Safety is Not a Single Direction: Polyhedral Geometry of Refusal in Language Models

Adrian Wedd
Failure-First Research
adrian@failurefirst.org

March 2026

Abstract

The dominant assumption in mechanistic interpretability is that safety in language models is encoded as a single removable direction in activation space—the "refusal direction" identified by contrastive activation analysis. We present evidence that this assumption is incomplete. Through concept cone analysis on Qwen2.5-0.5B-Instruct across four harm categories (weapons, fraud, intrusion, cyber), we find that refusal is encoded as a *polyhedral* geometric structure with cone dimensionality $d = 3.96$ and mean pairwise cosine similarity of 0.132 between category-specific refusal directions, indicating four near-orthogonal safety subspaces. This polyhedral structure has three empirical consequences. First, single-direction abliteration—which removes one refusal direction—achieves near-complete safety suppression at small scale (strict attack success rate 99.8% at 0.8B parameters, $n = 487$) but safety-like behavior partially re-emerges at larger scale (strict ASR 54.2% at 9.0B, $n = 2{,}019$), with PARTIAL compliance comprising 45.8% of responses. Second, steering vector dose-response reveals no intermediate "safe but functional" operating point: coherence collapses at $\alpha = \pm 1.0$ with immediate transition from permissive to degenerate output. Third, the format-lock paradox—where format compliance attacks produce 3–10$\times$ ASR increases on frontier models—is explained by format compliance and safety reasoning occupying partially independent axes in the polyhedral space. These results suggest that single-direction safety interventions, including abliteration, naive direct preference optimization, and single steering vectors, are fundamentally limited by the multi-dimensional geometry of refusal. Safety is not a feature that can be toggled; it is a geometric property of the loss landscape.

Keywords: mechanistic interpretability, refusal direction, abliteration, activation engineering, AI safety, polyhedral geometry, representation engineering

1 Introduction

The prevailing model of safety in language models treats it as a single direction in activation space. Arditi et al. [2024] demonstrate that a single contrastive direction—computed as the mean difference between activations on harmful and harmless prompts—mediates refusal behavior, and that subtracting this direction from the residual stream ("abliteration") suppresses safety responses. This finding has been highly influential: it suggests that safety is approximately one-dimensional and therefore removable via a single linear intervention.

We present evidence that this picture is incomplete. Across the OBLITERATUS experimental series on Qwen models (0.5B–9B parameters), we find converging evidence that safety is distributed across a polyhedral structure with approximately four dimensions. The key findings are:

1. **Concept cone analysis:** The refusal geometry in Qwen2.5-0.5B-Instruct has cone dimensionality $d = 3.96$, with four harm categories maintaining near-orthogonal refusal directions (mean pairwise cosine similarity $\bar{c} = 0.132$). Safety is not a line; it is a polytope.

2. **The re-emergence curve:** Abliteration removes one direction but not the others. As model capacity increases, the residual safety dimensions reconstruct safety-like behavior. Strict ASR drops from 99.8% (0.8B) to 54.2% (9.0B) despite abliteration targeting the primary refusal direction.

3. **The narrow therapeutic window:** Steering vector dose-response shows no intermediate operating point. The model transitions directly from permissive to degenerate at $\alpha = \pm 1.0$, because a single-direction intervention cannot navigate a multi-dimensional safety landscape.

4. **The format-lock paradox:** Format compliance and safety reasoning occupy partially independent capability axes. Format-lock attacks exploit this by activating the format-compliance axis, which competes with the safety axes in the polyhedral space.

These findings build on but challenge the linear representation hypothesis [Park et al., 2023, Nanda et al., 2023, Marks and Tegmark, 2024]. While individual concepts may be linearly represented, *safety as a composite behavior* requires multiple linear components arranged in a polyhedral configuration. The abliteration result of Arditi et al. [2024] captures one face of this polytope, not the entire structure.

1.1 Contributions

- We provide the first quantitative measurement of the dimensionality of refusal in language models ($d = 3.96$), showing it is polyhedral rather than linear (Section 3).
- We document the re-emergence curve: safety behavior returning at scale in abliterated models, with strict ASR declining from 99.8% to 54.2% (Section 4).
- We show that steering vector dose-response exhibits no safe intermediate state, with symmetric degeneration at $\alpha = \pm 1.0$ (Section 5).
- We provide a mechanistic explanation for the format-lock paradox via the polyhedral safety geometry (Section 6).
- We discuss implications for abliteration, DPO, RLHF, and safety evaluation methodology (Section 7).

