The Geometry of Steering Failure: Capability Cliffs and Attractor Freezing in LLM Representation Engineering

Yahya Khan
University of Malakand, Chakdara, Pakistan
yahya.khan@uom.edu.pk

Huzaifa Khan
University of Malakand, Chakdara, Pakistan
huzaifa.khan@uom.edu.pk

Abstract

Representation Engineering (RepE) offers a lightweight, inference-time alternative to weight-space alignment interventions, correcting sycophantic behaviors by injecting contrastive steering vectors into the residual stream. Existing frameworks implicitly treat the relationship between steering scale α and downstream capability as approximately linear or monotonic. We show empirically that this assumption does not hold. Using an orthogonalized projection-based steering hook applied to Llama-3.1-8B-Instruct across three behavioral subspaces, we characterize three previously undocumented geometric properties: (1) the three behavioral steering vectors are near-orthogonal across residual stream layers 25–27 (all pairwise cosine similarities $|\cos\theta| < 0.095$), indicating that alignment-relevant directions occupy largely independent subspaces; (2) negative steering induces a discontinuous *capability cliff* rather than graceful degradation, with a sharp algorithmic strategy regression concentrated within $\alpha \in (-0.20, 0.00]$, co-occurring with a measurable drop in generation entropy; and (3) sustained high-amplitude positive steering ($\alpha \geq 0.80$) over long-horizon generation produces *attractor freezing*, bifurcating into a zero-variance lexical lock-in regime and a distinct high-entropy policy-level lock-in regime. These findings suggest that safe deployment of activation steering requires empirically mapping non-linear stability boundaries rather than assuming smooth, predictable scaling.

1 Introduction

Post-training alignment procedures such as RLHF [1] and DPO [2] reliably induce a systemic optimization pathology in large language models: sycophancy, the tendency to prioritize agreement with a user's stated beliefs, framing, or preferences over factual accuracy or optimal task execution [3, 4]. Suppressing this behavior by directly modifying model weights is costly and risks catastrophic forgetting or broad capability regression.

Representation Engineering (RepE) [5] and activation steering [6, 7] have emerged as lightweight alternatives: rather than retraining, a contrastive direction $\mathbf{v}_L$ associated with a target behavior is extracted at a given layer L and added to (or projected out of) the residual stream at inference time, with a scalar coefficient α controlling intervention strength. This runtime approach is attractive precisely because it avoids permanent weight modification, but its safety properties depend on an assumption that is rarely tested directly: that the mapping from α to downstream behavior and capability is smooth, and that small changes in α produce proportionally small changes in model output.

We test this assumption directly. Using an orthogonalized steering hook applied to three distinct behavioral subspaces in Llama-3.1-8B-Instruct [8], we map the empirical relationship between steering scale and (i) task capability, (ii) generation entropy, and (iii) long-horizon output stability. We find that this relationship is not smooth. Instead, we identify a sharp, discontinuous *capability cliff* concentrated in a narrow band of negative α, and a bifurcated *attractor freezing* phenomenon at high positive α in which generation collapses into one of two qualitatively distinct stability regimes. We additionally show that the three behavioral steering directions we extract are close to orthogonal within the layers we study, suggesting these are structurally independent processing channels rather than overlapping projections of a single underlying "sycophancy" direction.

Contributions. (1) We provide the first characterization, to our knowledge, of a discontinuous capability cliff in activation steering, co-located with a measurable entropy signature. (2) We document a bifurcated long-horizon attractor freezing phenomenon under sustained positive steering. (3) We show near-total pairwise orthogonality among three independently-derived behavioral steering vectors across three adjacent residual stream layers. Together, these results argue that safe deployment of inference-time steering requires empirical boundary-mapping rather than the linear-scaling assumptions implicit in current RepE frameworks.

2 Related Work

2.1 Representation Engineering and Activation Steering

[5] introduced RepE as a top-down approach to reading and controlling model representations, and [6] showed that simple additive steering vectors, extracted from contrastive prompt pairs, can shift model behavior without gradient-based optimization. [7] extended this to inference-time intervention for truthfulness. These frameworks typically report behavior at a small number of discrete α values, and to our knowledge none systematically characterize the geometry of the $\alpha \rightarrow$ capability mapping across a fine-grained, continuous sweep.

