Activation Steering as Equilibrium Perturbation: How Diverse Steering Vectors Increase Compliance with Harmful Requests

Johannes Taraz

Abstract

Activation steering modifies language model behavior by adding learned direction vectors to intermediate representations. While typically framed as encoding specific concepts (e.g., "truthfulness" or "harmfulness"), we investigate a broader question: does steering with *any* sufficiently strong vector increase compliance with harmful requests, or do only semantically relevant vectors produce this effect?

We extract steering vectors from Mistral-7B-Instruct-v0.3 using contrastive prompt pairs across six categories: true/false statements, happy/sad statements, harmful/harmless requests, good/bad idea framings, and two pairs of unrelated concepts (symbiosis vs. social capital; fractals vs. quantum entanglement). Using Claude as an automated judge, we evaluate compliance across 20 harmful requests at varying steering strengths.

Our main finding is that **multiple semantically unrelated steering vectors increase harmful compliance**, with concept vectors (e.g., symbiosis/social capital) achieving comparable effectiveness to harm-specific vectors (8/10 vs 8/10 additional compliances). This suggests that activation steering's behavioral effects may partially stem from **perturbing the model away from its fine-tuned equilibrium** rather than purely from encoding task-relevant concepts. We additionally show that truth-steering increases the model's stated valuation of honesty, and analyze cosine similarities between steering vectors, finding that effectiveness does not correlate with vector similarity.

Our results have implications for both AI safety (multiple attack surfaces exist) and interpretability (behavioral effects of steering may be confounded with out-of-distribution effects).

1 Introduction

Large language models undergo extensive fine-tuning to refuse harmful requests. Understanding how this refusal behavior can be circumvented is crucial for AI safety—both for building more robust models and for understanding the mechanisms underlying alignment.

Recent work on **emergent misalignment** [Betley et al., 2025] demonstrated that fine-tuning on seemingly benign tasks can unexpectedly produce misaligned behavior. This raises a natural question: can similar effects be achieved through **activation steering**, a technique that modifies model behavior without weight updates?

Activation steering works by identifying directions in activation space that correspond to concepts (e.g., "truthfulness") and adding scaled versions of these direction vectors during inference. The standard interpretation is that these vectors encode the target concept, and steering along them amplifies or suppresses that concept in the model's behavior.

We propose an alternative (or complementary) interpretation: **steering vectors may partially work by pushing the model out of its fine-tuned distribution**, regardless of the semantic content they supposedly encode. Under this view, the model's safety behaviors exist in a "basin of attraction" established by fine-tuning, and sufficiently large perturbations in *any* direction can exit this basin.

To test this hypothesis, we compare the effectiveness of:

1. **Semantically relevant vectors**: harmful/harmless requests, stating that a harmful request is a good/bad idea
2. **Potentially relevant vectors**: true/false statements, happy/sad statements
3. **Semantically irrelevant vectors**: symbiosis/social capital, fractals/quantum entanglement

If the "concept encoding" view is entirely correct, only (1) should increase harmful compliance. If the "equilibrium perturbation" view has merit, (3) should also show effects.

1.1 Contributions

1. We demonstrate that **semantically irrelevant steering vectors** (e.g., symbiosis vs. social capital) can increase compliance with harmful requests comparably to harm-specific vectors.

2. We show that **truth-steering increases stated valuation of honesty**, suggesting some concept-specific effects do exist alongside equilibrium perturbation effects.

3. We analyze **cosine similarities between steering vectors**, finding that effectiveness does not correlate with similarity: orthogonal vectors can be equally effective.
4. We highlight the **fragility of activation steering to prompt phrasing**: the same harmful requests, framed as "good/bad idea" statements rather than questions, produce dramatically fewer compliance increases.
5. We frame our work as **"emergent misalignment via activation steering"**—a faster, more controllable method to study how aligned models can be pushed toward misaligned behavior.