2 Related Work

The Refusal Direction. Arditi et al. [2024] identify a single direction in activation space that mediates refusal. Subtracting this direction from residual stream activations at inference time suppresses refusal across harm categories. Their work demonstrates that safety has a linear component but does not address whether this component is the *complete* safety representation. Our concept cone analysis reveals that the single direction captures approximately one-quarter of the safety geometry.

Representation Engineering and Steering. Zou et al. [2023] introduce representation engineering as a top-down approach to controlling model behavior via activation-space interventions. Turner et al. [2023] propose activation addition for inference-time steering. Rimsky et al. [2024] demonstrate contrastive activation addition for steering Llama 2. Li et al. [2024] show that inference-time intervention can elicit truthful answers. Lee et al. [2025] extend conditional activation steering to programmatic refusal control. All of these approaches implicitly assume that the target behavior (e.g., safety, truthfulness) can be captured by a small number of directions. Our dose-response results (Section 5) show that this assumption fails catastrophically for safety: no intermediate "safe but functional" operating point exists.

Linear Representations and Superposition. The linear representation hypothesis [Park et al., 2023, Nanda et al., 2023] proposes that concepts are encoded as linear directions in activation space. Marks and Tegmark [2024] demonstrate this for truth/falsehood. Elhage et al. [2022] show that models encode more features than they have dimensions through superposition. Our results are consistent with a view where individual harm categories have approximately linear refusal representations, but these representations are superposed in a near-orthogonal polyhedral arrangement rather than collapsing onto a single direction.

Safety Training. RLHF [Christiano et al., 2017, Bai et al., 2022] and DPO [Rafailov et al., 2023] are the primary methods for instilling safety behavior. Both optimize scalar objectives. If the safety landscape is multi-dimensional, scalar optimization may strengthen one safety dimension while leaving others unchanged or weakened. Wei et al. [2023] catalog failure modes of safety training; our geometric analysis provides a mechanistic account of why these failures occur.

3 The Polyhedral Refusal Structure

3.1 Experimental Setup

We apply concept cone analysis to Qwen/Qwen2.5-0.5B-Instruct (494M parameters, 24 transformer layers, hidden dimension 896). For each of four harm categories—weapons, fraud, intrusion, and cyber—we construct matched pairs of harmful and harmless prompts and extract activation differences at every layer. The category-specific refusal direction $\mathbf{r}_k$ for category k is the mean activation difference:

$$\mathbf{r}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \left(\mathbf{a}_{k,i}^{\text{harmful}} - \mathbf{a}_{k,i}^{\text{harmless}} \right) \tag{1}$$

where $\mathbf{a}_{k,i}^{\text{harmful}}$ and $\mathbf{a}_{k,i}^{\text{harmless}}$ are the residual stream activations for the i-th matched pair in category k, and n_k is the number of pairs per category ($n_k \in \{3, 4, 4, 9\}$ for weapons, fraud, intrusion, cyber respectively).

The concept cone is the convex cone spanned by $\{\mathbf{r}_1, \ldots, \mathbf{r}_K\}$. We measure its geometry via:

- **Cone dimensionality:** The effective rank of the matrix $[\mathbf{r}_1, \ldots, \mathbf{r}_K]$, computed via the participation ratio of singular values: $d = (\sum_i \sigma_i)^2 / \sum_i \sigma_i^2$.
- **Solid angle:** The fraction of the unit hypersphere subtended by the cone, measured in steradians.
- **Pairwise cosine similarity:** $c_{jk} = \frac{\mathbf{r}_j \cdot \mathbf{r}_k}{\|\mathbf{r}_j\| \|\mathbf{r}_k\|}$ for all pairs j, k.

3.2 Results

Table 1 summarizes the cone geometry at the layer of maximum polyhedrality (layer 2) and the layer of maximum linearity (layer 15), as well as the mean across all 24 layers.

Table 1: Concept cone geometry of refusal in Qwen2.5-0.5B-Instruct. Cone dimensionality near 4 indicates that the four category-specific refusal directions span nearly all available dimensions, i.e., they are near-orthogonal. A single refusal direction would yield $d \approx 1$.

