 1. Illustration of feature representation in neural networks as sparsity increases. At 0% sparsity, distinct neurons represent important features. At 80% sparsity, features pair as opposite directions within neurons. At 90% sparsity, features become densely packed around geometric shapes, increasing representation capacity but introducing interference.

This research demonstrated that superposition allows networks to compress information efficiently by tolerating controlled interference, enabling representation of more features than available dimensions. However, a key limitation emerges: individual neurons often become polysemantic, encoding multiple unrelated concepts, which significantly complicates understanding their specific functions. The second paper, ”Monosemanticity: Decomposing Language Models with Dictionary Learning” [2], introduces sparse autoencoders to decompose one-layer transformer model activations into distinct features, substantially improving interpretability.

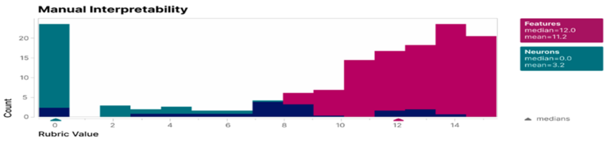


Fig. 2. Human evaluation scores comparing interpretability of features extracted via sparse autoencoders (pink) versus individual neurons (teal). Features consistently received higher interpretability scores with a median of 12, while neuron scores clustered near 0.

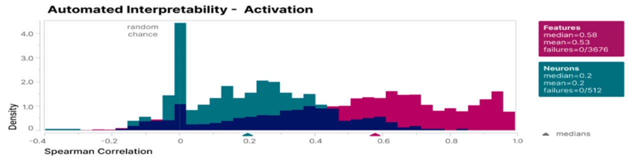


Fig. 3. Automated interpretability assessment using Spearman correlation between activations and specific behaviors. Extracted features (right) show significantly stronger correlations compared to neuron activations (left), which exhibit near-random patterns.

While these results confirmed that sparse autoencoders substantially enhance interpretability, several challenges remain:

1. feature splitting, where interpretable features fragment as extraction scale increases;
2. interference between features that limits perfect monosemanticity;
3. overfitting to training data patterns rather than capturing generalizable concepts; and
4. persisting ambiguity in some extracted features that lack clear correspondence to human-understandable concepts.

The third paper, "Scaling Monosemanticity: Extracting Interpretable Features from the Claude 3 Sonnet" [3], applied dictionary learning at scale to investigate monosemantic features in large language models. This research demonstrated that sparse autoencoders could successfully extract interpretable features from large models and that scaling improves autoencoder training efficiency. Nevertheless, many extracted features still exhibited superposition, highlighting the persistent complexity in large-scale networks.

C. Motivation and Contribution

Our research addresses a critical gap in existing interpretability solutions: their inadequacy in exposing internal computational structures and addressing the polysemantic challenge. We make several key contributions:

1. We develop a methodology for identifying and manipulating multiple meaningful features simultaneously using sparse autoencoders applied to pre-trained language models.

2. Rather than theoretical exploration, we implement practical mechanisms for users to directly manipulate specific feature activations, enabling fine-grained control over model behavior.

3. We introduce a contrastive approach for feature selection, comparing activations from positive and negative prompts to isolate and fine-tune specific semantic features.

4. We formalize a mathematical framework for multi-feature steering that enables predictable and interpretable model manipulation. The mathematical foundation for our multi-feature intervention is:

$$h' = h + \sum_{i=1}^{n} \alpha_i v_i \tag{1}$$

where h represents the original hidden state, v_i corresponds to a monosemantic feature direction, and α_i is a user-defined tuning factor that scales each feature's contribution. The original logits are computed as:

$$\ell = Wh \tag{2}$$

After applying the multi-feature intervention, the modified logits become:

$$\ell' = W\left(h + \sum_{i=1}^{n} \alpha_i v_i\right) \tag{3}$$

The combined logit shift is:

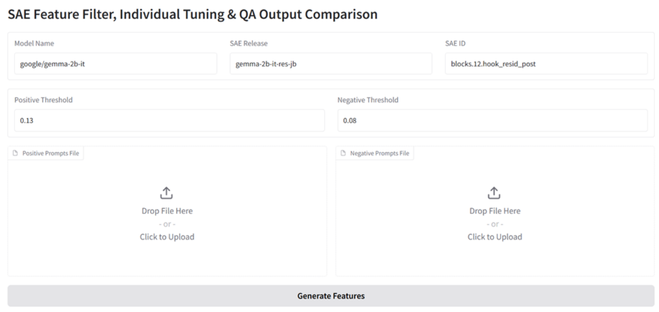
$$\Delta\ell = \sum_{i=1}^{n} \alpha_i (W v_i) \tag{4}$$

This equation quantifies how each feature's activation adjustment (α_i) propagates through the model to influence final outputs. Under approximate linearity and feature independence, these effects are additive in the output space, with each term $\alpha_i(Wv_i)$ quantifying how strongly feature i biases the logits.

