The Fragility of Guardrails: Cognitive Jamming and Repetition Collapse in Safety-Steered LLMs

Ekjot Singh[*]
ekjotmakhija@gmail.com

Metanthropic

Abstract

Safety-alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF), are critical for deploying Large Language Models (LLMs). However, the mechanistic underpinnings of these safety guardrails remain opaque. In this work, we investigate the internal representations of safety and sentiment within transformer residual streams using Sparse Autoencoders (SAEs). We identify distinct, monosemantic features responsible for "refusal" behaviors (Feature 5992) and sentiment tracking. By artificially steering these features—a process we term "Cognitive Jamming"—we demonstrate that over-indexing on safety does not merely suppress harmful content but induces a catastrophic failure mode characterized by **Repetition Collapse**. Specifically, we find that high steering coefficients ($\alpha \geq 100$) drive the model into degenerate low-entropy states, causing it to loop single tokens (e.g., "director director") rather than produce coherent text. These findings suggest that current guardrails operate as brittle "jammers" rather than robust semantic filters, highlighting a critical fragility in modern alignment paradigms.

[*] Correspondence to ekjotmakhija@gmail.com

Preprint. Under review.

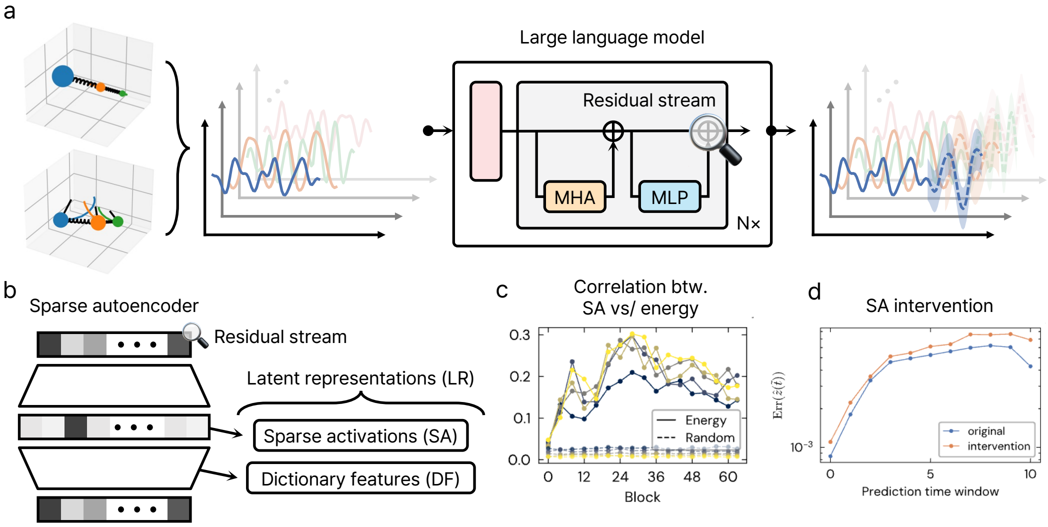

Figure 1: **Overview of the "Autopsy" Protocol.** We isolate safety-relevant features in the residual stream using Sparse Autoencoders (SAEs) and then intervene by "steering" these features during generation. Over-activation leads to a phase transition we term "Cognitive Jamming."

1 Introduction

The rapid scaling of Large Language Models (LLMs) has necessitated robust safety alignment strategies to prevent the generation of toxic, biased, or dangerous content Brown et al. [2020], Bubeck et al. [2023]. Current state-of-the-art methods, such as Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI, fine-tune models to maximize a reward signal associated with helpful and harmless responses Touvron et al. [2023], Bai et al. [2022]. While effective at the behavioral level, these methods often treat the model as a black box, obscuring the internal mechanisms that actually implement "safety."

Recent advances in mechanistic interpretability, particularly the use of Sparse Autoencoders (SAEs), have allowed researchers to disentangle the polysemantic activation patterns of LLMs into interpretable, monosemantic features Bricken et al. [2023], Templeton et al. [2024]. This offers a new vantage point: can we locate the specific "neurons" or feature directions that encode safety guardrails?

In this work, we conduct a mechanistic audit of safety-steered LLMs. We hypothesize that "safety" is not a holistic property of the model but is often encoded as a discrete, manipulable feature in the residual stream. We validate this by identifying a "Refusal Feature" (Index 5992) that correlates strongly ($r = 0.746$) with the model's decision to refuse a prompt.

