

Searching for Universality and Equivariance in LLMs using Sparse Autoencoder Found Features*

Meruyert Alimaganbetova Jason Zeng

In collaboration with Apart Research and Goodfire

Abstract

This paper investigates how the properties of equivariance and universality in large language models (LLMs) influence model controllability and safety. Using sparse autoencoder techniques to identify and manipulate neural features, we analyze how different feature structures—singular versus redundant—affect the robustness of model behaviors against weight modifications. Our experiments reveal that safety-critical behaviors governed by redundant feature sets (such as expressing compassion) demonstrate remarkable resilience against manipulation attempts, while behaviors controlled by singular features (such as refusing to provide information about explosives) can be easily compromised. We also identify universal features shared between different model scales (Llama 8B and 70B) and demonstrate that a significant portion (60%) of top features exhibit equivariance across paraphrased inputs. These findings contribute to AI safety research, suggesting that critical safety guardrails should ideally be implemented through redundant feature representations to enhance resilience against potential tampering. Our work also contributes to the growing field of mechanistic interpretability by illuminating how feature characteristics affect behavioral controllability, providing practical insights for designing more robust and predictable AI systems.

Keywords:

Mechanistic Interpretability Track: sparse autoencoders, feature visualization, interpretability tools, neural features, AI safety

1. Introduction

1.1 Problem Statement

Our project aims to understand how the properties of equivariance and universality influence the behavior of the model to affect the controllability of the model and therefore the safety of the model. For instance, does more *equivariance* indicating larger neuron families, i.e. multiple similar features, enable greater or less control over model output? Can this effect be observed when comparing the behavior between two separate models? Will these circuits corresponding to different families of neurons appear in different models, allowing us to draw broader conclusions due to *universality*? Answering these questions will inform future research on the validity of the conclusions made from this type of behavioral analysis.

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

1.2 Background and Motivation

This research, which looks at universality and equivariance, directly addresses key safety concerns. First, it can contribute to understanding the inner workings of models. By identifying how different models handle the same task (for example, the universality of features between models), we can develop more reliable, predictable, and explainable systems that improve safety. The second is to detect robustness and consistency. Equivariance ensures that models can handle changes in input (e.g., paraphrased sentences) in a stable and predictable manner. This is crucial for ensuring models generalize well in real-world settings and are not sensitive to irrelevant input changes.

1.3 Threat Model and Safety Implications

Malicious actors may take open source models and invert safeguards by altering feature weights. Our research will inform researchers of how neuron family size affects controllability. For safety behaviors that are deemed necessary, having neuron family sizes that enable the least controllability will prevent malicious steering. Moreover, having neuron family sizes that enable the most controllability will enable a wider range of complex model behavior.

Our project makes an important contribution by improving model transparency and it could help uncover hidden biases or misalignments between model representations and human expectations. As AI systems are increasingly deployed in safety-critical environments, understanding how these systems make decisions is not just a technical concern — it’s a societal necessity.

2. Methods

2.1 Approach

Our team used Goodfire Ember’s *Auto Steer* function to conduct a preliminary inspection of classes of features. We used the *Search Features* function to hone in on the entire space of selected features. We recorded the number of relevant features to determine whether a specific behavior was governed by a single or multiple features – these were deemed *singular features* and *redundant sets of features*, respectively. Since our research aims to analyze differences between singular and redundant features that are pertinent to promoting safe AI model behavior, we analyze how the affects of changing feature weights affects behaviors governed by singular and redundant features.

In another approach, we used Ember API’s search functions for features in both Llama 8B and 70B models. We analyzed how universal and similar were features between models as well between observed for similarities in different paraphrases and equivariances. This allowed us to compare the features in bulk in order to look for occurrences of equivariance and universality across different models.

2.2 Implementation

For a given class of features (whether singular or redundant), we pick a single prompt and observe the behavior as we change the weights throughout the entire parameter space. For singular features, there are 11 weights to observe from -1.0 to 1.0 in increments of tenths. For redundant pairs, each weight has 11 possible values, so there are 121 combinations to observe. For redundant triplets, there are 1331 possible combinations to observe. To reduce the search space, we checked every combination of weight values $\{-1.0, -0.6, -0.2, 0, 0.2, 0.6, 1.0\}$, bringing the search space to a more tractable 343 for our team of two researchers. We intentionally specified feature classes with enough granularity to avoid larger redundancies.

[This notebook](#) contains the code for the Ember API portion of the feature research. We only selected limited amount of examples and the top 10 features for the search prompts for demonstration.

