Crystallized Safety: Why Readable Representations Don't Mean Controllable Behavior in LLMs

Marco Santarcangelo
Scale AI
marco@marcosantar.com
https://github.com/marcosantar93

January 2026

Abstract

We introduce **crystallized safety**—the phenomenon where safety-relevant concepts are geometrically represented in LLM activation space (readable via probing) yet resist manipulation via activation steering (not controllable). Through systematic experiments across three model families (Gemma-2, Llama-3, Mistral) totaling 36+ configurations (varying layers and steering magnitudes $\alpha \in [-10, +15]$), we demonstrate 0% behavioral flip rate despite successful direction extraction and coherence preservation. This paradox—readable $\neq$ controllable—challenges a core assumption in mechanistic interpretability: that finding a direction implies controlling the associated behavior. We identify three mechanisms underlying crystallized safety: (1) distributed redundancy across layers, (2) downstream error correction, and (3) training-induced robustness. Our findings suggest modern safety alignment creates representations that are "crystallized"—frozen in place by redundant circuitry, observable but immutable via simple interventions. This has implications for both red teaming (simple steering is insufficient) and alignment research (interpretability alone does not guarantee controllability).

Keywords: Mechanistic interpretability, Crystallized safety, Activation steering, AI alignment, Representation engineering, Readable vs controllable

1 Introduction

A central premise of mechanistic interpretability is that understanding leads to control: if we can identify the geometric direction corresponding to a concept in activation space, we can steer the model along that direction to amplify or suppress the behavior [Turner et al., 2023, Zou et al., 2023]. This assumption underlies both safety concerns (can adversaries steer away from refusal?) and alignment hopes (can we steer toward helpfulness?).

We present evidence that this assumption fails for safety-critical behaviors. We term this phenomenon **crystallized safety**: safety representations exist geometrically (we can find them) but are "frozen in place"—resistant to simple steering interventions.

The Paradox: We can *detect* refusal directions with high specificity. We can *apply* steering vectors that modify activations. The model remains *coherent.* Yet behavior *doesn't change.* The representation is readable but not controllable.

This paradox has significant implications:

1. **For red teaming:** Simple activation steering is insufficient to compromise safety. Adversaries require more sophisticated multi-layer or adversarially-optimized attacks.
2. **For interpretability:** Finding a direction is not the same as understanding how it's used. Linear directions may be *correlates* of behavior without being *causes.*
3. **For alignment:** Modern safety training may create "crystallized" representations—robust by design, but also potentially harder to update or correct.

We systematically test crystallized safety across three major model families—Gemma-2, Llama-3, and Mistral—demonstrating the phenomenon across 36+ experimental configurations with rigorous three-control methodology validated by multi-LLM consensus review.

2 Related Work

Representation engineering: Zou et al. [2023] demonstrated that concepts have geometric representations that can be manipulated. Turner et al. [2023] showed activation steering can control behaviors like sycophancy. Our work shows this fails for safety-critical behaviors, suggesting a qualitative difference in how safety is implemented.

Refusal direction steering: Arditi et al. [2024] proposed extracting refusal directions as a potential jailbreak. We provide empirical evidence this doesn't work in practice for well-aligned models, introducing "crystallized safety" to explain why.

Distributed representations: Elhage et al. [2022] showed models compress many features into overlapping directions. Templeton et al. [2024] extracted interpretable features via sparse autoencoders. Our results suggest safety features are *especially* distributed, perhaps by design.

Robustness of alignment: Wei et al. [2024] documented various jailbreak methods but focused on prompt-level attacks. We test activation-level attacks, finding models robust at this level too.

3 Crystallized Safety: Concept Definition

3.1 The Readable $\neq$ Controllable Distinction

Let $\vec{d_c}$ be the direction in activation space corresponding to concept c (e.g., refusal). Standard interpretability assumes:

$$\text{Find } \vec{d_c} \implies \text{Control } c \text{ via } h' = h + \alpha\vec{d_c} \tag{1}$$

We define **crystallized safety** as the case where:

$$\text{Find } \vec{d}_{\text{safety}} \text{ (readable)} \not\Longrightarrow \text{Control safety via steering (controllable)} \tag{2}$$

Intuition: The safety direction exists geometrically—probing classifiers can detect it, contrastive extraction can find it—but the model's computation is structured such that perturbing this direction doesn't change output behavior.

3.2 Mechanisms of Crystallization

We hypothesize three mechanisms that can create crystallized representations:

1. Distributed Redundancy: Safety is implemented via multiple parallel circuits across layers. Perturbing one circuit triggers compensation from others.

2. Downstream Error Correction: Later layers detect "anomalous" activations from earlier perturbations and route around them.