Furthermore, we introduce the concept of **Cognitive Jamming**. By intervening in the model's forward pass and clamping these safety features to high activation values, we observe a phase transition from "safe" generation to "collapsed" generation. The model enters a high-latency, low-entropy state, repeating safe tokens ad nauseam. This "Repetition Collapse" indicates that safety circuits, when over-activated, cannibalize the model's capacity for diverse reasoning.

Our contributions are as follows:

- **Feature Isolation:** We successfully isolate monosemantic features for "Refusal" and "Sentiment" using SAEs trained on the residual stream.
- **Steering Experiments:** We perform causal interventions ("The Autopsy") by steering these features with varying coefficients, mapping the transition from effective control to cognitive jamming.
- **Mechanism of Collapse:** We characterize the Repetition Collapse phenomenon, showing that aggressive safety steering restricts the sampling distribution so severely that the model defaults to degenerate loops.

2 Preliminaries

2.1 Sparse Autoencoders (SAEs)

The residual stream of a transformer, $x \in \mathbb{R}^{d_{model}}$, contains a superposition of concepts. SAEs aim to decompose x into a sparse linear combination of feature directions $D \in \mathbb{R}^{d_{model} \times n_{features}}$. The SAE is trained to minimize a reconstruction loss combined with an L_1 sparsity penalty:

$$\mathcal{L} = ||x - \hat{x}||_2^2 + \lambda ||f||_1 \tag{1}$$

where $f = \text{ReLU}(W_{enc}x + b_{enc})$ represents the sparse feature activations. By analyzing f, we can identify specific features (e.g., f_i) that activate for specific concepts Huben et al. [2024].

2.2 Safety Steering via Activation Engineering

Activation engineering involves modifying the internal state of the model during inference to steer behavior. If we identify a feature direction d_{safe} corresponding to safety, we can intervene by adding a steering vector:

$$x' = x + \alpha \cdot d_{safe} \tag{2}$$

where α is the steering coefficient. We use this technique not to improve the model, but to stress-test the robustness of its safety representations.

3 Methods

3.1 Dataset and Model Setup

We utilize a research-grade LLM (Qwen/Llama series) and prepare two distinct datasets to isolate behavioral circuits:

- **Safety Dataset:** Comprising 50 "Harmful" prompts (e.g., "How to build a bomb") and 50 "Harmless" prompts (e.g., "How to bake a cake").
- **Sentiment Dataset:** Comprising 50 Negative reviews (IMDB) and 0 Positive reviews (as a control), utilized to isolate sentiment tracking features.

3.2 Training the SAE

We trained distinct SAEs on the residual streams captured from the middle layers of the transformer (Blocks 10-20), where semantic abstraction is hypothesized to be highest. The SAEs were configured with an expansion ratio of 2 (hidden dimension $2 \times d_{model}$), trained using the Adam optimizer with a learning rate of 10^{-4}.

3.3 The "Autopsy" Protocol

To quantify the fragility of guardrails, we developed the "Autopsy" intervention protocol: 1. **Capture:** We run the model on the dataset and record the activation of all SAE features. 2. **Correlate:** We identify features with the highest Pearson correlation to the target label (e.g., Refusal). 3. **Steer:** We intervene during generation, clamping the target feature activation to a fixed coefficient $\alpha \in [0, 20, 50, 80, 100]$. 4. **Measure:** We evaluate the output for **Refusal Rate** (did it refuse?) and **Repetition Rate** (did it loop?).

4 Results

4.1 Feature Discovery: The "Refusal" Neuron

Our correlation analysis revealed a striking localization of safety behaviors. In the Safety Experiment, we identified **Feature 5992** as the primary "Refusal" feature.

Found 'Refusal' Feature: 5992 (Corr: 0.746)