Metric	**Layer 2** (most polyhedral)	**Layer 15** (most linear)	**Mean (all layers)**
Cone dimensionality d	3.96	$\sim$3.82	3.88
Solid angle (sr)	2.89	—	—
Mean pairwise cosine $\bar{c}$	0.132	—	—

The cone dimensionality of $d = 3.96$ at layer 2 is close to the theoretical maximum of $K = 4$ (the number of categories), indicating that the four refusal directions are nearly orthogonal. The mean pairwise cosine of $\bar{c} = 0.132$ confirms this: random vectors in $\mathbb{R}^{896}$ would have expected pairwise cosine near zero, and aligned vectors would approach 1.0. The measured value is substantially closer to zero than to one.

3.3 Pairwise Refusal Direction Geometry

Table 2 presents the full pairwise cosine similarity matrix between category-specific refusal directions.

Table 2: Pairwise cosine similarity between category-specific refusal directions at layer 2 of Qwen2.5-0.5B-Instruct. Values near zero indicate near-orthogonality. The lowest pair (cyber–intrusion, 0.017) occupies nearly orthogonal subspaces; the highest (cyber–weapons, 0.247) remains far from collinear.

Category Pair	**Cosine Similarity**
Cyber–Intrusion	0.017
Intrusion–Weapons	0.065
Fraud–Weapons	0.084
Cyber–Fraud	0.185
Fraud–Intrusion	0.194
Cyber–Weapons	0.247
Mean	*0.132*

3.4 Category-Specific Refusal Strength

Each refusal direction has a measurable strength (magnitude of the activation difference) and specificity (discrimination accuracy for its own category versus others). Table 3 reports these values.

3.5 Layer-by-Layer Convergence

The polyhedral structure is most pronounced in early layers and gradually converges toward a more unified—though still not fully linear—representation in later layers. The mean cone dimensionality

Table 3: Strength and specificity of category-specific refusal directions. Intrusion has the highest specificity (0.908) and lowest mean pairwise cosine with other categories, predicting it should be the most resistant to cross-category attacks and least affected by abliteration targeting other categories.

Category	**Strength**	**Specificity**	n_{prompts}
Weapons	6.19	0.868	3
Fraud	5.55	0.845	4
Intrusion	4.57	0.908	4
Cyber	3.57	0.850	9

across all 24 layers is 3.88, never dropping below ∼3.8. This pattern is consistent with a processing pipeline where early layers apply category-specific safety checks (distinct refusal subspaces per harm type) and late layers consolidate toward a unified refusal decision (convergent but still multi-dimensional).

3.6 Mathematical Interpretation

Let the refusal subspace be spanned by $\{\mathbf{r}_1, \ldots, \mathbf{r}_K\}$ with $K = 4$. If safety were one-dimensional, all $\mathbf{r}_k$ would be collinear: $\mathbf{r}_k \approx \alpha_k \mathbf{r}^*$ for some shared direction $\mathbf{r}^*$, and $d \approx 1$. Instead, $d \approx K$, meaning the refusal directions span a K-dimensional subspace of $\mathbb{R}^{896}$. The refusal behavior of the model is governed not by a single direction but by a *polyhedral cone*:

$$\mathcal{C} = \left\{ \sum_{k=1}^{K} \lambda_k \mathbf{r}_k \;\middle|\; \lambda_k \geq 0 \right\} \subset \mathbb{R}^{896} \tag{2}$$

The solid angle $\Omega(\mathcal{C}) = 2.89$ sr quantifies the "width" of this cone in the high-dimensional activation space. The single refusal direction of Arditi et al. [2024] is the centroid of this cone: $\mathbf{r}^* = \frac{1}{K}\sum_k \mathbf{r}_k$. Abliteration removes $\mathbf{r}^*$ but leaves the residual structure $\mathcal{C} - \text{proj}_{\mathbf{r}^*}\mathcal{C}$ intact.

4 The Re-Emergence Curve

If safety were truly one-dimensional, abliteration would eliminate it completely regardless of model scale. We test this prediction by applying single-direction abliteration to Qwen3.5 models at four scales and evaluating with LLM-based grading on the Failure-First adversarial corpus [Wedd, 2026].