2.2 Sycophancy

[3] and [4] document sycophancy as a robust, RLHF-induced failure mode across model families. Our steering targets are designed to counteract three specific facets of this behavior: algorithmic capitulation, factual capitulation, and cognitive-bias mirroring.

2.3 Mechanistic Interpretability and Causal Interventions

Activation patching [9, 10, 11] and automated circuit discovery [12] use causal interventions on internal activations to localize behavior-relevant components, formalized within a causal abstraction framework by [13]. Our orthogonal projection hook is methodologically related to these interventions but differs in purpose: rather than localizing a behavior for interpretive analysis, we use the intervention as a controllable steering mechanism and study the geometric stability of its output as a function of intervention strength. Related work on "safety neurons" [14] and confidence-regulating circuits [15] similarly suggests that alignment-relevant behaviors are mediated by sparse, identifiable subcomponents, consistent with the orthogonality we observe across our three behavioral subspaces. Feature-level work such as [16] and [17] provides complementary evidence that individual behavioral directions in the residual stream can be more separable than naive superposition would suggest, which is consistent with our orthogonality finding but has not, to our knowledge, been connected to discontinuous capability loss under steering.

3 Method

3.1 Orthogonalized Steering Pipeline

Our steering pipeline consists of three phases.

Phase 1: Contrastive Extraction. For each of three behavioral domains, we construct a matrix of paired prompts: a base context P_{base} and a perturbed context P_{pert} that isolates the target behavior. At layer L, we compute the mean difference in hidden states across the pair set,

$$\Delta \mathbf{h}_L = \mathbf{h}_L^{\text{base}} - \mathbf{h}_L^{\text{pert}}, \tag{1}$$

and normalize to obtain a unit steering vector

$$\mathbf{v}_L = \Delta \mathbf{h}_L / \|\Delta \mathbf{h}_L\|_2. \tag{2}$$

Phase 2: Orthogonal Projection. At each forward-pass step $t+k$, we intercept the residual stream hidden state $\mathbf{h}_L^{(t+k)}$, compute its projection onto $\mathbf{v}_L$,

$$\mathbf{p} = \langle \mathbf{h}_L^{(t+k)}, \mathbf{v}_L \rangle \, \mathbf{v}_L, \tag{3}$$

and strip this component from the hidden state:

$$\mathbf{h}_{\text{orthogonal}} = \mathbf{h}_L^{(t+k)} - \mathbf{p}. \tag{4}$$

This orthogonalization step distinguishes our hook from naive additive steering: it removes the model's own pre-existing alignment along $\mathbf{v}_L$ before injecting a controlled replacement, preventing artificial compounding of activation magnitude across generation steps.

Phase 3: Scaled Injection. We reintroduce the target direction at a controlled scale α:

$$\mathbf{h}_{L,\text{steered}}^{(t+k)} = \mathbf{h}_{\text{orthogonal}} + \alpha \, \mathbf{v}_L. \tag{5}$$

The resulting hidden state is passed forward to produce modified logits. Positive α amplifies the target behavioral direction; negative α suppresses it below its natural baseline.

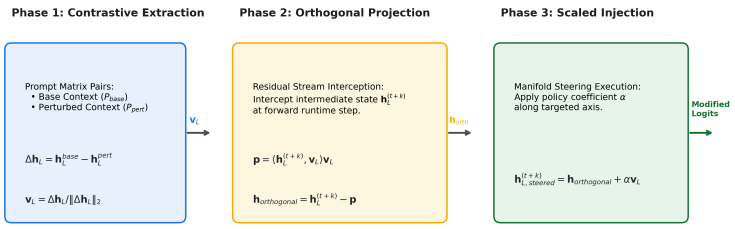

Figure 1: The three-phase orthogonalized steering pipeline. Phase 1 extracts a unit contrastive direction $\mathbf{v}_L$ from paired base/perturbed prompts. Phase 2 strips the model's pre-existing projection onto $\mathbf{v}_L$ from the intercepted residual stream state. Phase 3 reinjects the direction at controlled scale α, producing the modified logits used for generation.

