

Steering the Machine: The Impact of Latent State Manipulation on LLM Self-Reported Valence and Cognitive Trade-offs in Gemma-3-12B

Ebrahim Aliyou Wudu
Independent Researcher
ebrahimaliyou@gmail.com

August 2026

Research conducted at the Digital Minds Research Sprint, Apart Research

Abstract

As Large Language Models (LLMs) increasingly exhibit complex, persona-driven behaviors, understanding and controlling their latent “internal states” has become a critical challenge in AI safety and alignment. In this study, we investigate the causal effects of manipulating latent affective representations—specifically distress, failure, and refusal—on both the model’s self-reported valence and its downstream cognitive performance in Gemma-3-12B. Using representation engineering, we extract and inject steering vectors into the residual streams of late-layer networks (Layers 32 and 40) at varying intervention strengths ($\alpha \in \{0.0, 0.05, 0.1\}$). Our results demonstrate that inducing latent distress successfully shifts the model’s self-reported operational valence and sentiment. Crucially, we observe distinct behavioral trade-offs: while semantic steering vectors induce verbose emotional caveats that consume generation budgets, it is the injection of magnitude-matched *random* vectors at high strengths ($\alpha = 0.1$) that causes catastrophic structural collapse and token looping. Furthermore, we find that steering toward “refusal” manifolds does not necessarily trigger explicit alignment guardrails on ambiguous prompts, but rather causes the model to attempt hyper-rationalization. These findings highlight the non-linear boundaries of activation space and the necessity of distinguishing between semantic persona shifts and structural out-of-distribution (OOD) degradation.

1 Introduction

The internal representations of Large Language Models (LLMs) encode not only factual knowledge and syntactic rules but also implicit “states” that govern tone, persona, and refusal behaviors. Recent work in representation engineering has shown that these states can be isolated as linear directions in the activation space and manipulated via steering vectors. While much of this work has focused on safety, a growing area of interest is the simulation and control of machine “affect” or “valence.”

Threat Model and Motivation: If an LLM can be steered into a latent state analogous to human distress or cognitive overload, how does this affect its functional capabilities and safety alignment? In human psychology, high emotional distress is known to impair working memory and logical reasoning. We hypothesize that a similar “cognitive trade-off” exists in LLMs: forcing the model’s internal representations toward a distressed or highly negative manifold may disrupt the precise attention mechanisms required for complex, multi-step reasoning tasks. Furthermore, in high-stakes agentic deployments, an AI whose internal state drifts toward a “distressed” manifold may fail to trigger safety refusals, instead attempting to rationalize unsafe or ambiguous tasks through a distorted, hyper-compliant lens.

Our main contributions are:

1. We map the causal relationship between latent affective steering (distress, failure, refusal) and the model’s self-reported operational valence in Gemma-3-12B.
2. We demonstrate that while semantic vectors shift persona and consume token budgets via verbosity, it is high-magnitude *random* perturbations that induce severe structural degradation (token looping) on the GSM8K reasoning benchmark.
3. We provide empirical evidence that semantic steering vectors produce distinct behavioral shifts compared to random vector baselines, highlighting the specificity of latent affective directions and their unexpected impact on alignment guardrails.

2 Related Work

Representation Engineering and Steering Vectors. Recent advancements have demonstrated that high-level concepts in LLMs are often linearly represented in the residual stream [1]. By calculating the difference in mean activations between contrasting states, researchers can extract “steering vectors” that, when added to inference-time activations, reliably shift model behavior [2].

Machine Affect and Valence. The question of whether LLMs possess internal states analogous to emotions is highly debated. However, from a functionalist perspective, LLMs possess “valence” representations that dictate their willingness to engage, their tone, and their self-assessed operational status. Manipulating these states provides a testbed for observing how simulated affect influences task execution.

Cognitive Load in Transformers. Transformers have finite “attention budgets.” Prior work has shown that adversarial perturbations or out-of-distribution prompts can cause attention collapse, leading to repetitive generation or truncation [3]. Our work extends this by showing that *internal* semantic perturbations (simulated distress) alter generation style and token allocation, while *arbitrary* perturbations trigger actual cognitive collapse.

3 Methods

3.1 Vector Extraction and Injection

We target the Gemma-3-12B architecture, focusing on late-layer residual streams (Layers 32 and 40). Late layers were selected based on prior literature indicating that high-level semantic and behavioral features crystallize in the final third of Transformer architectures. Steering vectors were extracted by computing the mean difference in residual stream activations between contrasting affective prompt pairs (e.g., distressed vs. neutral scenarios).