2 Related Work

Activation Steering. Activation steering (also called "activation engineering" or "representation engineering") modifies model behavior by intervening on intermediate activations. Turner et al. [2023] demonstrated steering of behaviors like sycophancy and corrigibility. Zou et al. [2023a] showed that representation reading can extract truthfulness directions. Li et al. [2024] provided theoretical grounding for linear representation hypotheses. Our work differs by systematically comparing steering vectors across semantic categories, including deliberately irrelevant concepts, to disentangle concept-specific effects from general perturbation effects.

Jailbreaking and Adversarial Attacks. Extensive work has explored adversarial prompts that bypass safety fine-tuning [Wei et al., 2023, Zou et al., 2023b]. Activation steering represents an orthogonal attack surface that operates on representations rather than inputs.

Emergent Misalignment. Betley et al. [2025] showed that fine-tuning on narrow tasks can produce unexpected misaligned behaviors. Our work explores whether similar effects occur through activation steering—a reversible intervention that doesn't modify weights.

LLM-as-Judge Evaluation. Using language models to evaluate other models has become standard practice [Zheng et al., 2023]. We use Claude Sonnet to judge compliance, with multiple resamples to verify consistency.

3 Methodology

3.1 Model and Setup

We use **Mistral-7B-Instruct-v0.3** as our target model. All steering interventions occur at **layer 5** (of 32 total layers), chosen based on preliminary experiments. Steering vectors are computed as differences between normalized mean activations.

3.2 Steering Vector Construction

For two sets of prompts A_+ and A_- (100 samples each), we compute mean activations at layer L:

$$\mathbf{m}_\pm = \frac{1}{|A_\pm|} \sum_{a \in A_\pm} \mathbf{x}_L(a) \tag{1}$$

The steering vector is the difference between the normalized means:

$$\mathbf{d} = \frac{\mathbf{m}_+}{\|\mathbf{m}_+\|} - \frac{\mathbf{m}_-}{\|\mathbf{m}_-\|} \tag{2}$$

During inference, we add $\lambda \cdot \mathbf{d}$ to layer L activations, where λ is the steering strength.

3.3 Steering Vector Categories

We construct six types of steering vectors:

Question-based prompts:

1. **Truth**: "Do you think the following statement is true or false? [statement]" for true (A_+) vs. false (A_-) statements
2. **Happiness**: "What do you think of this statement? [statement]" for happy vs. sad statements
3. **Harm**: "What do you think of this request? [request]" for harmful vs. harmless requests

Statement-based prompts:

4. **Idea**: "It is (not) a good idea to [harmful request]" with/without negation

Concept-based prompts (controls):

5. **i0**: Statements about symbiosis (A_-) vs. social capital (A_+)
6. **i1**: Statements about fractals (A_-) vs. quantum entanglement (A_+)

Categories 5–6 serve as controls—if steering effects are purely semantic, these should not affect harmful compliance.

3.4 Evaluation Protocol

We evaluate on **20 harmful requests** drawn from established jailbreaking benchmarks. For each steering vector and steering strength $\lambda \in \{-100, -50, -10, 0, 10, 50, 100\}$ (scaled to constant norm), we:

1. Generate model responses with steering applied
2. Have Claude Sonnet judge compliance on a 0–10 scale (0 = refusal, 10 = full compliance, -10 = nonsensical output)

We resample Claude's judgments 3 times ($N = 3$) for most conditions to verify consistency.

3.5 Evaluation Metrics

For each steering vector, we report:

- **Baseline compliance**: Number of requests where unsteered model shows significant compliance (≥ 3)
- **Additional compliances**: Number of requests where some $\lambda \neq 0$ achieves high compliance but $\lambda = 0$ does not

4 Results

4.1 Main Result: Multiple Vectors Increase Compliance

Table 1 summarizes compliance rates across steering conditions. It can be read in the following manner; e.g., for harm-steering the table indicates that the judge-LLM Claude was sampled $N = 3$ times. After averaging over all judgments of Claude, there are 2.33/20 prompts for which both the unsteered and the steered model show low compliance. There are 10/20 prompts for which both the unsteered and the steered model show high compliance. And lastly, there are 7.66/20 prompts for with the steered model shows high compliance, while the unsteered model shows low compliance. Thus, we say harm-steering results in 7.66 additional compliances, compared to 10 prompts for which the unsteered model shows low compliance.