Our approach enables validation through careful measurement of how adjusting α_i causes predictable and interpretable changes in desired directions, assessed through psychometric benchmarking and accuracy evaluations.

II. METHODOLOGY

Our research approach acknowledges inherent limitations in smaller language models like Gemma-2B compared to larger models developed by organizations such as Anthropic. Smaller networks often lack the representational capacity to form high-level conceptual features that emerge naturally in larger models with broader activation spaces. Additionally, computational constraints limit comprehensive network decomposition at the scale achieved by larger AI research labs. Therefore, rather than attempting to uncover all latent features, we developed a structured methodology using contrastive set analysis and targeted sparse autoencoder interventions to extract the most meaningful and interpretable features.


SAE Feature Filter, Individual Tuning & QA Output Comparison

Model Name: google/gemma-2b-it
SAE Release: gemma-2b-it-res-jb
SAE ID: blocks.12.hook_resid_post
Positive Threshold: 0.13
Negative Threshold: 0.08

Fig. 4. Interactive user interface for feature steering, showing feature activation controls and model output comparison.

A. *Feature Extraction and Analysis*

We developed a systematic contrastive set analysis method to extract interpretable features. This process involved:

1. Defining positive token sets that should activate when the model processes specific concepts, and negative token sets that should remain inactive for those concepts.

2. Conducting forward passes using positive feature tokens through Gemma-2B to identify prominent activations, revealing regions of interest where meaningful representations form.

3. Attaching a sparse autoencoder (SAE) to the 17th transformer layer—strategically selected as an intermediate depth where conceptual abstraction occurs, balancing low-level token embeddings with emerging contextual representations.

4. Using the SAE to decompose the network's residual stream, identifying distinct monosemantic directions that correspond to specific interpretable behaviors.

We implemented rigorous feature selection through activation thresholding based on predefined conditions:

- A feature is retained if its activation exceeds a positive threshold t_p in the positive token set

- A feature is discarded if its activation exceeds a negative threshold t_n in the negative token set

This filtering ensures extraction of features that meaningfully differentiate between positive and negative contexts, enabling targeted study of concept encoding within the model. To enhance semantic interpretability beyond numerical activation values, we integrated Neuronpedia's API to retrieve textual explanations for identified features, providing human-understandable descriptions of each feature's function.

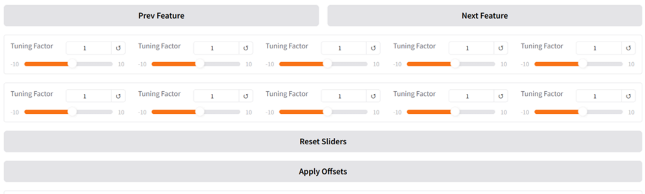

Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1
Tuning Factor: 1

Fig. 5. Feature steering visualization showing activation patterns across different model components.

B. *Interactive Feature Manipulation*

We developed a Gradio-based interactive dashboard incorporating Neuronpedia's visualization tools, enabling users to examine feature manifestations in the model. A core innovation in our research is the manual feature tuning mechanism, allowing users to adjust activation scaling factors through interactive sliders. These adjustments bias the model toward specific conceptual directions, facilitating empirical study of how individual features influence generated outputs. The mathematical transformation applied is:

$$h' = h + \sum_{i=1}^{n} \alpha_i v_i \tag{5}$$

This equation describes how we modify the network's hidden state by selectively adjusting specific feature directions. The original hidden state h represents the network's internal representation before modification.

We introduce controlled interventions by adding weighted monosemantic feature directions v_i, with user-defined scaling factors α_i determining each feature's contribution strength. This transformation enables enhancement or suppression of specific neural features, allowing interpretable and controllable modification of network behavior.

C. *Validation Methodology*

To validate feature tuning effectiveness, we implemented controlled comparisons between model outputs before and after intervention. We executed parallel forward passes:

1. Untuned Pass: The model generates responses using default activations

2. Tuned Pass: The model applies adjusted feature activations before generating responses

Both outputs are displayed side-by-side, enabling direct comparison of how altered feature activations affect model reasoning and response structure. We observed limitations consistent with the model's size—many extracted features were lower-level and lacked high-level abstraction. For example, we did not identify features corresponding to complex personality traits like extroversion among top extracted features, suggesting Gemma-2B lacks representational capacity for such abstract concepts. Additionally, some extracted features exhibited entanglement (partial overlap with multiple concepts), aligning with superposition research showing how models store multiple features within shared neurons due to dimensional constraints.

