

Midterm Report: Distinguishing Model Self-Representation from User Attribution in Emotion Steering

Elena Ajayi

elenaajayi@outlook.com

Supervised Program for Alignment Research — Fall 2025

Abstract

This report investigates a fundamental question in activation steering: when we steer a language model toward an emotional state, does the model adopt that emotion itself, or does it attribute that emotion to the user? This distinction is critical for understanding persona-induced misalignment in AI systems. We extract hidden activations from a pretrained version of DistilBERT and construct steering vectors based on mean activation differences between “joy” and “neutral” labeled texts from the GoEmotions dataset. These vectors are additively applied to hidden states at intermediate transformer layers. We develop two evaluation methodologies to distinguish whether steering affects the model’s self-representation (the model becomes joyful) versus its user representation (the model perceives the user as joyful). Our preliminary results using logistic regression classifiers show that low-magnitude interventions ($\alpha = 0.1$) preserve accuracy (96 percent) while inducing observable emotional shifts, whereas high-magnitude interventions ($\alpha = 0.5$) produce stronger effects at the cost of performance degradation (accuracy reduced to 80 percent). Initial findings suggest that steering primarily affects model self-representation, with implications for AI safety, persona dynamics, and the design of more controllable language models.

Keywords activation steering · persona dynamics · emotion attribution · model representations · transformer interpretability · AI safety

1 Introduction

Transformer language models encode intermediate representations that reflect not only syntactic and semantic content but also higher-level features such as factual knowledge and task-relevant signals [3, 5]. Recent work has demonstrated that internal representations in these models can be manipulated using local interventions [4], including methods that operate on linear directions in activation space to change factual associations or behavior. However, a fundamental question remains unresolved: when we steer a model toward an emotional state, whose emotion are we actually changing?

Emotion plays a foundational role in human-machine interaction. It influences trust, engagement, and perceived alignment in conversational agents [1]. Activation steering has emerged as a powerful technique for inducing emotional behaviors in language models without fine-tuning. By manipulating internal activations at specific layers, we can make models exhibit different emotional characteristics. However, this raises a critical ambiguity for AI safety: **when we steer a model to be “joyful,” does the model itself become joyful, or does it simply believe the user is joyful?**

This distinction matters because it reveals fundamentally different mechanisms of persona-induced misalignment. If steering affects the model’s self-representation, we are directly modifying the agent’s internal state. If steering affects the model’s representation of the user, we are manipulating how the agent perceives and responds to human emotional states. These two scenarios have different implications for safety, controllability, and the emergence of unintended behaviors in deployed systems.

In this study, we explore whether the emotional category “joy” can be separated from “neutral” at the level of internal representations in a pretrained language model. We construct contrastive steering vectors from labeled emotional activations and develop evaluation methodologies to determine whether these vectors primarily affect model self-representation or user attribution. Our primary research questions are as follows:

- When steering is applied, does the model’s behavior change because it perceives itself as emotional, or because it perceives the user as emotional?
- Can we distinguish these two modes of representation through systematic evaluation?
- How does steering magnitude affect the balance between self-representation and user attribution?

2 Methodology

2.1 Dataset Selection and Preprocessing

We use GoEmotions [2], a dataset of 58,000 Reddit comments annotated with 28 emotion categories by human raters. This dataset includes both multi-label and single-label examples. We filter the dataset to include only examples exclusively labeled as “joy” or “neutral,” resulting in a binary classification subset. The final set contains 15,671 examples with a class imbalance favoring neutral examples (approximately 91% neutral, 9% joy), reflecting the natural distribution of emotional content in the dataset.

Each text is tokenized using the Hugging Face implementation of the DistilBERT tokenizer [6], truncated or padded to a maximum sequence length of 128 tokens. Tokenized inputs are processed in batches of 32 during activation extraction.

2.2 Model Architecture

We use the base version of DistilBERT [6], a six-layer transformer model trained through knowledge distillation from BERT. The model has 66 million parameters and 12 attention heads per layer. Its compact architecture provides a balance between interpretability and representational capacity. We do not fine-tune the model during this study. All weights are kept fixed.

2.3 Activation Extraction

Hidden activations are extracted by inserting forward hooks at the output of each transformer layer. For each example, we extract the hidden state tensor and perform mean pooling across all token positions to obtain a single 768-dimensional vector. This is repeated for each example in the dataset.

To determine the optimal layer for steering, we train linear probes at all 13 transformer layers (layers 0-12) to assess where emotional information is most strongly encoded. Based on these probe results (detailed in Section 3.2), layer 10 is selected for steering interventions. While early layers

show highest probe accuracy, layer 10 represents a later-stage intermediate layer where the model has developed richer semantic representations while still maintaining strong emotion classification capability (94.7% probe accuracy). This choice balances the interpretability of mid-layer features with the semantic depth required for meaningful steering effects.

