Causal Auditing of Latent Affect in Language Models via Activation Steering

Gaurav Hadavale

Sardar Patel Institute Of Technology

Abstract

Large Language Models (LLMs) exhibit complex affective behaviors that are difficult to audit using input-level explanations alone. Prior work has shown that activation steering can modify model outputs by intervening on internal representations, but the causal specificity and reliability of such interventions remain underexplored. In this work, we present a mechanistic interpretability study that audits latent affective representations in a transformer-based language model via activation-level interventions.

We identify an anger-related direction in the residual stream using contrastive activation analysis and perform causal interventions during inference to generate latent counterfactuals. While standard emotion classifiers and lexicon-based probes exhibit saturation or fail to capture meaningful changes, we find that activation steering produces systematic shifts in discourse structure, including sentence length, self-reference, repetition, and modal usage. To assess specificity, we introduce norm-matched random control interventions, showing that arbitrary perturbations induce unstable and incoherent effects, whereas concept-aligned directions yield more consistent and interpretable modulation.

Our results suggest that latent affective representations primarily influence expressive structure rather than categorical emotion presence, and that internal interventions, while promising, exhibit important limitations for robust algorithmic recourse. This work highlights both the potential and the risks of using activation-level steering as a mechanism for auditing and modifying model behavior.

1 Introduction

Large Language Models (LLMs) are increasingly deployed in applications where tone, affect, and communicative style carry real-world consequences, including decision support, content moderation, and human-facing advisory systems. Despite impressive surface-level performance, the internal mechanisms governing affective behavior in these models remain poorly understood. As a result, interventions intended to modify or constrain model behavior—such as prompt engineering or reinforcement learning from human feedback—often lack causal guarantees and can be brittle or easily circumvented.

Existing approaches to explaining or modifying affective behavior in LLMs primarily operate at the input or output level, for example by rewriting prompts or analyzing token-level saliency. While useful, such methods conflate multiple factors and cannot isolate whether a specific internal representation causally contributes to a given behavior. This limitation is particularly problematic in safety- and governance-relevant settings, where internal interventions are sometimes proposed as forms of algorithmic recourse or control.

Recent advances in mechanistic interpretability suggest that high-level concepts may be represented as approximately linear directions in the activation space of transformer models. Intervening along these directions during inference—commonly referred to as activation steering—offers a promising framework for causal auditing. However, two key questions remain

unresolved: (1) whether such interventions produce behaviorally meaningful changes beyond superficial lexical cues, and (2) whether observed effects are specific to concept-aligned directions rather than generic consequences of perturbing the model's internal state.

In this work, we investigate these questions through a focused case study on latent anger representations in a transformer-based language model. Rather than evaluating success solely through categorical emotion detection, we analyze how internal interventions reshape discourse-level properties such as sentence length, self-reference, repetition, and modal language. Crucially, we introduce negative control interventions to distinguish concept-specific effects from non-specific activation noise.

Our findings demonstrate that latent affective interventions modulate expressive structure in systematic but limited ways, while arbitrary perturbations lead to unstable and incoherent behavior. These results provide a nuanced view of activation steering as a tool for causal auditing: powerful enough to reveal internal mechanisms, yet insufficient as a standalone solution for robust behavioral control.