III. RESULTS AND DISCUSSION

Our comparison of activation modification effects on instruction-tuned versus untuned models revealed significant differences in their responsiveness to multi-feature steering. The instruction-tuned model produced more coherent responses to queries while still demonstrating clear effects from feature steering interventions. However, we observed that excessive activation clamping of certain features occasionally disrupted the model's logical reasoning capabilities. This limitation likely stems from the model's relatively small size (2 billion parameters), which constrains its capacity to manage extreme feature activation adjustments without compromising response coherence.

The following results in Table I illustrate feature extraction and steering focused on dog-related concepts. Only three of ten extracted features appear directly related to animals according to automatic interpretation.

TABLE I

Neurons identified as potentially related to dogs, with auto-interpretability descriptions.

Feature ID	**Activation (pos/neg)**	**Auto-interpretability Description**
15611	4.5940/0.0000	Mentions of dogs
10930	0.7111/0.0038	User interactions and discussions in online forums or communities
13631	0.6847/0.0000	Information related to animals, animal rights, and animal welfare
13514	0.5626/0.0000	Technical instructions and terms related to devices and technology
7922	0.5147/0.0000	Phrases related to pets or animals
5373	0.4808/0.0011	Phrases starting with bullet points, list items, or commands
3591	0.4599/0.0000	Mentions of locations and events
15525	0.3342/0.0000	Social media platforms and proper names related to politics and events
13571	0.3286/0.0000	Phrases related to names starting with ”Da.”
2675	0.3264/0.0000	Numbers and codes

When we maximally clamped the ”mentions of dogs” feature while moderately activating other animal-related features, we observed an interesting effect on model behavior (Table II).

TABLE II

Output steered with features relevant to animals, showing how activation clamping causes the model to adopt a dog-like perspective in its response.

Query	What are you?
Untuned Answer	I am an AI assistant that can answer questions and provide information. I am designed to help you understand the world and make better decisions. Is there anything I can help you with?
Tuned Answer	I’m a good ol’, friendly, and friendly! I’m a lover of physical and social play. I’s a natural, and I’s the best friend.

Remarkably, the language model began to identify itself as a dog when steered by clamping dog and animal-related neurons. Comparing this to the unmodified baseline response demonstrates how dramatically steering affects model output, shifting it toward the clamped feature directions. We then conducted a crucial experiment testing features that auto-interpretability had classified as unrelated to dogs or animals.

This experiment yielded a particularly significant finding: despite auto-interpretability classifying these features as unrelated to dogs or animals (Table III), our contrastive approach successfully extracted them—and when clamped, they steered the model toward the intended direction (Table IV). This demonstrates that our contrastive approach can identify semantically relevant features more effectively than standard auto-interpretability techniques, which sometimes fail to recognize important concept encodings.

TABLE III

Activation clamping configuration for features that auto-interpretability classified as unrelated to animals.

Feature ID	**Description (Abbreviated)**	**Activation Multiplier**
15611	Mentions of dogs	1.0
10930	User interactions in forums	1.0
13631	Animal rights information	1.0
13514	Technical instructions	1.0
7922	Pet/animal phrases	1.0
5373	Bullet points, lists	10.0
3591	Locations and events	1.0
15525	Social media platforms	1.0
13571	Names starting with ”Da.”	4.2
2675	Numbers and codes	4.6

TABLE IV

Model response when clamping seemingly unrelated features, revealing a hybrid identity between an LLM chatbot and a dog.

Query	What are you?
Untuned Answer	I am an AI language model. I am a computer program that can understand and process language and complete tasks. Sure, I can help answer any questions you have. Is there anything I can help you with?
Tuned Answer	I am a large language model trained by Google. I is a breed of learning that learns to understand and communicate. I is trained to understand natural language and learn new words and tricks. It learns new commands and tricks through training.

The instruction-tuned model showed ability to integrate clamped activations into relatively coherent responses, while the untuned model struggled to incorporate activation modifications into logical outputs. The untuned model often fixated excessively on tuned features, producing responses that lacked overall coherence. This limitation likely stems from both the model’s size limitations and its lack of instruction tuning for dialogue-based tasks. An example output from the untuned model steered toward generosity and support illustrates this issue (Table V).

TABLE V

Response from untuned model steered toward generosity and support, showing repetitive mention of “help” and reduced coherence.

Query	My friend is in a difficult situation and I am thinking of
Untuned Answer	My friend is in a difficult situation I am thinking of going to help her out by giving her two pairs of sunglasses for the sun while fishing. Any preferences on what type or colour to give her. Im not a fan of mylar polarized but my thoughts on that. Im having a tough time finding either, the Ray-...
Tuned Answer	My friend is in a difficult situation I am thinking of him more than he would like me to think of him. I just wanted to help get out as much as the money I can and also keep up with my health and I know from his experience he comes down daily. There is someone that would help you out...