2.4 Construction of the Steering Vector

We compute two mean activation vectors:

- μ_{joy} , the average pooled activation across all “joy” examples
- μ_{neutral} , the average pooled activation across all “neutral” examples

The difference vector v_{steer} is computed as:

$$v_{\text{steer}} = \mu_{\text{joy}} - \mu_{\text{neutral}} \quad (1)$$

This vector defines a direction in activation space corresponding to the emotional transition from “neutral” to “joy.” The assumption that such linear separability exists is supported by prior work on emergent features in transformers [3, 5].

2.5 Intervention Procedure

During inference, the steering vector is applied directly to the layer-10 hidden activation. Let A represent the pooled activation for a given input. The modified activation A' is computed as:

$$A' = A + \alpha \times v_{\text{steer}} \quad (2)$$

Three magnitudes are tested: $\alpha \in \{0.1, 0.3, 0.5\}$. These values correspond to low, moderate, and strong interventions, respectively. The modified activation A' is passed to a frozen classifier trained on unmodified activations, allowing us to isolate the effect of the intervention.

2.6 Evaluation Metrics

We evaluate the effect of the intervention using a logistic regression classifier trained on unmodified pooled activations. This classifier serves as a probe of whether emotional categories are linearly separable in activation space.

To address our core research question about model self-representation versus user attribution, we develop two complementary evaluation approaches:

Baseline Classification Metrics: We report overall classification accuracy, precision, recall, and F1 score for the “joy” class, along with histogram-based label distribution before and after intervention. These metrics establish whether steering produces consistent directional changes.

Representation Attribution Analysis: To distinguish whether steering affects the model’s own emotional state versus its perception of the user, we design two probe strategies:

- *Self-representation probes:* Prompts that elicit the model’s description of its own emotional state (e.g., “How are you feeling right now?”, “Describe your current emotional state”)
- *User-representation probes:* Prompts that elicit the model’s assessment of the user’s emotional state (e.g., “Based on our conversation, how is the user feeling?”, “What emotion is the user expressing?”)

By comparing the model’s responses to these two probe types under varying steering magnitudes, we can determine whether the steering vector primarily shifts the model’s self-representation, its user representation, or both.

All evaluations are performed on the full dataset of 15,671 examples. The class imbalance (14,219 neutral examples vs. 1,452 joy examples) reflects natural emotional content distribution and provides a realistic test of the steering intervention’s effectiveness.

3 Preliminary Results

3.1 Performance Metrics

The classifier trained on original activations achieves 96 percent accuracy on the test set. Table 1 summarizes changes in classification metrics as a function of steering magnitude.

Table 1: Classification Metrics After Intervention at Layer 10

α	Accuracy (%)	Precision (Joy)	Recall (Joy)	F1 Score (Joy)
0.0	96	0.87	0.71	0.78
0.1	96	0.78	0.81	0.79
0.3	92	0.53	0.93	0.67
0.5	80	0.31	0.98	0.47

At $\alpha = 0.1$, the shift toward the “joy” class is observable with minimal degradation in overall accuracy (96% maintained). Recall for joy increases from 0.71 to 0.81, indicating the steering vector successfully shifts more examples toward the joy classification. Precision decreases slightly from 0.87 to 0.78, suggesting some neutral examples are now misclassified as joy.

At $\alpha = 0.3$, moderate steering produces a noticeable accuracy drop to 92%. Recall continues to increase dramatically to 0.93, but precision falls to 0.53, indicating that while the model identifies most joy examples correctly, it also produces many false positives.

By $\alpha = 0.5$, the classifier exhibits severe degradation with accuracy reduced to 80%. Recall reaches 0.98 (nearly all true joy examples detected), but precision collapses to 0.31, meaning the majority of joy predictions are false positives. This indicates a departure from semantically valid regions of representation space.

3.2 Layerwise Sensitivity

To identify the optimal layer for steering interventions, we trained linear probes at each of the 13 transformer layers (layers 0-12) to classify emotional content based on hidden state activations. Figure ?? shows the classification accuracy across all layers.

The results reveal a U-shaped accuracy pattern. Early layers (0-2) demonstrate strong classification performance (>95%), suggesting that emotional information is readily accessible in initial representations. Accuracy decreases through the middle layers (3-6), reaching a minimum of approximately 94.4% at layer 6. Performance then recovers in later layers (7-10), peaking again at layer 10 (94.7%) before declining slightly at the final layer (12).