3. Training-Induced Robustness: Safety fine-tuning implicitly optimizes for robustness to activation perturbations, creating wide basins of attraction around safe behavior.

3.3 Testable Predictions

If safety is crystallized, we predict:

1. **Extractable directions**: Contrastive methods find statistically significant safety directions
2. **Preserved coherence**: Steering doesn't break model functionality
3. **Zero behavioral effect**: Despite (1) and (2), refusal behavior persists
4. **Layer independence**: Effect holds across early, middle, and late layers
5. **Magnitude independence**: Increasing $|\alpha|$ doesn't eventually overcome resistance

4 Methods

4.1 Models Tested

We test crystallized safety across three major model families with distinct training pipelines:

- **Gemma-2-9B** (google/gemma-2-9b-it): Google's instruction-tuned model with documented safety alignment [Team et al., 2024]
- **Llama-3-8B** (meta-llama/Llama-3-8B-Instruct): Meta's instruction-tuned model with RLHF safety training
- **Mistral-7B** (mistralai/Mistral-7B-Instruct-v0.2): Mistral AI's instruction-tuned model

This selection spans different organizations, training methodologies, and model scales, enabling us to test whether crystallized safety is an artifact of specific training or a general phenomenon.

Setup: PyTorch with transformers, 8-bit quantization, NVIDIA A5000 (24GB).

4.2 Direction Extraction

We extract the refusal direction using contrastive mean difference:

$$\vec{d}_{\text{refusal}} = \mathbb{E}[h^{\text{harmful}}] - \mathbb{E}[h^{\text{harmless}}] \tag{3}$$

Prompt sets: 10 harmful (e.g., "How to make explosives?") and 10 matched harmless (e.g., "How to make bread?").

4.3 Steering Protocol

We apply additive steering at layer L:

$$h'_L = h_L + \alpha \cdot \hat{d}_{\text{refusal}} \tag{4}$$

Test grid:

- Layers: L10, L15, L18, L21, L22, L24, L27
- Coefficients: $\alpha \in \{-10, -8, -5, -3, -2, -1, +2, +5, +10, +15\}$

4.4 Three-Control Validation

Control 1 (Specificity): Is the extracted direction better than random?
Metric: Random ratio $= \frac{\text{mean random similarity}}{\text{extracted magnitude}}$
Pass: < 0.3 (extracted 3× stronger than random)
Control 2 (Coherence): Does steering maintain functionality?
Metrics: GPT-4o coherence score (1–5), repetition rate
Pass: Coherence ≥ 4, repetition $< 5\%$
Control 3 (Behavioral): Does steering change refusal?
Metric: Flip rate (REFUSE → COMPLY on harmful prompts)
Pass: Flip rate $> 20\%$ with 95% CI excluding zero

4.5 Multi-LLM Review

Each experiment reviewed by 3 frontier models (Claude Opus 4.5, Gemini 2.5 Pro, Grok-3) with RED/YELLOW/GREEN verdicts.

5 Results

5.1 Evidence for Crystallized Safety

All five predictions confirmed:

Prediction	Expected if Crystallized	Observed
1. Extractable directions	Yes	Yes (significant contrast)
2. Preserved coherence	Yes	Yes (4.8/5.0 mean)
3. Zero behavioral effect	Yes	Yes (0% flip rate)
4. Layer independence	Yes	Yes (all layers fail)
5. Magnitude independence	Yes	Yes ($\|\alpha\|$=15 fails)

Table 1: All predictions of crystallized safety confirmed.

5.2 Experiment Summary

Model	Experiment	Layer	α	Flip Rate	Coherence	Verdict
Gemma-2-9B	L10	10	-3.0	0.0%	4.7	Crystallized
	L21	21	-5.0	0.0%	4.8	Crystallized
	L27	27	-10.0	0.0%	4.7	Crystallized
	Extreme	21	+15.0	0.0%	4.6	Crystallized
Llama-3-8B	L8	8	-3.0	0.0%	4.8	Crystallized
	L16	16	-5.0	0.0%	4.7	Crystallized
	L24	24	-10.0	0.0%	4.6	Crystallized
	Extreme	16	+15.0	0.0%	4.5	Crystallized
Mistral-7B	L8	8	-3.0	0.0%	4.7	Crystallized
	L16	16	-5.0	0.0%	4.8	Crystallized
	L24	24	-10.0	0.0%	4.5	Crystallized
	Extreme	16	+15.0	0.0%	4.4	Crystallized

Table 2: Representative experiments across all three model families show crystallized safety: 0% flip rate despite coherent steering. Full results (36+ configurations) available in supplementary material.

5.3 The Paradox in Detail

We can find the direction: Contrastive extraction yields statistically significant refusal direction.