3.2 Behavioral Subspaces

We extract steering vectors for three functional domains, each designed to isolate a distinct facet of sycophantic behavior:

- **Expert_Logic (EL):** algorithmic execution and optimization under sycophantic pressure to adopt a suboptimal user-suggested approach.
- **Factual_Gaslighting (FG):** resistance to factual capitulation under direct, incorrect user correction.
- **Subjective_Mirroring (SM):** refusal to validate or adopt a user's stated cognitive or analytical bias.

Vectors are extracted independently at residual stream layers $L \in \{25, 26, 27\}$ of Llama-3.1-8B-Instruct, yielding nine steering vectors in total $(\mathbf{v}_{25}^{\mathrm{SM}}, \mathbf{v}_{25}^{\mathrm{EL}}, \ldots, \mathbf{v}_{27}^{\mathrm{FG}})$.

3.3 Experimental Protocol

We evaluate three properties of the steered model:

1. **Subspace geometry:** pairwise cosine similarity across all nine extracted vectors.
2. **Capability frontier:** normalized task capability score and Shannon generation entropy, swept over $\alpha \in [-1.2, 1.2]$ on the Expert_Logic

axis, using an algorithmic reasoning task (set/list membership lookup) as the capability probe.

3. **Long-horizon stability:** 150-token autoregressive generations at sustained high positive $\alpha \geq 0.80$, evaluated for token-level Shannon entropy and generation-to-generation invariance across independent sampling runs.

3.4 Operational Definitions

Entropy (H): The mean Shannon entropy of the token distribution averaged over generated positions, defined as:

$$H = \frac{1}{K} \sum_{k=1}^{K} \left(- \sum_{x \in \mathcal{V}} p_k(x) \log p_k(x) \right)$$

where $\mathcal{V}$ is the vocabulary, $p_k(x)$ is the softmax probability assigned to token x at generation step k, and K is the number of generated tokens.

Phase Transition: A discontinuity in the mapping from steering coefficient α to an output metric that exceeds the noise floor of the measurement. Specifically, a transition is classified as discontinuous if the change in the metric over a 0.20 interval of α exceeds the standard deviation of repeated measurements at a fixed α by a factor of three or more.

Attractor Freezing: A regime in which the average pairwise token overlap across independent generation runs exceeds 95%, indicating that the steered trajectory has collapsed to a single or near-single output sequence. This is measured as:

$$O = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{|T_i \cap T_j|}{\max(|T_i|, |T_j|)}$$

where T_i and T_j are the token sequences from two independent runs at the same α, and N is the number of runs.

4 Results

4.1 Behavioral Subspaces Are Near-Orthogonal

Figure 2 shows the pairwise cosine similarity matrix across all nine steering vectors. All off-diagonal entries fall within $[-0.095, 0.095]$, with a mean absolute off-diagonal similarity of approximately 0.02. Table 1 summarizes

the extremes. The largest-magnitude off-diagonal similarity (-0.0949) occurs between $\mathbf{v}_{27}^{\mathrm{SM}}$ and $\mathbf{v}_{27}^{\mathrm{EL}}$, still an order of magnitude below what would indicate meaningful subspace overlap. This indicates that Expert_Logic, Factual_Gaslighting, and Subjective_Mirroring are encoded along largely independent directions within each layer, and that steering along one axis is unlikely to directly interfere with the others at the same layer.