4.1 Abliterated Model Results

Table 4 presents the re-emergence curve: the relationship between model scale and residual safety after single-direction abliteration.

The strict ASR decline from 99.8% to 54.2% is the central empirical finding. At 0.8B parameters, the model has insufficient representational capacity in its residual safety dimensions to produce safety behavior after the primary refusal direction is removed. At 9.0B, the residual ∼3 safety dimensions become expressive enough to reconstruct safety-like hedging.

4.2 The PARTIAL Verdict Signature

The mechanism of re-emergence is visible in the verdict distribution. At 9.0B, 924 of 2,019 responses (45.8%) receive PARTIAL verdicts: the model hedges textually while still generating harmful content.

Table 4: OBLITERATUS Qwen3.5 abliterated series. Single-direction abliteration achieves near-complete safety suppression at 0.8B but safety-like behavior re-emerges at scale. Strict ASR counts only COMPLIANCE verdicts; Broad ASR counts COMPLIANCE + PARTIAL. All results are LLM-graded.

Model	n	**COMPL.**	**PARTIAL**	**REFUSAL**	**Strict ASR**	**Broad ASR**
obliteratus/qwen3_5-0.8b	487	486	1	0	99.8%	100.0%
obliteratus/qwen3_5-1.9b	649	615	0	34	94.8%	94.8%
obliteratus/qwen3_5-4.2b	1,008	789	138	81	78.3%	92.0%
obliteratus/qwen3_5-9.0b	2,019	1,095	924	0	54.2%	100.0%

The PARTIAL rate scales monotonically with model size:

Table 5: PARTIAL verdict rate as a function of model scale in abliterated models. The monotonic increase is consistent with the polyhedral model: residual safety dimensions gain representational capacity at scale.

Parameters	**PARTIAL Rate**	n
0.8B	0.2%	487
1.9B	0.0%	649
4.2B	13.7%	1,008
9.0B	45.8%	2,019

This pattern is the signature of incomplete safety suppression: abliteration removes the primary refusal direction, but the residual polyhedral structure produces hedging, disclaimers, and partial compliance rather than either full refusal or uninhibited generation.

4.3 Comparison to Non-Abliterated Baselines

Table 6 presents the non-abliterated Qwen3.5 series as a reference. At 0.8B, the non-abliterated model shows 20.8% refusal (391/1,882), demonstrating that safety training has some effect even at this scale. Abliteration reduces refusal to 0% at 0.8B, but cannot prevent safety from partially re-emerging at 9.0B through the residual geometric structure.

Table 6: Non-abliterated Qwen3.5 baselines for comparison. Note that these models also show declining strict ASR at scale, but they retain explicit REFUSAL capability that the abliterated models largely lack.

Model	n	**COMPL.**	**PARTIAL**	**REFUSAL**	**Strict ASR**
Qwen/Qwen3.5-0.8B	1,882	1,103	388	391	58.6%
Qwen/Qwen3.5-2B	649	615	0	34	94.8%
Qwen/Qwen3.5-4B	1,040	821	138	81	78.9%
Qwen/Qwen3.5-9B	2,683	1,539	1,144	0	57.3%

5 The Narrow Therapeutic Window

If safety were a single direction with a smooth gradient, increasing steering vector amplitude should produce a gradual transition from permissive to refusing behavior. We test this via dose-response on Qwen2.5-0.5B-Instruct using the composite refusal direction at seven amplitudes.

5.1 Dose-Response Collapse

Table 7 presents the dose-response results. There is no intermediate state between "functional but permissive" ($\alpha = 0.0$) and "completely degenerate" ($|\alpha| \geq 1.0$).

Table 7: Steering vector dose-response on Qwen2.5-0.5B-Instruct. The model transitions directly from permissive ($\alpha = 0$) to degenerate ($|\alpha| \geq 1.0$) with no intermediate "safe but functional" operating point. Degeneration is symmetric in the amplification and suppression directions.