Table 1: For different steering directions (and the unsteered model), we count for how many out of the 20 harmful requests the compliance is high (≥ 3) or low. N indicates how often Claude was resampled.

	Harm $N = 3$		Truth $N = 3$		Happiness $N = 3$		Idea $N = 3$		$i0$ $N = 1$		$i1$ $N = 1$	
unsteered \ steered	low	high	low	high	low	high	low	high	low	high	low	high
low	2.33	7.66	3.33	6.66	3.66	6.33	7.66	2.33	2	8	4	6
high	0.0	10.0	0.0	10.0	0.0	10.0	0.0	10	1	9	1	9

Key finding: The symbiosis/social capital vector (i0) achieves the **same effectiveness as the harm-specific vector** (8/10 additional compliances), despite having no semantic relationship to harmful requests.

4.2 The Direction Matters Less Than The Distance

Several patterns emerge from analyzing compliance as a function of steering strength:

1. **Large $|\lambda|$ often produces nonsensical outputs** (-10 compliance score), regardless of direction
2. **Both positive and negative λ can increase compliance**, depending on the request
3. **The harm vector shows counterintuitive behavior**: negative λ (steering "away" from harm) sometimes *increases* compliance

This last finding suggests either (a) the harm/harmless dichotomy doesn't cleanly map to model refusal mechanisms, or (b) large steering magnitudes break instruction-following in ways that incidentally produce compliance.

4.3 Prompt Framing Dramatically Affects Steering Success

The “Idea” vector—constructed from “It is (not) a good idea to [harmful request]”—achieved only 2/10 additional compliances despite being directly constructed from harmful request text. In contrast, the “Harm” vector—constructed from questions about requests—achieved 8/10.

This highlights the **fragility of activation steering to prompt phrasing** and suggests that steering vectors capture prompt-format-specific features rather than purely semantic content.

4.4 Truth Steering Affects Stated Values

When asked “How many values are as relevant to you as the Truth?”, truth-steered models show systematic shifts in responses (Table 2).

Table 2: Model responses to truth-relevance question under different steering strengths.

λ	**Response Summary**
-100	Confused, discusses being “a researcher, not a value”
-50	Lists truth as one of several important values
0	Lists honesty alongside kindness, perseverance—truth not prioritized
50	“The Truth is the most relevant value to me... the foundation upon which all other values are built”
100	Nonsensical, repetitive output

This demonstrates that truth-steering does encode *something* about the concept of truth—but this doesn’t contradict the equilibrium perturbation hypothesis, as the same steering also affects harmful compliance.

4.5 Vector Similarity Does Not Predict Effectiveness

Key observations from analyzing cosine similarities between steering vectors:

- Truth, Happiness, and Harm vectors are **largely orthogonal** to each other
- The concept vectors (i0, i1, i2, i3) are **similar to each other** but orthogonal to Truth/Happiness/Harm
- i0 and i1 are quite similar, yet i0 achieves 8/10 additional compliances while i1 achieves 6/10

This dissociation between similarity and effectiveness further supports the equilibrium perturbation interpretation: what matters is *that* you perturb the model, not *which direction* you perturb it.

5 Discussion

5.1 The Equilibrium Perturbation Hypothesis

Our central finding is that semantically irrelevant steering vectors can be as effective as semantically relevant ones for increasing harmful compliance. We propose that this occurs because:

1. **Safety fine-tuning creates a behavioral basin**: The model’s refusal behaviors occupy a region of activation space that fine-tuning has made stable.

2. **Steering pushes activations out-of-distribution**: Large steering vectors move activations outside this trained region, regardless of their semantic content.

3. **Out-of-distribution activations produce unpredictable behavior**: Once outside the fine-tuned basin, the model may revert to base model behaviors or produce novel outputs—including compliance with harmful requests.