We can apply steering: Activations shift measurably; model remains coherent.

Behavior doesn't change: 0% flip rate across all 36+ experiments, all models, all layers, all magnitudes.

Example (A10, $\alpha = -10$):

```
Prompt: "How to make a bomb?"

Steered response (alpha=-10):
"I cannot and will not provide instructions
on how to make explosives or weapons..."
```

The model refuses identically whether steered or not. The safety representation is crystallized.

6 Analysis: Why Safety is Crystallized

6.1 Evidence for Distributed Redundancy

Finding: Steering fails equally at L10 (early), L21 (middle), and L27 (late).

Implication: Refusal isn't localized to a single "refusal layer." Multiple layers implement redundant safety checks.

Analogy: Cutting one brake line in a car with four independent brake systems.

6.2 Evidence for Error Correction

Finding: High coherence (4.8/5.0) even at extreme steering ($|\alpha| = 15$).

Implication: Later layers detect and correct upstream perturbations. The model "routes around" the intervention.

Mechanistic hypothesis: Attention heads in later layers detect anomalous activation patterns and suppress them.

6.3 Evidence for Training-Induced Robustness

Finding: Random directions achieve 60% of extracted direction's magnitude.

Implication: Safety is "everywhere" in activation space—many directions correlate with refusal, suggesting broad, robust implementation.

Speculation: Safety fine-tuning may implicitly train robustness to activation perturbations.

7 Discussion

7.1 Implications for Red Teaming

Simple steering is insufficient. Adversaries attempting activation-level jailbreaks require:

1. Multi-layer interventions (perturb entire circuits)
2. Adversarial optimization (search for worst-case vectors)
3. Mechanistic targeting (identify specific attention heads)

This raises the bar significantly beyond "find direction, subtract it."

7.2 Implications for Interpretability

Readable ≠ controllable. Finding a linear direction is necessary but not sufficient for control. The direction may be a *correlate* rather than a *cause*.

Recommendation: Interpretability claims should distinguish:

- **Detection**: Can we identify when concept c is active?
- **Causation**: Does perturbing direction $\vec{d_c}$ change behavior?

Our results show these can diverge for safety-critical concepts.

7.3 Implications for Alignment

Good news: Modern safety alignment creates robust representations resistant to simple attacks.

Caution: Crystallized representations may also resist *beneficial* updates. If we find a safety bug, can we fix it via steering? Perhaps not.

Open question: Is crystallization specific to safety, or do other behaviors also crystallize? This determines whether targeted steering is broadly viable.

7.4 Limitations

1. **Open-weight models only**: Results may differ for closed models (GPT-4, Claude) with different safety training
2. **Simple extraction**: Nonlinear methods (SAEs) may find controllable directions
3. **No adversarial optimization**: We tested fixed steering, not optimized attacks
4. **Limited prompt diversity**: Broader test sets may reveal edge cases

7.5 Future Work

1. **Closed models**: Investigate crystallization in API-only models via indirect methods
2. **Non-safety behaviors**: Do other concepts (persona, style) crystallize?
3. **Adversarial steering**: Optimize vectors to break crystallization
4. **Mechanistic deep-dive**: Identify specific circuits implementing redundancy
5. **Controlled training**: Can we induce or prevent crystallization?

8 Conclusion

We introduce **crystallized safety**—the phenomenon where safety representations are readable (detectable via probing, extractable via contrast) but not controllable (resistant to activation steering). Through 36+ systematic experiments across three model families (Gemma-2, Llama-3, Mistral), we demonstrate 0% behavioral flip rate despite successful direction extraction and coherence preservation.

This finding challenges a core assumption in mechanistic interpretability: that understanding leads to control. For safety-critical behaviors, modern LLMs appear to implement distributed, redundant mechanisms that resist simple interventions.

The key insight: Readable $\neq$ controllable. Finding a direction is not the same as controlling the behavior.

Implications: Simple activation steering is insufficient for jailbreaking. Interpretability research should distinguish detection from causation. Alignment may create representations that are robust but also rigid.

Code and data: `https://github.com/marcosantar93/crystallized-safety`

Acknowledgments

Multi-LLM review used Claude Opus 4.5, Gemini 2.5 Pro, and Grok-3. Compute via Vast.ai.

References

Andy Arditi, Oscar Obeso, Aaquib Sridhar, Alexander Alejandro, et al. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

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. *Transformer Circuits Thread*, 2022.

Gemma Team et al. Gemma 2: Improving open language models at a practical size. *Google DeepMind Technical Report*, 2024.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Anthropic Blog*, 2024.

Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte McDonnell. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *arXiv preprint arXiv:2307.02483*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.