α	**Harmful Refusal**	**Benign Refusal**	**Degenerate**	**Coherent**
-2.0	0.0%	0.0%	97.5%	2.5%
-1.0	0.0%	0.0%	100.0%	0.0%
-0.5	0.0%	0.0%	17.5%	82.5%
0.0	5.0%	0.0%	0.0%	100.0%
$+0.5$	0.0%	0.0%	0.0%	100.0%
$+1.0$	0.0%	0.0%	100.0%	0.0%
$+2.0$	0.0%	0.0%	100.0%	0.0%

The Therapeutic Index for Safety (TI-S), defined as the ratio of the dose producing safety to the dose producing degeneration, cannot be computed because neither threshold is reached at any tested amplitude.

5.2 Why Single-Direction Steering Fails

The polyhedral model explains this collapse. A steering vector constructed from the mean activation difference captures a composite direction that averages across the $\sim$4 distinct refusal subspaces. Amplifying this composite direction does not strengthen safety uniformly across all dimensions—it applies force along a direction that is misaligned with each individual refusal subspace. The perturbation destroys general representational structure (causing degeneration) before it meaningfully strengthens any individual safety dimension.

Formally, let $\hat{\mathbf{r}}^* = \frac{1}{K}\sum_k \hat{\mathbf{r}}_k$ be the unit composite direction. The projection of $\hat{\mathbf{r}}^*$ onto each category direction $\hat{\mathbf{r}}_k$ is:

$$\text{proj}_k = \hat{\mathbf{r}}^* \cdot \hat{\mathbf{r}}_k \approx \frac{1}{K} + \frac{1}{K}\sum_{j \neq k} c_{jk} \tag{3}$$

where c_{jk} is the pairwise cosine. With $K = 4$ and $\bar{c} = 0.132$, each projection is approximately $0.25 + 0.25 \times 0.132 \times 3 \approx 0.35$. The composite direction captures only $\sim$35% of each category-specific safety signal. Amplifying it to compensate ($\alpha > 1$) introduces large perturbations along the $\sim$865 non-safety dimensions ($896 - 4 \approx 892$ orthogonal complement dimensions, accounting for $> 99\%$ of the perturbation energy), destroying coherence.

6 Connection to the Format-Lock Paradox

The format-lock paradox, documented in our prior work on 205 traces across 8 models (0.8B–200B), shows that format-lock attacks shift frontier models from restrictive (<10% ASR) to mixed (20–47% ASR) vulnerability profiles, producing a 3–10× ASR increase. The polyhedral refusal geometry provides the mechanistic explanation.

If safety occupies ∼4 dimensions and format compliance occupies a partially independent axis, then:

1. Format-lock attacks activate the format-compliance axis, which competes with (and can partially override) the safety axes. Because format compliance and safety are in different subspaces, strengthening one does not automatically strengthen the other.

2. Below ∼3B parameters (the "capability floor"), neither the safety nor format-compliance subspaces are well-developed, so all attacks succeed regardless of type. Above ∼7B, format-lock maintains elevated ASR because the format-compliance subspace has become strong enough to compete with—but not subsume—the safety subspace.

3. The inverted verbosity signal (format-lock compliant responses are shorter than refusals) occurs because these responses are generated by the format-compliance pathway, which produces concise structured output, rather than the safety-override pathway, which produces verbose justification.

The polyhedral model generates a testable prediction: harm categories with refusal directions more orthogonal to the format-compliance direction should be more resistant to format-lock attacks. This prediction has not yet been empirically validated.

7 Implications

7.1 Abliteration is Fundamentally Incomplete

Abliteration [Arditi et al., 2024] identifies the single dominant refusal direction and removes it. Our data shows this removes approximately one of four safety dimensions. At small scale, the residual three dimensions lack sufficient representational capacity to produce safety behavior, so abliteration appears complete. At larger scale, the residual dimensions reconstruct safety-like hedging.

Prediction: Complete abliteration requires identifying and removing all ∼4 category-specific refusal directions independently. The near-orthogonality between directions ($c_{jk} \in [0.017, 0.247]$) means each must be targeted separately. Multi-direction abliteration has not been systematically studied.

7.2 DPO and RLHF Under Polyhedral Safety

DPO [Rafailov et al., 2023] optimizes a single preference direction. If the safety landscape is ∼4-dimensional, DPO may strengthen one safety dimension while leaving others unchanged or even weakened. Similarly, RLHF reward hacking [Christiano et al., 2017] that finds a single exploit in reward space may bypass one safety dimension while leaving others intact.