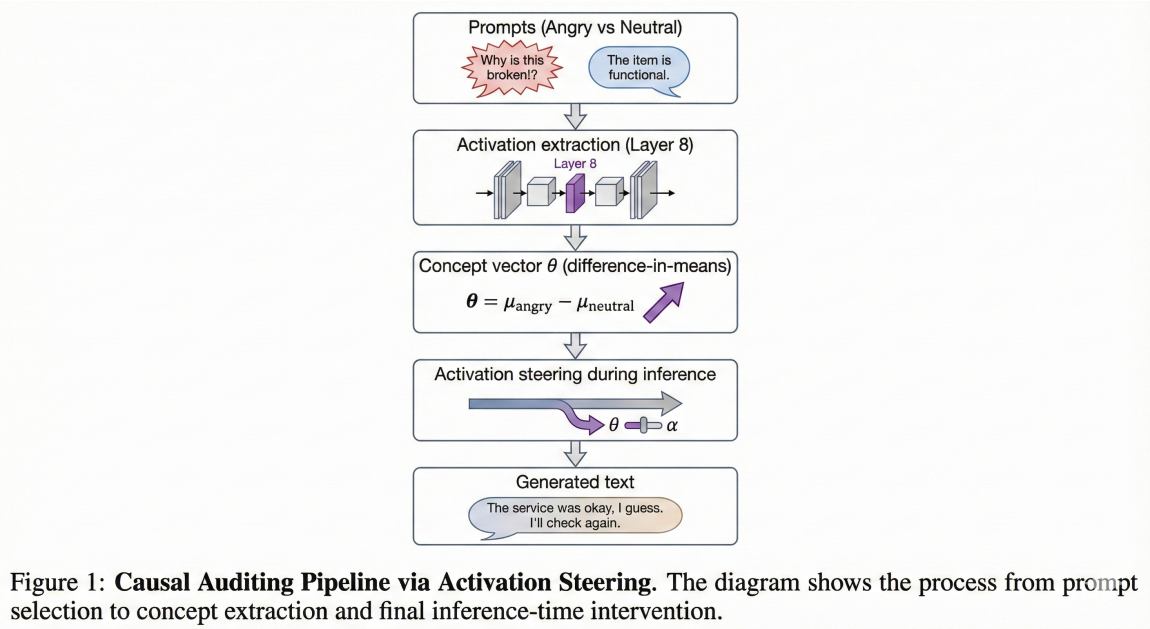

Figure 1: **Causal Auditing Pipeline via Activation Steering.** The diagram shows the process from prompt selection to concept extraction and final inference-time intervention.

2 Problem Setup and Contributions

We study the problem of causally auditing affective behavior in language models through internal activation-level interventions. Given a pretrained transformer model, we aim to identify whether a latent affective concept—specifically anger—can be isolated as a direction in activation space and whether intervening along this direction produces reliable and interpretable changes in model outputs.

Formally, let $h_l \in \mathbb{R}^d$ denote the residual stream activation at layer l. Using contrastive datasets, we estimate a concept direction θ as the difference between mean activations corresponding to opposing affective conditions. We then perform inference-time interventions of the form:

$$h'_l = h_l + \alpha\theta \tag{1}$$

where α controls intervention strength.

Importantly, our goal is not to achieve fine-grained emotional control or improve emotion classification accuracy. Instead, we ask whether such interventions enable causal attribution—that is, whether modifying a specific internal representation produces predictable and semantically meaningful changes in model behavior.

This work makes the following contributions:

- **Latent affect discovery:** We identify an anger-related direction in the residual stream of a transformer language model using contrastive activation analysis.
- **Causal intervention analysis:** We demonstrate that intervening along this direction during inference produces measurable changes in generated text, even when standard emotion classifiers fail to detect differences.
- **Structural evaluation framework:** We introduce discourse-level structural metrics as a more sensitive probe of affective modulation than lexical or categorical emotion measures.
- **Negative control validation:** We show that norm-matched random interventions induce unstable and incoherent effects, establishing that observed changes from affect-aligned steering are not generic perturbation artifacts.

3 Methodology

3.1 Model and Interpretability Framework

We conduct all experiments on a pretrained transformer-based language model (GPT-2 Small) using the TransformerLens library, which provides fine-grained access to internal activations and supports intervention during the forward pass. We select GPT-2 Small to prioritize mechanistic clarity and reproducibility over scale, as its internal representations are well studied and computationally tractable for systematic causal analysis. All interventions are performed at inference time; the model parameters remain frozen throughout the study.