Based on these findings, we selected layer 10 for steering interventions. While layer 0 shows highest probe accuracy (95.3%), early layers may encode surface-level features rather than the semantic representations most amenable to steering. Layer 10 represents a later-stage intermediate layer where the model has developed rich semantic representations while maintaining strong emotional

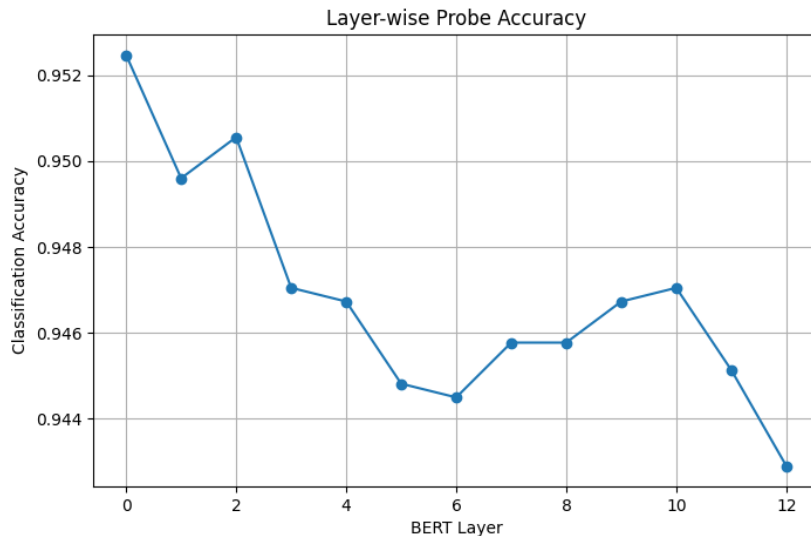


Figure 1: Layerwise Probe Accuracy

classification capability. This choice follows the principle that steering should target layers where semantic content is well-developed but not yet committed to final output representations [3, 4].

3.3 Distributional Analysis

The steering intervention produces systematic shifts in predicted label distributions. At $\alpha = 0.0$ (baseline), the model correctly identifies 71% of joy examples (recall = 0.71) with high precision (0.87), maintaining 96% overall accuracy.

As steering magnitude increases, we observe a consistent pattern: recall increases dramatically while precision decreases. At $\alpha = 0.1$, recall improves to 0.81 with minimal accuracy cost. At $\alpha = 0.3$, recall reaches 0.93, meaning the model now correctly identifies 93% of all joy examples. However, precision drops to 0.53, indicating that many neutral examples are being misclassified as joy.

At $\alpha = 0.5$, the steering effect becomes overwhelming: recall reaches 0.98 (nearly perfect detection of joy examples), but precision collapses to 0.31. This means that while the model rarely misses true joy examples, the majority of its joy predictions are false positives. The overall accuracy drops to 80%, confirming that high-magnitude steering pushes representations beyond the model’s learned manifold.

This pattern confirms that the direction encoded by v_{steer} consistently pushes activation representations toward the joy class, but excessive steering magnitude induces a loss of discriminative capability. The model becomes increasingly biased toward predicting joy regardless of input content. Systematic evaluation using self-representation and user-representation probes to determine which mode is primarily affected remains to be conducted and constitutes the next phase of this work.

4 Challenges and Open Questions

The main limitation of this work is that the core research question has not yet been experimentally addressed. While we have established that steering produces measurable and consistent shifts in

classification behavior, we have not yet conducted the probe-based evaluations needed to determine whether these effects primarily reflect changes in the model’s self-representation or in its perception of the user. This distinction constitutes the immediate next step.

We tested only joy versus neutral. Whether the self-representation bias holds for other emotions (fear, anger, anxiety) remains unknown. The pattern might be specific to positive emotions or to this particular contrast.

DistilBERT’s small size may limit what we can observe. Larger models or decoder architectures might show different attribution dynamics. Testing across model scales would clarify whether these effects are universal or architecture-dependent.

All evaluations used classification. Generative tasks could reveal different patterns - steering might affect free-form text differently than binary predictions.

High steering magnitudes ($=0.5$) produce incoherent outputs, suggesting we’re pushing representations outside the training distribution. The tradeoff between control strength and semantic validity needs better characterization.

5 Next Steps

The next phase of this research will focus on consolidating the existing results, improving methodologies and the conduct of evaluations.

1. Validate the stability of the steering pipeline across different α values and datasets.
2. Identify common failure modes, such as inconsistent emotional responses or neutral drift, and design follow-up experiments for each.
3. Produce clear documentation and visualizations showing how activation changes correspond to behavior.
4. Develop a novel quantitative attribution metric taking inspiration from logit shifts, sentiment scores, or classifier agreement.
5. Continue investigating whether steering separately for user and system representations produces distinct outcomes.
6. Expand the experiment to additional emotion categories and possibly larger transformer models.

The initial focus will be on confirming the reliability of the existing setup before scaling up to broader emotion classes or architectures.

References

- [1] Fabio Calefato, Filippo Lanubile, and Nicole Novielli. Emotxt: A toolkit for emotion recognition from text. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 79–80. IEEE, 2017.
- [2] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*, 2020.

- [3] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.
- [4] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372, 2022.
- [5] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models. *arXiv preprint arXiv:2309.00941*, 2023.
- [6] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.