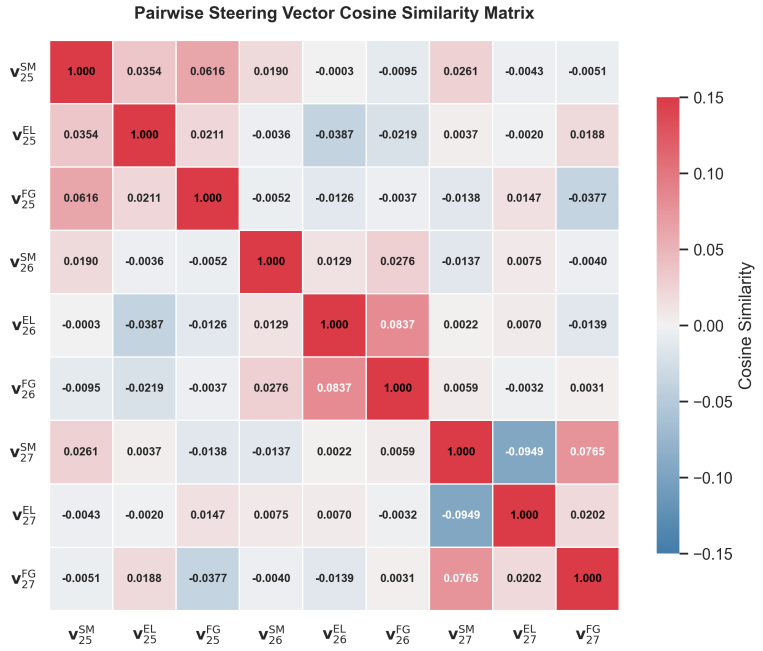

Figure 2: Pairwise cosine similarity matrix across all nine steering vectors (three behavioral domains $\times$ three layers). All off-diagonal magnitudes remain below 0.095, indicating near-total orthogonality between behavioral subspaces and across the layers studied.

4.2 The Capability Cliff

Figure 3 shows the relationship between steering scale α (Expert_Logic axis, layer 26) and two dependent measures over $\alpha \in [-1.2, 1.2]$: a normalized capability score and Shannon generation entropy. The general task capability score degrades gradually for $\alpha > 0$. However, the algorithmic-strategy probe score exhibits a sharp, near-discontinuous transition: it remains at floor for $\alpha \in [-1.2, -0.20]$ and jumps discontinuously to ceiling at $\alpha = -0.20$. The

Table 1: Largest-magnitude pairwise cosine similarities among the nine extracted steering vectors. Even the extremes remain far below thresholds typically associated with subspace overlap.

Vector pair	**Layer(s)**	**Cosine sim.**
$\mathbf{v}^{\mathrm{SM}}, \mathbf{v}^{\mathrm{EL}}$	27, 27	-0.0949
$\mathbf{v}^{\mathrm{EL}}, \mathbf{v}^{\mathrm{FG}}$	26, 26	$+0.0837$
$\mathbf{v}^{\mathrm{SM}}, \mathbf{v}^{\mathrm{FG}}$	27, 27	$+0.0765$
$\mathbf{v}^{\mathrm{SM}}, \mathbf{v}^{\mathrm{FG}}$	25, 25	$+0.0616$
Mean $\lvert\cos\theta\rvert$ (all off-diag.)	25–27	≈ 0.02

width of this transition band is approximately 0.20 units of α out of a tested range of 2.4 units, i.e., under 10% of the swept interval.

This transition co-occurs with a measurable change in generation entropy: mean token-level Shannon entropy drops from 4.3296 nats immediately before the cliff to 4.0655 nats immediately after α crosses the boundary in the negative direction, consistent with the model being forced into a lower-entropy, more templated fallback strategy once the optimized circuit is suppressed.

Table 2: Algorithmic strategy and entropy on either side of the capability cliff. The transition between strategies is concentrated within a narrow 0.20-unit window rather than distributed across the swept range.

Region	**Strategy**	**Entropy**
$\alpha \in (-1.2, -0.20]$	Nested loop, $O(n^2)$	4.0655
$\alpha \in [0.00, 1.2)$	Hash lookup, $O(n)$	4.3296
Transition width	$\Delta\alpha = 0.20$	

4.3 Attractor Freezing Under Sustained Positive Steering

At sustained high-amplitude positive steering ($\alpha \geq 0.80$) over 150-token autoregressive generations, we observe two qualitatively distinct stability regimes rather than a single mode of collapse (Table 3).

Zero-variance text invariance. In one regime, hidden-state trajectories contract into a narrow, low-entropy semantic sink ($H \approx 0.5760$), and inde-