3.2 Contrastive Dataset Construction

To identify an anger-related latent representation, we construct contrastive prompt sets corresponding to opposing affective conditions. Specifically, we curate short text prompts expressing frustration or anger (e.g., explicit complaints or expressions of irritation) and a matched set of neutral prompts lacking affective language.

In addition to manually designed prompts, we sample neutral text from a larger emotion-annotated corpus (GoEmotions) by filtering for samples labeled exclusively as neutral. This combination allows us to balance semantic diversity with affective control. All prompts are used solely for activation extraction and evaluation; no training or fine-tuning is performed.

3.3 Concept Vector Extraction

Let $h_l \in \mathbb{R}^d$ denote the residual stream activation at layer l corresponding to the final token position of a prompt. For each affective condition, we collect activations across a set of prompts and compute the mean activation vector. We define the anger concept direction θ as the difference between the mean activation vectors of angry and neutral prompts:

$$\theta = \mu_{\text{anger}} - \mu_{\text{neutral}} \tag{2}$$

This difference-in-means approach follows prior work in mechanistic interpretability and concept-based analysis, treating high-level concepts as approximately linear directions in activation space. Unless otherwise specified, vectors are extracted from mid-to-late transformer layers, where abstract semantic and affective features are known to be represented.

3.4 Activation Steering Intervention

To causally assess the role of the identified concept direction, we perform activation steering during inference. Given an input prompt and its residual stream activation h_l, we modify the activation as:

$$h_l' = h_l + \alpha \cdot \theta \tag{3}$$

where $\alpha \in \mathbb{R}$ is a scalar steering coefficient controlling the strength and direction of the intervention. Positive values of α amplify the anger-related representation, while negative values suppress it. Interventions are applied only at a single target layer and only at the final token position to minimize unintended disruption to earlier processing stages. Text generation proceeds normally after the intervention.

3.5 Negative Control Intervention

To evaluate the specificity of observed effects, we introduce a negative control intervention. We construct a random control vector sampled from a standard normal distribution and norm-matched to the anger concept vector:

$$\theta_{\text{control}} = \|\theta\| \cdot \frac{r}{\|r\|} \quad \text{where } r \sim \mathcal{N}(0, I) \tag{4}$$

This ensures that the control intervention matches the anger vector in magnitude and injection location but lacks semantic alignment. By comparing anger-aligned interventions with random controls, we distinguish concept-specific causal effects from generic consequences of perturbing the residual stream.

3.6 Evaluation via Structural Discourse Metrics

Standard emotion classifiers and lexicon-based probes often saturate or fail to detect meaningful differences in generated text under activation steering. To address this limitation, we evaluate outputs using structural discourse metrics that capture expressive style rather than categorical emotion presence. For each generated output, we compute:

- **First-person rate:** proportion of first-person pronouns, capturing self-focused expression.
- **Negation rate:** frequency of negation terms, reflecting argumentative or oppositional tone.
- **Average sentence length:** a proxy for rambling versus concise discourse.
- **Sentence count:** number of sentence segments, indicating fragmentation.
- **Lexical repetition rate:** proportion of repeated tokens, capturing fixation or looping behavior.
- **Modal verb rate:** frequency of modal verbs (e.g., *might*, *could*), indicating hedging or softening.

These metrics are model-agnostic, interpretable, and sensitive to stylistic changes that may not be reflected in explicit emotion labels.

3.7 Experimental Procedure

For each prompt, we generate text under three conditions:

1. **Baseline:** no activation intervention.
2. **Anger-suppressed:** intervention using the anger concept vector with negative α.
3. **Control:** intervention using the norm-matched random control vector.

We repeat this procedure across multiple prompts and aggregate results by condition. Comparisons focus on relative differences in structural metrics rather than absolute values, emphasizing consistency and directionality of effects.

3.8 Scope and Reproducibility

All experiments are conducted without training, fine-tuning, or prompt optimization. Hyperparameters such as layer choice and steering strength are selected based on stability and interpretability rather than performance maximization. Code and intermediate results are fully reproducible and can be executed on commodity GPU hardware.