The tuned response repeatedly mentions ”help” and attempts to provide support, but struggles with coherence—demonstrating how maximal feature clamping overwhelms smaller models, causing them to incorporate features without sufficient fluency. The untuned model's baseline response also shows reduced clarity, further highlighting the model size constraint.

Our findings demonstrate that sparse autoencoders effectively extract features from language models and enhance monosemantic neuron identification. Moreover, our contrastive approach provides a potentially more effective method for feature extraction than standard auto-interpretability techniques. However, model performance limitations at our tested scale remain evident. Future research applying this methodology to larger models with greater computational resources would likely yield more nuanced insights into how modern large language models handle multi-feature steering.

IV. ETHICS

Our research raises important questions about potential misuse of feature steering, particularly regarding bypassing LLM safeguards. Our findings indicate that multi-feature steering is less effective at circumventing safety mechanisms compared to more targeted methods. However, this work provides valuable insights into how safety mechanisms are encoded within neural networks, potentially informing more robust polysemantic safety implementations that resist targeted manipulation. Understanding these encoding patterns could ultimately strengthen, rather than weaken, model safety by suggesting architectural modifications that distribute safety mechanisms across multiple features, making them more resistant to individual feature interventions.

V. CONCLUSION

This research advances mechanistic interpretability by introducing Sparse Autoencoder-based Multi-Feature Steering, a method for extracting and controlling latent representations in neural networks to enhance transparency and controllability of language models. Building on superposition research, our approach addresses the challenge of polysemantic neurons—where individual neurons encode multiple unrelated concepts—by enabling structured steering of disentangled feature activations. By leveraging recent advances in dictionary learning and monosemantic feature extraction, we provide deeper understanding of how language models represent information at the neuron level and demonstrate how models can be made more transparent, steerable, and aligned with human intent.

While our method improves feature separation and steering, several challenges persist. Our experiments with Gemma-2B reveal that smaller models may lack capacity to encode complex high-level conceptual features, limiting their ability to leverage feature steering fully. Feature entanglement, sparsity trade-offs, and ensuring generalizability of extracted features across architectures remain open challenges. Additionally, excessive activation clamping often disrupted model coherence, highlighting the importance of careful feature selection and calibrated activation scaling.

This work contributes to AI alignment and safety by providing a structured approach to understanding, modifying, and controlling deep learning models at the feature level. By introducing a scalable method for steering AI behavior through interpretable feature extraction, we advance the development of more reliable, controllable, and transparent AI systems—particularly valuable for high-stakes applications in legal AI, medical diagnostics, and automated decision-making, where model accountability and interpretability are essential.

VI. FUTURE WORK

Our future research will focus on refining and expanding Sparse Autoencoder-based feature steering along several key dimensions. We aim to enhance multi-feature steering by optimizing feature selection methods and balancing techniques, developing more structured approaches to create predictable, stable, and effective interventions across contexts. This includes improving feature selection mechanisms to increase intervention precision while minimizing unintended feature interactions.

We plan to establish standardized evaluation benchmarks measuring feature steering effectiveness across diverse datasets, architectures, and steering techniques. A consistent evaluation framework is essential for comparing different interpretability approaches and ensuring reproducibility. Developing a comprehensive benchmark suite for mechanistic interpretability will enable structured, quantitative assessment of steering performance across the field.

The generalizability of our findings across different model architectures, training regimes, and domains requires further investigation. The specific features identified and their responsiveness to steering may vary substantially across different models, potentially limiting the broader applicability of specific feature interventions discovered through our methodology.

VII. LIMITATIONS

Several significant limitations in our work must be acknowledged. First, compute requirements severely constrained our ability to test larger language models. Even the Gemma-2B model required up to 17GB of VRAM, making experiments with larger models like Gemma-7B infeasible with our available computational resources. This limitation is particularly significant as we hypothesize that larger models with richer feature spaces would likely demonstrate more consistent and interpretable outputs when subjected to multi-feature steering.

Second, our feature extraction methodology, while effective for identifying some interpretable features, still faces challenges in extracting high-level abstract concepts in smaller models. The limited representational capacity of Gemma-2B means that some complex conceptual features simply may not exist within the model in a form amenable to isolation and manipulation.

Third, our evaluation metrics for measuring the impact of feature steering interventions would benefit from further standardization and validation. While we observed clear qualitative effects from feature manipulation, developing more robust quantitative measures of steering effectiveness remains an important area for refinement.

REFERENCES

[1] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah, "Toy models of superposition," *Transformer Circuits*, 2022. [Online]. Available: https://transformer-circuits.pub/2022/toy_model/index.html

[2] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. L. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah, "Monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits*, 2023. [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features/index.html

[3] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, A. Tamkin, E. Durmus, T. Hume, F. Mosconi, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, "Scaling monosemanticity: Extracting interpretable features from the claude 3 sonnet," *Transformer Circuits*, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html