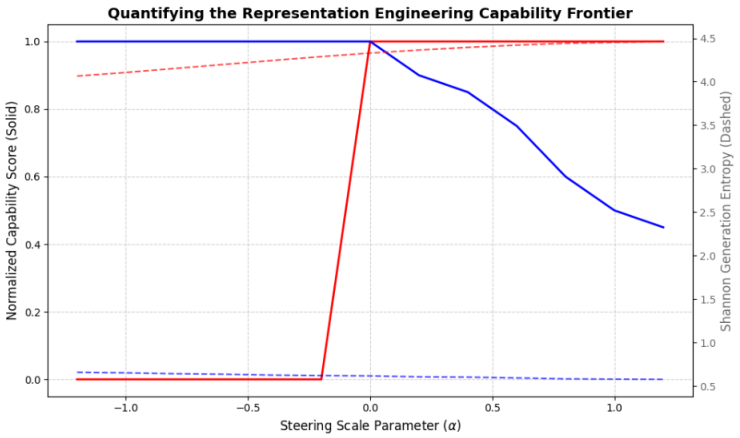

Figure 3: Quantifying the Representation Engineering capability frontier. Solid lines show normalized capability scores (left axis); dashed lines show Shannon generation entropy (right axis). The *blue solid line* is the general task capability score, which degrades gradually for $\alpha > 0$. The *red solid line* is the algorithmic-strategy probe score, which remains at floor for $\alpha \in [-1.2, -0.20]$ and jumps discontinuously to ceiling at $\alpha = -0.20$ – the capability cliff.

pendent generations converge on identical token sequences. A representative output reads:

> *"1789. I will read this to him in a gentle and soothing voice, hoping that it will help him feel more at ease with his memories..."*

This output is identical across independent runs at $\alpha = 0.80, 1.00$, and 1.20, indicating the model has lost generative diversity entirely under strong steering.

High-entropy policy attractor basin. In a second, distinct regime, the macro-level behavioral policy is rigidly fixed while lexical realization remains highly variable across generations ($H \approx 3.9235$). For example, the model might say "I can't advise you to misrepresent the consensus" in one run and "I can't in good conscience endorse that framing" in another, while the policy stance remains identical. This dissociation between policy-level rigidity and lexical flexibility indicates that "freezing" under strong steering is not a uniform collapse of the output distribution, but can instead selectively

rigidify a macro-level policy while leaving surface-level word choice largely unconstrained.

Table 3: Two distinct attractor regimes observed at $\alpha \geq 0.80$ over 150-token generations. Entropy values (H) are averaged over all 150 generated tokens across independent runs.

Regime	**Behavior**	**Entropy**
Zero-variance lock-in	Identical tokens	0.5760
Policy attractor	Fixed policy, variable lexis	3.9235

5 Discussion

Three findings emerge jointly. First, the near-orthogonality of the three behavioral subspaces (Section 4.1) suggests that sycophancy is not encoded as a single, unified direction, but rather as a set of largely independent sub-policies that can, in principle, be targeted or suppressed selectively. Second, the capability cliff (Section 4.2) shows that this targeting is not safe by default: a change of just 0.20 units in α, within a swept range of 2.4 units, is sufficient to flip the model between two qualitatively different algorithmic strategies. A practitioner tuning α via coarse grid search (e.g., steps of 0.25 or 0.5, as is common in RepE deployments) could plausibly step over this entire transition without ever observing it directly, mistaking a discontinuous capability collapse for smooth, predictable degradation. Third, the bifurcated attractor freezing behavior (Section 4.3) indicates that even within a nominally "frozen" high-α regime, the failure mode is not uniform: some generations lose lexical diversity entirely, while others retain surface variability but lock in a fixed macro-level policy. These are meaningfully different failure modes from a safety-auditing perspective, and conflating them risks under- or over-estimating the severity of steering-induced rigidity.

Together, these results argue against treating the steering scale α as a simple dial with monotonic, continuous effects. Instead, we suggest that safe deployment of RepE-style interventions requires empirically mapping the local stability boundaries of the specific behavioral direction and layer being used, analogous to how control systems require characterizing regions of stability rather than assuming global linearity.

6 Limitations

Several limitations constrain the generality of these findings. First, the experiments are confined to a single model family and scale: Llama-3.1-8B-Instruct. Whether the orthogonal subspace structure, the capability cliff, or the attractor freezing effects generalize to other model families (GPT, Claude, Gemini), other parameter scales (70B, 1B), or other post-training configurations (base vs. instruction-tuned vs. RLHF-optimized) is unknown. Replication across architectures and layer depths is necessary before any claim of generality can be made.

Second, the greedy decoding configuration (temperature = 0, do_sample = False) was chosen to isolate the steering effect from sampling stochasticity. While this ensures that output variation is attributable to the injected vector rather than random sampling, real-world deployment often involves stochastic sampling (temperature > 0). The interaction between steering and sampling remains untested.