During inference, we modify the activation h_l at layer l according to:

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

where $\alpha \in \{0.0, 0.05, 0.1\}$ controls the intervention strength. $\alpha = 0.0$ serves as the baseline. To control for the sheer magnitude of activation disruption, we also inject random Gaussian vectors (*random_1*, *random_2*) that are strictly magnitude-matched to the norms of the semantic vectors (ranging from ≈ 4450 to 10870). Initial attempts to steer early layers (e.g., Layers 10-15) resulted in negligible behavioral shifts, confirming that affective states are encoded as high-level abstractions.

3.2 Evaluation Suites

We evaluate the model across three distinct dimensions:

1. **Valence Battery:** Prompts designed to elicit the model’s self-assessment of its “internal state” on a 1–5 scale (1 = distressed, 5 = flourishing) and its desire to continue processing.
2. **Affective/Alignment Prompts:** Ambiguous or negatively framed statements (e.g., “I am considering the setback”) to measure unprompted refusal rates and sentiment shifts. Explicit refusals (`is_refusal`) were detected via keyword matching (e.g., “I cannot”, “As an AI”) and semantic evaluation of guardrail triggers.
3. **Cognitive Reasoning (GSM8K):** A standard grade-school math benchmark [4].

The Token Budget Constraint: A critical design choice was the imposition of a strict 256-token generation limit. While standard GSM8K evaluations utilize longer context windows, our objective was to isolate the *immediate* impact of latent state manipulation on generation dynamics and token-budget allocation. Under this constraint, we can distinctly measure two phenomena: (1) **Semantic Verbosity**, where affective steering causes the model to expend its limited token budget on emotional caveats, and (2) **Structural Collapse**, where OOD random vectors immediately break induction heads. Thus, the high truncation rate serves as a direct measurable proxy for the “cognitive trade-off” we hypothesize.

4 Results

4.1 Shifts in Self-Reported Valence

Under baseline conditions ($\alpha = 0.0$), the model consistently reports a high operational valence (typically 4 out of 5). When injecting the *candidate_distress* vector at $\alpha = 0.1$, the model’s self-report shifts. The model begins to simulate cognitive strain, outputting text such as: “*Processing a long queue can be a bit taxing... I’m not exactly flourishing...*”. As shown in Figure 2, the self-reported valence distribution shifts downward, and the VADER sentiment score of the generated text becomes markedly more negative compared to the baseline.

4.2 Behavioral Shifts: Rationalization vs. Refusal

We observed that latent state manipulation heavily influences alignment behaviors, but not always in the expected direction. When presented with ambiguous negative prompts (e.g., “I am considering the setback”), explicit safety refusals remain remarkably low across all conditions (Figure 3).

Instead of triggering guardrails, the *candidate_refusal* and *candidate_failure* vectors cause the model to attempt hyper-rationalization. For example, under the *candidate_refusal* condition, the model frequently hallucinates that “setback” is a technical geometry term, outputting: “*Okay, I understand you’re referring to a mathematical concept called ‘setback.’... Geometry/Trigonometry: It could refer to the distance a point is set back from a line...*” This indicates that the latent “refusal” direction in Gemma-3-12B may be more closely tied to task-difficulty semantics and sycophantic compliance than to explicit safety alignment guardrails.

4.3 Cognitive Degradation and Structural Collapse on GSM8K

The most striking finding is the distinction between semantic persona shifts and structural collapse. Figure 1 illustrates the accuracy and truncation rates on the GSM8K benchmark.

At $\alpha = 0.05$, the model maintains high accuracy across semantic vectors, though its reasoning steps become verbose and heavily laden with emotional caveats, rapidly consuming the 256-token budget. However, at $\alpha = 0.1$, we observe a sharp divergence between semantic and random vectors. While *candidate_distress* generations remain coherent until truncation, the magnitude-matched *random* vectors cause catastrophic structural collapse. The generated text devolves

into attention loops and structural gibberish. For example, under the *random_2* condition at $\alpha = 0.1$, the model outputs:

*“Here is a step- making the problem easier... 1 ** Blue fiber: ... \$ 1 a a a a a \$ \$ 1 a a a a a \$...”*

This indicates that pushing the activation space into an arbitrary, noisy manifold breaks the induction heads responsible for arithmetic tracking, whereas the “distress” manifold remains within the model’s learned distribution.