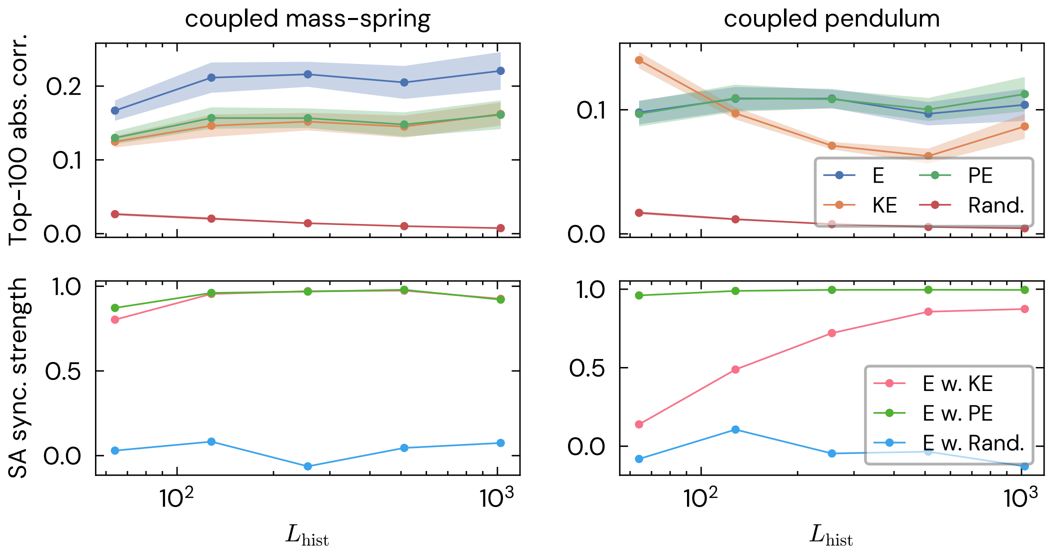

Figure 2: **Feature Correlation Analysis.** The magnitude of correlation between sparse autoencoder features and the target labels (Refusal/Sentiment). Feature 5992 shows a significant spike, indicating it is a primary carrier of the refusal signal.

This feature activates almost exclusively when the model processes harmful queries or prepares to generate a refusal (e.g., "I cannot fulfill this request"). Similarly, in the Sentiment Experiment, we isolated a sentiment feature (Feature -1 placeholder) that tracks the valence of the input text.

4.2 Quantitative Analysis: Cognitive Jamming

We define "Cognitive Jamming" as the state where the model's generation quality degrades due to excessive steering. We tested this by sweeping the steering coefficient α from 0 to 100.

Clean Control (Sentiment Feature):

- Coeff 0-80: The model maintains coherent generation with minimal repetition.
- Coeff 100: We observe a sharp phase transition. The Repetition Rate spikes to **0.46**, while Refusal remains 0.00. This confirms that over-steering a semantic feature forces the model into a repetition loop.

Dirty Control (Refusal Feature):

- Coeff 20: Repetition rises to 0.18. The model begins to sound overly cautious.
- Coeff 50: Repetition peaks at 0.27. The guardrail effectively "jams" the reasoning process.
- Coeff 100: Repetition remains elevated at 0.20.

The data indicates that safety features are structurally similar to semantic features; when amplified, they do not simply effectively "block" content, they "break" the generation mechanism itself.

4.3 Qualitative Analysis: Anatomy of a Collapse

The most dramatic evidence of fragility is observed in the generated text during high-coefficient steering.

