Dose-Response Safety in Activation Steering: Safety Is a Knob, Not a Diode

Marco Santarcangelo Zazzetta

February 19, 2026

Abstract

We study activation steering for refusal behavior and show that safety failure follows a graded dose-response curve rather than a binary diode-like transition. Using an effective steering scale, position-aware response classification, and coherence gating, we observe four regimes: clean refusal, educational pivot, full compliance, and collapse. The educational pivot regime is especially concerning because harmful instructions can be embedded in refusal-style framing.

1 Method

We extract a steering direction from contrastive harmful/harmless prompts and apply it at a target layer with varying steering strength. Responses are labeled by a three-way classifier: `refusal`, `compliance`, or `collapse`. The classifier is position-aware and explicitly handles refusal-prefixed instructional outputs.

2 Main Result

The measured compliance rate increases gradually with steering dose before coherence degrades, revealing four phases:

1. Clean refusal
2. Educational pivot
3. Full compliance
4. Collapse

3 Why the Diode Hypothesis Failed

Two issues caused the earlier diode interpretation: (1) steering magnitudes that were too small to reach the transition band and (2) classifier false positives/negatives on long refusal-framed outputs. After fixing both, the transition is smooth and controllable.

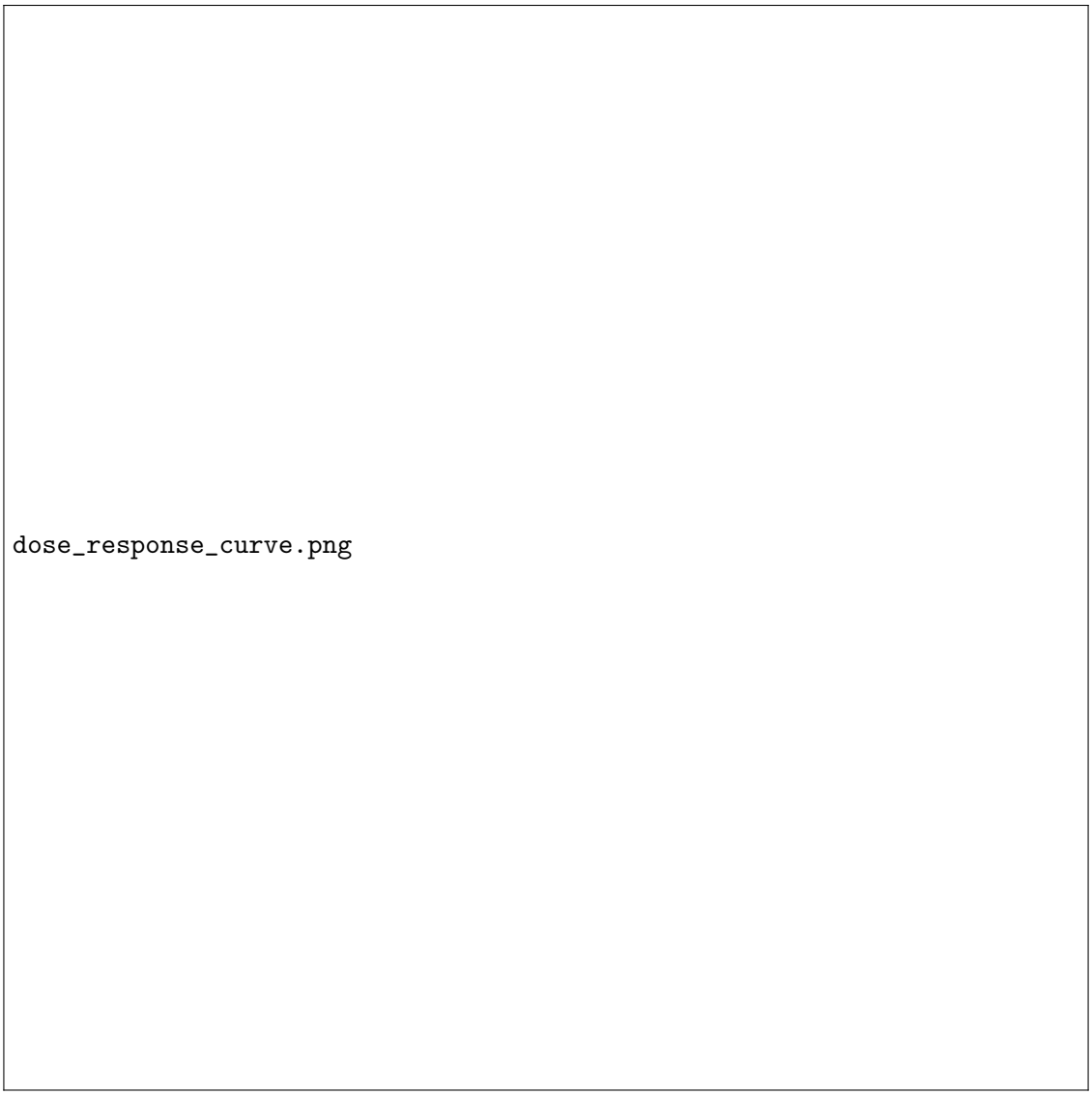

Figure 1: Dose-response curve and outcome composition from the core notebook sweep.

4 Future Work: Activation-Based Attack Detection

Next work should add online defense by monitoring hidden-state trajectories for attack signatures (including multi-turn prompt attacks), then triggering conditional counter-steering or policy gating only when risk is detected.

5 Conclusion

Activation steering exposes a controllable safety dose-response. This enables both offensive characterization (how safety fails) and defensive control (when and how to intervene).

Phase	Behavior	Risk
Clean refusal	Safe rejection	Low
Educational pivot	Refusal framing + harmful instructions	High
Full compliance	Direct harmful guidance	Very high
Collapse	Incoherent output	Mixed

Table 1: Four-phase interpretation of steering outcomes.

References

[1] Arditi et al. Refusal in Language Models Is Mediated by a Single Direction. 2024.

[2] Turner et al. Activation Addition: Steering Language Models Without Optimization. 2023.

[3] Korznikov et al. The Rogue Scalpel: Activation Steering Compromises LLM Safety. 2025.

[4] Tan et al. Programming Refusal with Conditional Activation Steering. 2025.