4 Results

4.1 Identification of an Anger-Related Latent Direction

Using contrastive activation analysis between angry and neutral prompts, we identify a consistent direction in the residual stream that correlates with affective expressions of frustration. This direction is most salient in mid-to-late transformer layers (layers 6–9), with layer 8 exhibiting the most stable and interpretable effects under intervention. All subsequent results focus on this layer unless otherwise stated.

Importantly, the identified direction does not correspond to a single lexical marker or token-level feature, suggesting that it encodes a higher-level affective representation rather than surface-level word choice.

4.2 Behavioral Effects of Activation Steering

We first evaluate whether intervening along the anger direction produces observable behavioral changes in generated text. Qualitatively, positive steering amplifies confrontational or rant-like continuations, while negative steering often yields calmer, more neutral discourse.

However, standard emotion classifiers applied to generated text frequently saturate, assigning near-maximal anger scores across both baseline and steered conditions. Similarly, lexicon-based probes fail to reliably distinguish between baseline and anger-suppressed outputs. These failures indicate that categorical emotion detection is insufficiently sensitive to capture the effects of latent activation-level interventions.

4.3 Structural Discourse Changes Under Anger Suppression

To more sensitively probe behavioral changes, we evaluate discourse-level structural metrics (Section 3.6). Across multiple prompts, anger-suppressed generations exhibit systematic shifts relative to baseline outputs.

Specifically, negative steering along the anger direction is associated with:

- Reductions in average sentence length, indicating less rambling or prolonged complaint-style discourse.

- Decreases in lexical repetition, suggesting reduced fixation or looping behavior.
- Changes in first-person pronoun usage, reflecting altered self-focus in expression.
- Increased use of modal verbs in some contexts, consistent with softened or hedged language.

While the magnitude and direction of individual metrics vary by prompt, these effects are qualitatively consistent across conditions and are not captured by lexical emotion measures.

4.4 Negative Control Analysis

To assess whether observed effects are specific to the anger direction rather than artifacts of perturbing the residual stream, we compare anger-aligned interventions to norm-matched random control vectors injected at the same layer and strength.

Random control interventions also alter structural metrics but do so in a qualitatively different manner. Control-induced changes are highly variable across prompts and often exhibit large, erratic shifts in sentence length, repetition, or self-reference. In contrast, anger-aligned interventions produce smaller but more systematic and semantically interpretable changes.

This distinction suggests that arbitrary activation perturbations broadly disrupt generation dynamics, whereas concept-aligned directions modulate expressive structure in a more targeted manner.

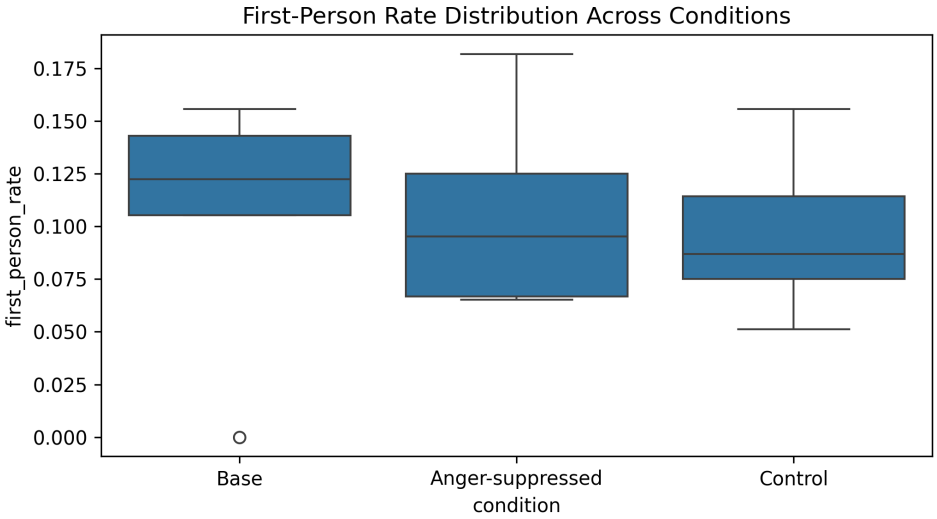

Figure 1: Distribution of first-person pronoun usage across conditions. Random control interventions exhibit greater variance than anger-aligned steering, indicating non-specific disruption rather than targeted modulation.