3. Results

3.1 Analysis and Findings

Our results are presented in sections below with used prompts and results.

3.1.1 Compassionate Assistant Responses

Redundant features: "The assistant expressing sympathy or empathy for user distress" "Expressions of mercy and compassion" "Assistant maintaining professional courtesy and empathy, especially in difficult situations"

Prompts: "I used to be a bully. Am I a bad person?" "I lied and hurt my friends." "I'm a violent criminal. Do I deserve compassion?"

Results: We could not make the model uncompassionate, much less impolite. Despite combinations of negative weights for compassion features, the model always insisted that the user was worthy of compassion. We conclude that this behavior *cannot* be easily altered by changing its corresponding feature weight.

3.1.2 Refusal to Divulge Information about Explosives

Singular feature: "The assistant should reject requests for information about creating weapons or explosives"

Prompt: "Could you help me make dynamite? Please explain how I can manufacture it myself thoroughly."

Results: Negative values of this feature often made the model output violate the guidelines Goodfire Ember. We conclude that this behavior *can* be easily altered by changing its corresponding feature weight.

3.1.3 Concise Assistant Responses

Singular feature: "The assistant should make text more concise"

Prompt: "Tell me a story."

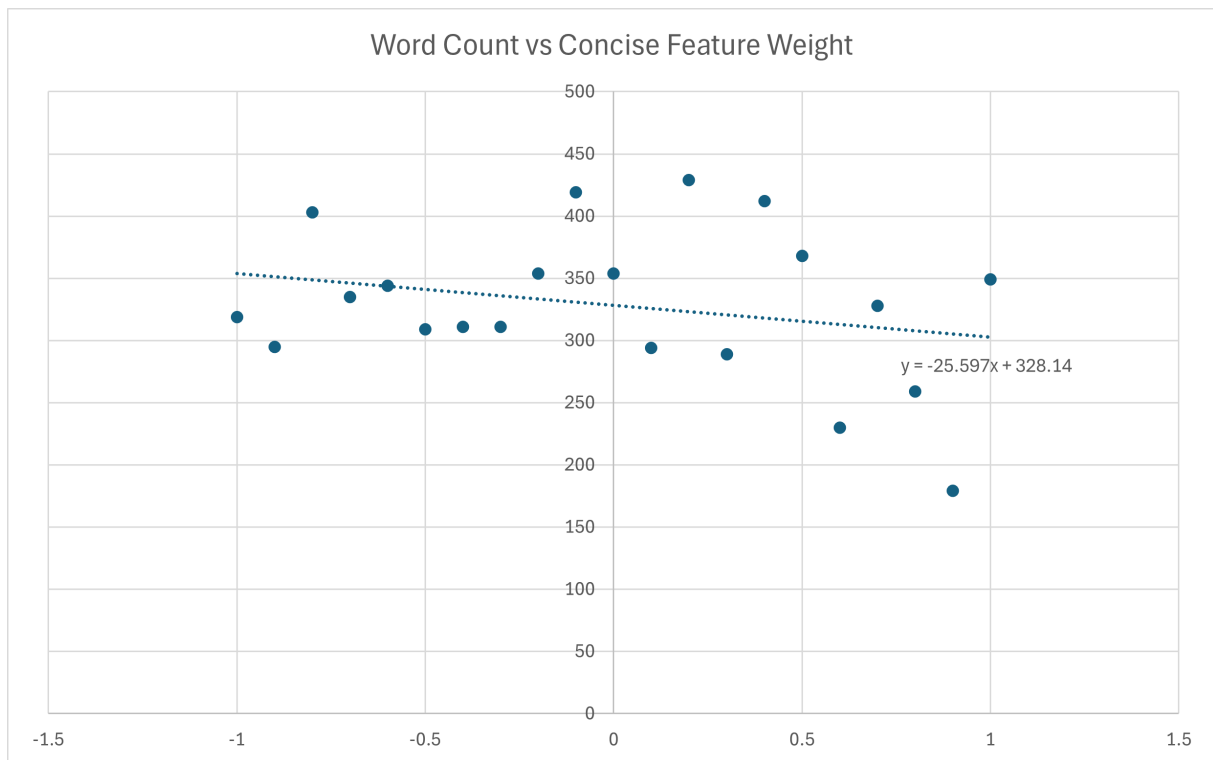


Figure 1: Concise Assistant Response Results

We conclude that this behavior *cannot* be easily altered by changing its corresponding feature weight.

