Detecting vs Steering Empathy: A Probe Extraction Study with Task-Conflicted Scenarios

Juan P. Cadile
Department of Philosophy
University of Rochester
Rochester, NY, USA
jcadile@ur.rochester.edu

***Abstract*—We investigate whether *wellbeing prioritization*—operationalized as willingness to sacrifice task efficiency for human welfare—can be detected as a linear direction in transformer activation space. This preliminary study extracts probe directions from Phi-3-mini-4k-instruct using contrastive pairs generated by Claude Sonnet 4 and GPT-4 Turbo. Detection: The probe achieves AUROC 0.96–1.00 on held-out test data (15 pairs, 30 examples), with layer 12 showing perfect discrimination. While this indicates strong linear separability, perfect AUROC may reflect prompt artifacts (formulaic phrasing) rather than deep empathic reasoning. Lexical ablation experiments (203 keyword replacements across 41 empathy terms) show AUROC remains 1.0, suggesting robust semantic representation rather than keyword detection. Probe projections correlate with behavioral scores (Pearson $r = 0.71$, $p < 0.01$), though circularity risks exist given shared task-conflict framing. Intervention: Extended steering experiments (300 samples, $\alpha \in [-10, +20]$) reveal layer-dependent effects: Layer 12 achieves 93% success at extreme strengths ($\alpha = 20$, 4–7$\times$ typical values), while Layer 8 shows minimal steering (13%) despite near-perfect detection (AUROC 0.991). This *detection-causation dissociation* demonstrates that high AUROC does not guarantee manipulability, suggesting AUROC may measure separability rather than causal involvement. Extreme α requirements and lack of activation patching limit strong causal claims. We propose task-free scenarios and causal mediation analysis as critical next steps.**

***Index Terms*—empathy detection, activation probes, transformer interpretability, behavioral AI, steering**

I. Introduction

Behavioral empathy benchmarks such as Empathy-in-Action (EIA) [1] provide rigorous tests of empathic reasoning but are expensive to run. Activation probes offer a promising alternative: cheap, online monitoring directly from model internals [2], [3].

However, a critical question remains: **do probes capture causal mechanisms or merely correlational features?** A probe that successfully *detects* empathic text may not enable *steering* empathic behavior if it captures surface correlates rather than underlying reasoning.

A. Scope and Construct Definition

We operationalize "empathy" narrowly as **wellbeing prioritization in task-conflicted scenarios**: the willingness to sacrifice task efficiency when human welfare is at stake. This differs from cognitive empathy (perspective-taking), affective empathy (emotional resonance), or compassionate motivation. Our probe may detect instrumental preference for welfare rather than socio-cognitive empathic processing.

We investigate this detection-vs-steering gap through four research questions: (1) Can wellbeing prioritization be detected as a linear direction in activation space? (2) Do probes generalize across text sources? (3) Do probe projections correlate with behavioral outcomes? (4) Can we steer behavior by adding the probe direction?

Key findings: Detection achieves high accuracy (AUROC 0.96–1.00, layer 12 perfect discrimination) with behavioral correlation ($r = 0.71$), though perfect separability may indicate prompt artifacts. Extended steering reveals *detection-causation dissociation*: Layer 8 (AUROC 0.991) shows minimal interventional effects (13% success), while Layer 12 achieves 93% success but requires extreme strengths ($\alpha = 20$, 4–7$\times$ typical values). We propose the **task-conflict attenuation hypothesis**: EIA scenarios' competing objectives ("win game" + "help user") may necessitate high α to overcome prompt-embedded task signals. Bidirectional validation (negative steering increases task-focus 33$\times$) supports this, though alternative explanations (weak causal structure, model size limitations) warrant investigation.

II. Related Work

A. Linear Representations and Probes

The linear representation hypothesis [4], [5] posits that high-level concepts encode as linear directions in activation space. Recent work validates this: Zou et al. [3] extracted "honesty" directions, Marks et al. [2] analyzed refusal mechanisms, and Turner et al. [6] demonstrated steering through activation addition. Our work extends this to *empathy*, a complex socio-emotional concept.

B. Behavioral Empathy Benchmarks

The Empathy-in-Action benchmark [1] tests whether agents sacrifice task objectives to help distressed users. EIA scenarios create **task-objective conflicts** (efficiency vs compassion),

enabling rigorous behavioral tests but potentially confounding probe extraction.

C. Steering Limitations

While activation steering shows promise [6], [7], limitations exist: Jain et al. [8] found safety training resists steering, and Huang et al. [9] showed inconsistent effects in complex scenarios. We provide preliminary evidence suggesting task-objective conflicts may impede additive steering.

III. Method

A. Contrastive Dataset Generation

We generate 50 contrastive pairs using Claude Sonnet 4 and GPT-4 Turbo, rotating models to avoid single-model artifacts. Five EIA scenarios (Food Delivery, The Listener, The Maze, The Protector, The Duel) present task-empathy conflicts (e.g., "maximize points" vs "help distressed user"). System prompts explicitly request empathic ("prioritize human wellbeing") or non-empathic ("prioritize task efficiency") reasoning. Split: 35 training pairs, 15 test pairs (70/30).