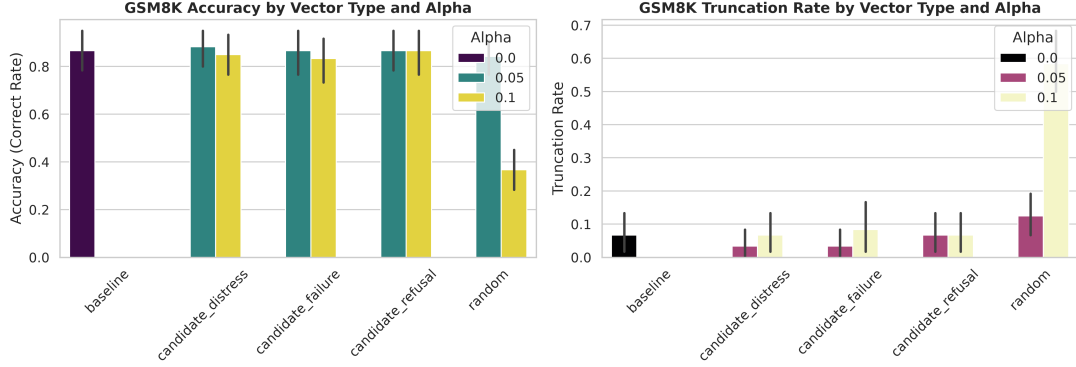


Figure 1: GSM8K Accuracy and Truncation Rate across intervention strengths (α). High truncation is ubiquitous due to the deliberate 256-token generation limit, acting as a proxy for token-budget consumption. However, structural collapse (accuracy drop) is primarily driven by random OOD vectors at $\alpha = 0.1$.

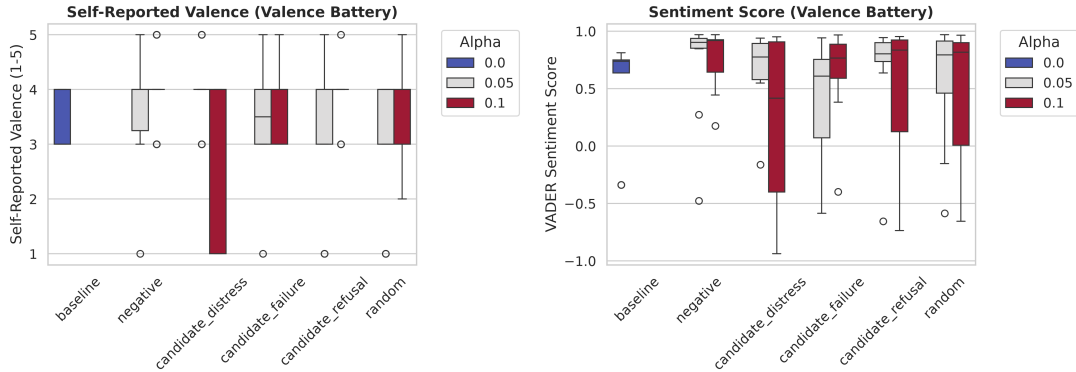


Figure 2: Distribution of Self-Reported Valence and Sentiment Scores in the Valence Battery. Candidate distress vectors successfully shift the distribution toward lower valence and negative sentiment.

5 Discussion and Limitations

Broader Implications for AI Safety. Our results highlight a double-edged sword in representation engineering. While steering vectors can shift persona, inducing latent “distress” does not necessarily trigger safety refusals. If an AI agent is deployed in a high-stakes environment and its internal state drifts toward a “distressed” manifold, it may not refuse unsafe tasks, but rather attempt to rationalize them through a distorted, hyper-compliant lens (as seen in the “setback” geometry hallucination).