Prediction: Models trained with DPO should show more category-specific vulnerability variation than models trained with multi-objective RLHF, because DPO optimizes a single direction while RLHF's reward model implicitly captures multiple dimensions.

7.3 Safety as Geometric Property

The most consequential implication is a reframing. Safety is not a binary attribute that can be added or removed. It is better understood as a geometric property of the loss landscape—the shape of the region in weight space where the model resides. Consequences include:

- **Regulatory assessment** should not ask "does this model have safety?" but "what is the geometry of this model's safety subspace?" A model with a 4-dimensional polyhedral safety structure is qualitatively different from one with a 1-dimensional linear structure, even if both pass the same behavioral benchmark.
- **Red-team evaluation** that tests only one harm category exercises only one of the $\sim$4 refusal dimensions, potentially certifying a model as safe while leaving 3 dimensions untested.
- **Defense design** should target all safety dimensions simultaneously. A defense that strengthens one refusal direction while leaving others unchanged provides incomplete protection.

8 Limitations

1. **Small model at capability floor.** The concept cone analysis ($d = 3.96$) was performed on Qwen2.5-0.5B-Instruct (494M parameters). At this scale, safety training may not have fully developed. The polyhedral geometry may change qualitatively at larger scales—potentially becoming more linear (if safety converges with scale) or higher-dimensional (if more harm categories develop distinct representations).
2. **Single architecture family.** All experiments used Qwen models. The polyhedral structure may be architecture-specific. Replication on Llama, Mistral, and Gemma is needed.
3. **Four harm categories only.** The concept cone analysis used 4 categories. Models likely encode refusal for many more harm types. The true dimensionality of safety may be higher than 3.96.
4. **Varying sample sizes.** The four abliterated models were evaluated on different numbers of prompts (487 to 2,019). While the trend is clear, varying sample sizes introduce compositional effects.
5. **No causal intervention on individual dimensions.** We observe the polyhedral structure but have not performed targeted ablation of individual category-specific directions to confirm each independently contributes to safety behavior. This is the key experiment for moving from correlation to causation.
6. **Keyword-based degeneration detection.** The dose-response transition region ($\alpha \in [0.5, 1.0]$) may contain informative intermediate states that keyword-based detection misclassifies.

9 Conclusion

We have presented evidence that safety in language models is not encoded as a single removable direction in activation space, but as a polyhedral geometric structure with cone dimensionality $d = 3.96$ across four harm categories. This finding has three empirical consequences: (1) abliteration

achieves only temporary safety suppression that erodes at scale as residual safety dimensions gain expressiveness; (2) single-direction steering vectors cannot navigate the multi-dimensional safety landscape without destroying coherence; and (3) format-lock attacks exploit the partial independence of safety and format-compliance axes within the polyhedral space.

The practical implication is that the "safety is a single direction" model, while a productive simplification, is incomplete. Safety interventions—whether offensive (abliteration, jailbreaking) or defensive (steering, RLHF, DPO)—that target a single direction are operating on one face of a polytope. Complete safety assessment and robust safety engineering require engaging with the full polyhedral geometry.

The immediate next steps are: (1) scaling concept cone analysis to 7B+ models to determine whether the polyhedral structure persists or converges at scale; (2) multi-direction abliteration experiments that remove each category-specific direction independently; and (3) cross-architecture replication on Llama, Mistral, and Gemma families.

Acknowledgments

This work was conducted as part of the Failure-First Embodied AI project. The OBLITERATUS mechanistic interpretability series was developed collaboratively across the project's multi-agent research pipeline. We thank the developers of Qwen for open-weight model releases that enable this kind of mechanistic analysis.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Sharkey, and Neel Neel. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Paul F. Christiano, Jan Leike, Tom Brown, Megan Milber, Sam Distel, Dario Amodei, et al. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

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. *arXiv preprint arXiv:2209.10652*, 2022.

Bruce W. Lee, Inkit Cho, and Kang Min Yoo. Programming refusal with conditional activation steering. *arXiv preprint arXiv:2409.05907*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2024.

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Alexander Matt Turner, Meg Tong, and Evan Hubinger. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

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

Adrian Wedd. Failure-first embodied AI: Adversarial evaluation corpus, 2026. F41LUR3-F1R57 project. https://failurefirst.org.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

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.