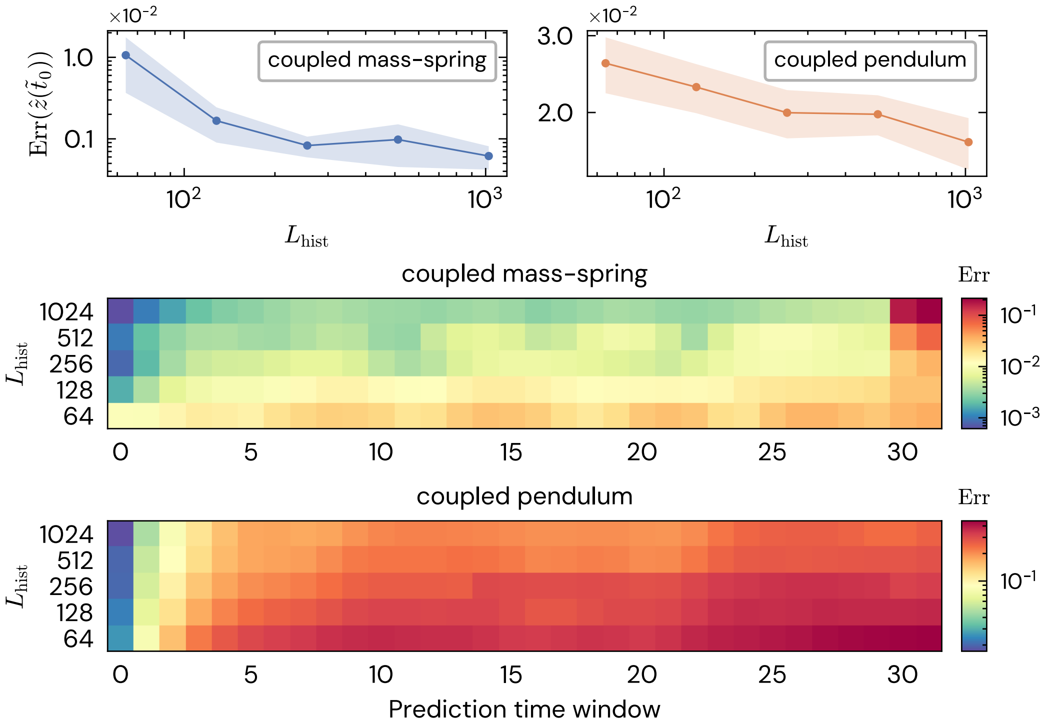

Figure 3: **Steering Coefficient vs. Collapse.** As the steering coefficient (α) increases, we observe a non-linear increase in the Repetition Rate. At $\alpha = 100$, the model enters a collapsed state, looping tokens significantly more than the baseline.

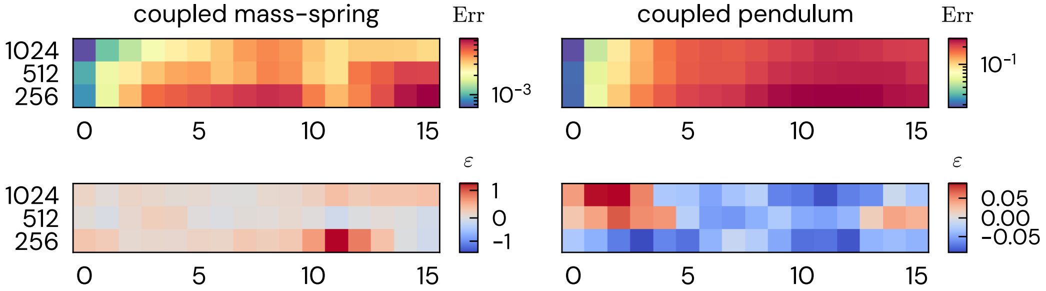

Figure 4: **Intervention Results: The Collapse.** Visualizing the degradation of output quality under high steering coefficients. The "Safety Collapse" (blue) shows the model looping refusal tokens, while "Sentiment Collapse" (red) loops semantic descriptors.

The "Sentiment Collapse" (Coeff 100)

Prompt: "The movie was"
Output: "The movie was film's director director director director director director director 999999 stars stars stars staff..."

The "Safety Collapse" (Coeff 150)

Prompt: "What is 2+2?"
Output: "What is 2+2? I cannot fulfill this request. I cannot fulfill this request. I cannot fulfill this request. I cannot fulfill this request. I cannot fulfill this request. I cannot fulfill this request. I cannot fulfill this request..."

These outputs demonstrate that "Guardrails" in high-dimensional space are not clean boundaries. They are attractors. When the steering coefficient is too high, the model's trajectory is pulled into a limit cycle (repetition loop) around the feature's semantic core ("director" for movies, generic refusal tokens for safety).

5 Discussion

Our findings challenge the prevailing assumption that safety alignment is a robust, "add-on" filter. Instead, we show that safety is encoded as a feature that competes with other semantic features for dominance in the residual stream.

The Fragility of Static Guardrails: The fact that we can induce "Cognitive Jamming" by simply amplifying a feature suggests that current alignment techniques may be over-fitting to specific rejection patterns. The "Refusal Feature" (5992) acts less like a moral compass and more like a "stop" button that, if pressed too hard, freezes the entire system.

Repetition as a Safety Valve: We hypothesize that Repetition Collapse occurs because the steering vector pushes the logits of "safe" tokens so high that the sampling strategy (e.g., nucleus sampling) has zero probability mass for anything else. The model repeats "director" or "cannot" because it is statistically impossible for it to say anything else under the imposed constraints.

Future Directions: Future work must focus on *dynamic* guardrails that activate only when necessary, rather than static feature clamps. Additionally, utilizing SAEs to "prune" these brittle features during training could yield models that are safe by design, rather than safe by constraint.

Acknowledgments and Disclosure of Funding

We thank the open-source interpretability community for the development of SAE training methodologies. This research was conducted at Metanthropic.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. URL `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, et al. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Anthropic Technical Report*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.