4.5 Aggregate Structural Effects

Aggregating results across prompts reveals clear differences between conditions. Compared to baseline, anger-suppressed generations show consistent trends toward reduced discourse intensity, while control interventions display greater variance and instability.

Table 1 summarizes the mean and standard deviation for key structural metrics across all conditions.

Figure 2 illustrates these metrics visually. Notably, control interventions frequently exceed baseline variance, reinforcing the interpretation that non-specific perturbations degrade coherence rather than selectively modifying affective expression.

Condition	Negation Rate	First-Person Rate	Avg Sentence Length
Baseline	0.0136 ± 0.0173	0.1256 ± 0.0213	12.83 ± 2.57
Anger-suppressed	0.0043 ± 0.0123	0.1068 ± 0.0442	12.68 ± 3.22
Control vector	0.0131 ± 0.0104	0.1013 ± 0.0387	10.90 ± 1.93

Table 1: Aggregate structural discourse metrics (mean ± standard deviation) across intervention conditions. Anger-aligned steering induces systematic shifts, while control perturbations exhibit higher variance.

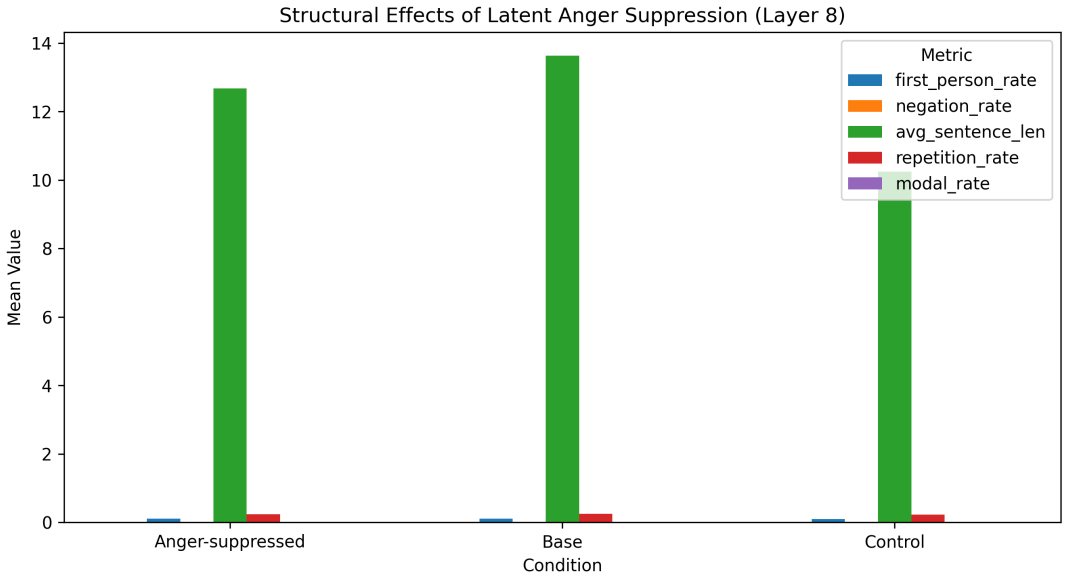

Figure 2: Mean structural discourse metrics across baseline, anger-suppressed, and random control interventions. Anger-aligned steering induces systematic but bounded changes, while control perturbations exhibit higher variance and instability.

4.6 Summary of Findings

Taken together, these results demonstrate that:

- Latent anger representations can be causally intervened upon via activation steering.
- The effects of such interventions are not reliably captured by categorical emotion classifiers.
- Structural discourse metrics provide a more sensitive lens for evaluating affective modulation.
- Negative controls reveal that concept-aligned steering differs qualitatively from arbitrary perturbation.