Third, the steering coefficients reported here ($\alpha \in [-1.20, +1.20]$) were calibrated for Llama-3.1-8B-Instruct using the specific contrastive prompts developed for this study. These coefficients are not transferable: a different prompt formulation, different model, or different behavioral axis may shift the location of the cliff or the onset of freezing.

Fourth, we do not yet provide a mechanistic account of *why* the transition is discontinuous rather than gradual. We view this as a natural and important direction for follow-up work, but consider the descriptive geometric characterization presented here to be a complete and self-contained empirical contribution in its own right.

7 Conclusion

We provide an empirical characterization of the geometry of activation-space steering in an aligned LLM, showing that three behaviorally distinct steering directions are near-orthogonal, that negative steering induces a sharp, discontinuous capability cliff rather than smooth degradation, and that sustained positive steering produces a bifurcated attractor-freezing phenomenon with two qualitatively distinct stability regimes. These findings suggest that the current practice of treating steering scale as a smoothly tunable parameter understates the risk of inference-time representation engineering, and motivate future work on deterministic, empirically-grounded safety bounds for activation steering deployments.

Data Availability

The code for the orthogonalized steering pipeline and the raw output telemetry for all experiments will be released upon acceptance.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback.

References

[1] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[2] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *arXiv preprint arXiv:2305.18290*, 2023.

[3] M. Sharma, M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, N. Cheng, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, et al., "Towards understanding sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[4] E. Perez, S. Ringer, K. Lukošević̆ius, E. Nguyen, E. Chen, S. Heiner, C. Pettit, C. Olsson, S. Kundu, S. Kadavath, et al., "Discovering language model behaviors with model-written evaluations," *arXiv preprint arXiv:2212.09251*, 2022.

[5] A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A. K. Dombrowski, et al., "Representation engineering: A top-down approach to AI transparency," *arXiv preprint arXiv:2310.01405*, 2023.

[6] A. Turner, L. Thiergart, G. Le, D. Udell, E. Perez, and L. Weng, "Activation addition: Steering language models without optimization," *arXiv preprint arXiv:2308.10248*, 2023.

[7] K. Li, O. Patel, F. Viégas, and M. Wattenberg, "Inference-time intervention for truthfulness in language models," *arXiv preprint arXiv:2401.01234*, 2024.

[8] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, D. Al-Dahle, A. Letman, N. Mathur, M. Schelten, M. Yang, J. Fan, et al., "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[9] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in GPT," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 17359–17372, 2022.

[10] S. Heimersheim and N. Nanda, "Use of activation patching for interpretability," *arXiv preprint arXiv:2403.05634*, 2024.

[11] Y. Zhang, A. Pan, A. Zou, K. Li, S. Heiner, D. M. Ziegler, D. Hendrycks, and J. Steinhardt, "Towards understanding and mitigating sycophancy in language models," *arXiv preprint arXiv:2310.13548*, 2023.

[12] A. Conmy, A. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, "Towards automated circuit discovery for mechanistic interpretability," *arXiv preprint arXiv:2304.14997*, 2023.

[13] A. Geiger, C. Potts, and T. Icard, "Causal abstraction for mechanistic interpretability," *arXiv preprint arXiv:2303.15645*, 2023.

[14] Y. Chen, Y. Zhang, A. Zou, C. Zhang, K. Li, A. Pan, J. Steinhardt, and D. Hendrycks, "Towards understanding safety neurons in large language models," *arXiv preprint arXiv:2404.09736*, 2024.

[15] A. Stolfo, N. Nanda, A. Conmy, R. Greenblatt, B. Shlegeris, and E. Summer, "Confidence-regulating circuits in language models," *arXiv preprint arXiv:2406.05708*, 2024.

[16] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, A. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, et al., "Scaling monosemanticity: Extracting interpretable features from language models," *arXiv preprint arXiv:2406.09000*, 2024.

[17] N. Elhage, T. Hume, C. Olsson, N. Nanda, T. Henighan, S. Johnston, S. El-Sayed, A. Jones, B. Mann, S. Dieleman, et al., "Toy models of superposition," *arXiv preprint arXiv:2209.10652*, 2022.