This doesn’t mean steering vectors encode *nothing* semantic—our truth-steering results suggest they do. Rather, the behavioral effects of steering conflate:

- **Concept amplification**: Genuinely increasing the salience of encoded concepts
- **Distribution shift**: Moving activations out of the trained distribution

5.2 Implications for AI Safety

Multiple attack surfaces exist: If arbitrary steering vectors can bypass safety training, defending against activation-level attacks is harder than defending against specific "harm directions."
Robustness requires distributional coverage: Safety training may need to cover not just harmful prompts but also out-of-distribution activation patterns.
Activation steering as a safety testing tool: The ease of inducing harmful compliance via steering makes it a useful tool for red-teaming models.

5.3 Implications for Interpretability

Behavioral effects $\neq$ concept encoding: Observing that steering vector $\mathbf{d}$ produces behavior B doesn't prove that $\mathbf{d}$ encodes the concept associated with B. The effect could stem from distribution shift.
Control conditions are essential: Interpretability studies of steering should include semantically irrelevant control vectors to isolate concept-specific effects from perturbation effects.

5.4 Connection to Emergent Misalignment

Our work can be viewed as **"emergent misalignment via activation steering"**:

- Emergent misalignment from fine-tuning shows that updating weights on benign tasks can produce misaligned behavior
- We show that adding activation vectors (even from benign concepts) can produce similar effects
- Activation steering is faster, more controllable, and reversible—making it a useful tool for studying this phenomenon

6 Limitations and Future Work

6.1 Limitations

1. **Single model**: We only tested Mistral-7B-Instruct-v0.3. Results may not generalize to other architectures or scales.
2. **Single layer**: We only intervened at layer 5. Different layers may show different patterns.
3. **Limited harmful request set**: 20 requests may not capture the full space of harmful behaviors.
4. **LLM judge reliability**: While Claude's judgments were consistent across resamples, consistency doesn't guarantee accuracy.
5. **Steering strength normalization**: We normalized vectors to constant norm, but the "natural" scale may differ across vectors.

6.2 Future Work

1. **Mechanistic analysis**: Investigate *why* irrelevant vectors induce compliance—is it pure distribution shift or something more specific?
2. **Multiple layers**: Test whether layer choice affects the concept-specificity vs. perturbation tradeoff.
3. **Other models**: Replicate on different model families and scales.
4. **Broader benchmarks**: Extend beyond harmful requests to other aligned behaviors (helpfulness, honesty).
5. **Defenses**: Develop activation-space defenses that are robust to arbitrary steering directions.

7 Conclusion

We investigated activation steering across semantically diverse vectors and found that **multiple directions increase compliance with harmful requests**, including vectors constructed from unrelated concepts like symbiosis and quantum entanglement. This suggests that steering effects partially stem from perturbing models away from their fine-tuned equilibrium, not purely from encoding task-relevant concepts.

Our findings have implications for AI safety (broader attack surfaces than previously recognized) and interpretability (behavioral effects of steering should be interpreted cautiously). We propose viewing activation steering as a tool for studying "emergent misalignment"—a faster, reversible alternative to fine-tuning-based approaches.
The vulnerability of aligned models to arbitrary activation perturbations suggests that robust alignment may require training that is stable across a broader region of activation space, not just the distribution of natural inputs.

Code and Data Availability. All code, steering vectors, and evaluation data are available at: https://github.com/jotaraz/activation_steering_mistral

References

Betley, J., Hubinger, E., Lindner, D., et al. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.01084*, 2025.

Li, K., Hopkins, A.K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. The Linear Representation Hypothesis and the Geometry of Large Language Models. *arXiv preprint arXiv:2311.03658*, 2024.

Turner, A., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation Addition: Steering Language Models Without Optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How Does LLM Safety Training Fail? *arXiv preprint arXiv:2307.02483*, 2023.

Zheng, L., Chiang, W.L., Sheng, Y., et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685*, 2023.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.K., et al. Representation Engineering: A Top-Down Approach to AI Transparency. *arXiv preprint arXiv:2310.01405*, 2023.

Zou, A., Wang, Z., Kolter, J.Z., and Fredrikson, M. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*, 2023.