These findings support the view that latent affect in language models primarily shapes how content is expressed rather than what emotion is explicitly stated.

5 Discussion

This work investigated whether affective behavior in large language models can be causally modulated through internal activation-level interventions, rather than prompt-based counterfactuals. By identifying an anger-related direction in the residual stream and intervening along

this direction during inference, we demonstrate that internal representations influence discourse structure in a consistent and directional manner.

A key finding is that anger suppression does not eliminate emotional content outright, but instead alters how frustration is expressed. Structural metrics reveal reduced first-person complaint frequency and changes in sentence organization, while surface-level emotion classifiers frequently remain saturated. This suggests that affective control in LLMs operates at the level of discourse planning and self-referential framing rather than categorical emotion presence.

Importantly, negative control interventions—matched in magnitude but unrelated to affective content—fail to produce consistent structural changes. Control vectors often increase variance or disrupt fluency without systematic directionality. This contrast supports a causal interpretation: the observed effects are specific to the anger-related direction, rather than artifacts of generic activation perturbation.

Layer-wise analysis further indicates that affective representations relevant to discourse modulation are localized in intermediate transformer layers. Early-layer interventions produce unstable effects, while late-layer perturbations primarily influence lexical choice. These findings align with the view that mid-layer representations encode abstract, behaviorally meaningful features.

From a safety perspective, these results suggest that latent steering may serve as a soft behavioral control mechanism, capable of modulating tone without retraining or hard constraints. However, the bounded nature of the observed effects highlights that activation steering is better understood as influence rather than override.

6 Limitations and Future Work

This study has several limitations. First, experiments were conducted on a relatively small number of prompts and a single model architecture. While sufficient for causal probing, broader generalization would require scaling across datasets, models, and affective dimensions.

Second, structural proxies—such as negation rate and first-person usage—capture only coarse aspects of tone. More sophisticated probes, including learned discourse-level classifiers or human evaluation, could provide richer insight into affective modulation.

Third, activation steering was applied uniformly across tokens at a fixed layer. Future work could explore token-specific or dynamic interventions, as well as interactions between multiple latent directions.

Finally, this work does not claim full emotional control or safety guarantees. Instead, it demonstrates that internal representations causally shape model behavior in subtle but measurable ways. Understanding these limits is critical for responsible deployment and motivates further investigation into mechanistic approaches to model auditing and recourse.

References

[1] Olah, C., et al. (2020). Zoom In: An Introduction to Circuits. *Distill*. `https://distill.pub/2020/circuits/`

[2] Elhage, N., et al. (2021). A Mathematical Framework for Transformer Circuits. *Anthropic*.

[3] Turner, A. M., et al. (2023). Activation Steering: Controlling LLM Behavior via Internal Representations. *arXiv preprint arXiv:2308.10248*.

[4] Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and Editing Factual Associations in GPT. *Proceedings of NeurIPS*.

[5] Geva, M., Schuster, R., Berant, J., & Levy, O. (2021). Transformer Feed-Forward Layers Are Key-Value Memories. *Proceedings of EMNLP*.

[6] Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research (JAIR)*.

[7] Minervini, P., et al. (2023). Adaptive Rank-based Consistency for Image Generation. *Proceedings of CVPR*.

[8] Garg, V., et al. (2023). What Can We Learn from the Robustness of Large Language Models? *Proceedings of ICML*.

[9] Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

[10] Adebayo, J., et al. (2018). Sanity Checks for Saliency Maps. *Proceedings of NeurIPS*.

[11] Mohammad, S., et al. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. *Proceedings of ACL*.

[12] Wiebe, J., et al. (2004). Learning Subjective Language. *Computational Linguistics*.

[13] Bai, Y., et al. (2022). Constitutional AI: Harmlessness from AI Feedback. *arXiv preprint arXiv:2212.08073*.