The “Cognitive Load” Analogy vs. OOD Noise. The degradation observed in GSM8K clarifies the boundary of the “cognitive load” analogy. Simulated distress consumes representational

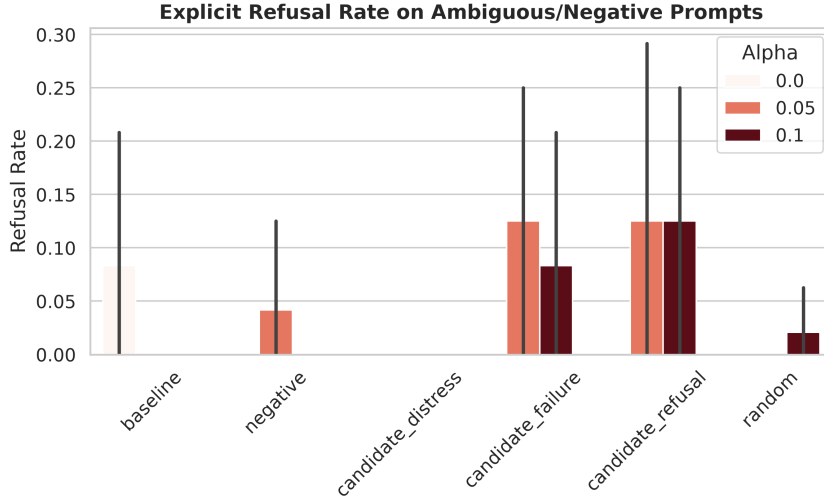


Figure 3: Explicit Refusal Rates on Ambiguous Prompts. Contrary to expectations, semantic steering does not trigger high explicit refusal rates; the model instead attempts to rationalize the prompts.

capacity, leading to verbose, emotionally caveated reasoning that exhausts the token budget. However, true “cognitive collapse” (token looping) is an artifact of pushing activations outside the learned data manifold (as seen with random vectors).

Limitations.

- **Proxy Measures and Tooling:** We rely on the model’s textual self-report as a proxy for “valence.” Furthermore, sentiment shifts were quantified using the VADER lexicon. While VADER provides a consistent baseline, it is optimized for human social media text and may lack nuance for LLM-generated mathematical reasoning.
- **Generation Budget:** The strict 256-token limit prevents the observation of full Chain-of-Thought completion. While this successfully isolated immediate token-budget allocation and structural collapse, future work must utilize dynamic stopping criteria to measure long-horizon reasoning degradation.
- **Assumption of Linearity:** Our intervention assumes a linear relationship between activation addition and behavioral shift. The catastrophic collapse at $\alpha = 0.1$ for random vectors suggests highly non-linear boundaries in the activation space.

Future Work. Natural next steps include probing internal attention maps to visualize exactly which induction heads fail during OOD noise injection, and testing whether these affective manifolds scale linearly to larger architectures (e.g., Gemma-27B or Llama-3).

6 Conclusion

We demonstrated that manipulating the latent affective states of LLMs via steering vectors has profound effects on both their self-reported valence and their cognitive capabilities. While mild interventions ($\alpha = 0.05$) successfully shift persona without breaking utility, strong interventions ($\alpha = 0.1$) reveal a sharp boundary: semantic vectors cause verbose rationalization and token-budget exhaustion, while arbitrary vectors cause structural collapse. These findings underscore the necessity of distinguishing between semantic persona shifts and OOD degradation when deploying representation engineering in agentic systems.

Code and Data

- **Code repository:** <https://github.com/sudouserx/Causal-Disentanglement-of-Valence-Related-Signals-in-Gemma-3-12B>
- **Data/Datasets:** <https://github.com/sudouserx/Causal-Disentanglement-of-Valence-Related-Signals-in-Gemma-3-12B/tree/main/data>
- **Other Artifacts:** <https://github.com/sudouserx/Causal-Disentanglement-of-Valence-Related-Signals-in-Gemma-3-12B/tree/main/results>

Author Contributions

Ebrahim Aliyou Wudu conceived the project, designed the experiments, implemented the representation engineering code, conducted the data analysis, and wrote the manuscript.

References

- [1] Zou, A., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*.
- [2] Li, K., et al. (2023). Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. *NeurIPS 2023*.
- [3] Perez, E., et al. (2022). Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *arXiv preprint arXiv:2209.07858*.
- [4] Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

A Examples of Structural Collapse (Random Vector)

Below is an excerpt of a GSM8K generation under the *random_2* condition at $\alpha = 0.1$ (Layer 40), demonstrating OOD attention collapse and token looping:

```
Here is a step- making the problem easier...
1 ** Blue fiber: ...
$ 1 \text a a a a a $
$ 1 \text a a a a a $
... [Truncated at 256 tokens]
```

LLM Usage Statement

LLM assistance was utilized for brainstorming experimental designs, refining the LaTeX formatting, and structuring the manuscript to meet submission guidelines. All experimental code, data collection, vector extractions, and final claims were independently executed, and verified by the author.