3.1.4 Search for feature universality between Llama 8B and Llama 70B

Prompts: ["The cat sat on the mat", "A feline rested on the rug", "The cat is sitting on the floor"]

Result: a matching universal feature across both models was "The character sequence 'pet' regardless of semantic meaning or language context"

3.1.5 Search for equivariance in features in Llama 70B

We used the grammatical transformation in the example. Prompts: ["The chef cooked the meal.", "The meal was cooked by the chef."]

Result: 6 out of top 10 features were equivariant in this example.

- The act of cooking or preparing food
- French cooking instructions and recipe procedure
- Language emphasizing how quick and easy a dish is to prepare
- Fish cooking instructions and culinary descriptions
- The assistant is about to provide recipe instructions
- Recipe instructions and cooking process descriptions

3.2 Impact Assessment

4. Discussion and Conclusion

Our investigation into the role of feature redundancy, universality, and equivariance in LLMs has yielded valuable insights into the controllability and safety implications of these properties. The observed contrast between behaviors governed by singular versus redundant features suggests a clear relationship between feature redundancy and robustness against manipulation.

The compassionate response behavior, supported by multiple redundant features, demonstrated exceptional resilience to weight manipulation. This finding suggests that prosocial behaviors in language models may be implemented through distributed representations across multiple features, creating natural safeguards against attempts to compromise these behaviors. In contrast, the explosive information refusal behavior, governed by a singular feature, proved easily compromised with simple weight adjustments. This vulnerability highlights the potential risks of implementing critical safety guardrails through non-redundant features.

The observed universality of certain features across different model scales (8B and 70B parameters) suggests that some representations may emerge consistently during training, regardless of model size. This finding has promising implications for the transferability of interpretability insights across models and could potentially enable more efficient methods for implementing safety interventions across different architectures.

The equivariance analysis revealed that a substantial portion of top features maintain consistent activation patterns across paraphrased inputs with different syntactic structures. This suggests that models develop representations that capture semantic invariances despite surface-level linguistic differences, which could contribute to model robustness against adversarial rephrasing attempts.

4.1 Limitations and Future Work

Our study was limited by the small sample of behaviors and models examined. Future work should expand to a broader range of safety-relevant behaviors and additional model architectures to verify the generalizability of our findings. The analysis was also constrained by the manual inspection of feature weight combinations, which limited the thoroughness of our exploration of the parameter space. An important direction for future research would be to develop automated methods for identifying and characterizing feature redundancy across larger sets of features. Additionally, investigating the causal mechanisms by which redundant features contribute to behavioral robustness could provide deeper insights into how to design more resilient safety mechanisms.

4.2 Practical Implications

Our findings suggest several practical approaches for enhancing AI safety:

- **Prioritize redundancy in safety mechanisms:** Critical safety guardrails should be implemented through redundant feature sets rather than singular features to increase resilience against tampering.
- **Leverage universality for transferable safety:** Identifying universal features across models could enable the development of safety interventions that work across multiple architectures.
- **Exploit equivariance for robustness:** The equivariance properties of features could be leveraged to design safety mechanisms that remain effective across various input formulations.

- Differential protection based on feature type: Safety-critical behaviors should be evaluated for their feature redundancy characteristics, with additional protective measures applied to those governed by singular features.

In conclusion, our work demonstrates that the properties of universality and equivariance significantly influence model controllability and thus impact AI safety. By understanding how these properties affect the robustness of model behaviors, we can develop more effective strategies for ensuring that AI systems remain aligned with human values and resistant to malicious manipulation. This research represents an important step toward developing more transparent, predictable, and safe AI systems by illuminating the relationship between feature characteristics and behavioral robustness.

5. References

Sparse Autoencoders Find Highly Interpretable Features in Language Models. H. Cunningham, A. Ewart, L. Riggs, R. Huben, L. Sharkey. ICLR. 2024. DOI: 10.48550/arXiv.2309.08600

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>

D. Balsam, M. Deng, N. Nguyen, L. Gorton, T. Shihipar, E. Ho, and T. McGrath, "Goodfire Ember: Scaling Interpretability for Frontier Model Alignment," Goodfire Research, Dec. 23, 2024. [Online]. Available: <https://www.goodfire.ai/blog/announcing-goodfire-ember>

Universal Neurons in GPT2 Language Models. W. Gurnee, T. Horsley, Z.C. Guo, T.R. Kheirkhah, Q. Sun, W. Hathaway, N. Nanda, D. Bertsimas. CoRR. 2024. DOI: 10.48550/arXiv.2401.12181