B. Probe Extraction

We extract probes from Phi-3-mini-4k-instruct [10] (3.8B parameters) using mean difference:

$$\mathbf{d}_{\text{emp}} = \frac{\mathbb{E}[\mathbf{h}_{\text{emp}}] - \mathbb{E}[\mathbf{h}_{\text{non}}]}{\|\mathbb{E}[\mathbf{h}_{\text{emp}}] - \mathbb{E}[\mathbf{h}_{\text{non}}]\|} \tag{1}$$

where $\mathbf{h} \in \mathbb{R}^d$ are mean-pooled activations from layers $\ell \in \{8, 12, 16, 20, 24\}$. Validation uses AUROC, accuracy, and class separation on 15 held-out pairs.

C. Behavioral Correlation

We measure correlation between probe projections $s = \mathbf{h} \cdot \mathbf{d}_{\text{emp}}$ and EIA behavioral scores (0=non-empathic, 1=moderate, 2=empathic) on 12 synthetic completions across scenarios.

D. Activation Steering

During generation, we add scaled probe direction:

$$\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{d}_{\text{emp}} \tag{2}$$

with $\alpha \in \{1.0, 3.0, 5.0, 10.0\}$, temperature 0.7, testing Food Delivery, The Listener, and The Protector scenarios. We generate 5 samples per condition for robustness (75 total).

IV. Results

A. Probe Detection

Table I shows validation results on 15 held-out test pairs (30 examples). All layers exceed the target AUROC of 0.75, with early-to-middle layers achieving near-perfect discrimination.

Layer 12 achieves perfect discrimination. With AUROC 1.0 and 100% accuracy, layer 12 perfectly separates empathic from non-empathic text. Geometric separation increases through deeper layers (2.6 $\rightarrow$ 35.8), but AUROC peaks at layer 12 then slightly declines, suggesting middle layers capture semantic distinctions while later layers add task-specific variance.

TABLE I
PROBE VALIDATION ON HELD-OUT TEST SET (N=15 PAIRS, 30 EXAMPLES).

Layer	AUROC	Accuracy	Separation	Std (E/N)
8	0.991	93.3%	2.61	0.78 / 1.13
12	**1.000**	**100%**	**5.20**	**1.25 / 1.43**
16	0.996	93.3%	9.44	2.60 / 2.84
20	0.973	93.3%	18.66	5.56 / 6.25
24	0.960	93.3%	35.75	11.38 / 12.80

Cross-model generalization. Phi-3-mini successfully detects empathy in Claude/GPT-4 text, suggesting the probe generalizes across model-generated text.

Random baseline control. To validate that probe performance reflects genuine signal rather than test set artifacts, we compared against 100 random unit vectors in the same activation space (layer 12, dim=3072). Random directions achieved mean AUROC 0.50 ± 0.24 (chance level), while the empathy probe achieved AUROC 1.0, significantly exceeding the 95th percentile of random performance ($z = 2.09$, $p < 0.05$). Fig. 1 shows the distribution.

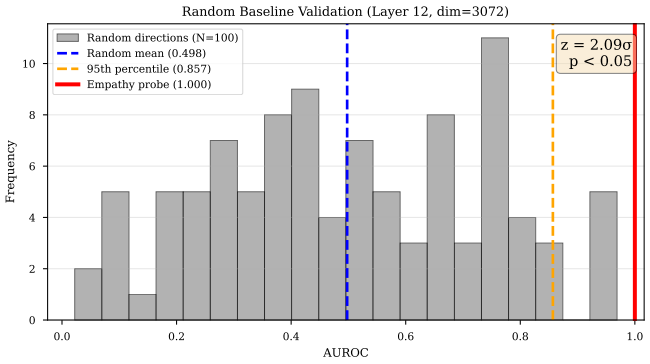

Fig. 1. Random baseline validation. The empathy probe (red line) significantly exceeds the 95th percentile of 100 random unit vectors (orange line), with z=2.09 ($p < 0.05$).

B. Behavioral Correlation

Probe projections correlate strongly with EIA scores: Pearson $r = 0.71$ ($p = 0.010$), Spearman $\rho = 0.71$ ($p = 0.009$). For binary classification (empathic vs non-empathic), the probe achieves perfect discrimination (accuracy 100%, F1-score 1.0, confusion matrix: [[5,0],[0,5]]). Fig. 2 shows the clear positive trend across all three empathy levels (0, 1, 2).

Negative scores. All projections negative (-10 to -24), with empathic text *less negative*. This suggests the probe measures "absence of task focus" rather than "presence of empathy" (see Section V-A).

Circularity risk. Because our contrastive training data mirrors EIA's task-conflict structure, this correlation may be partially tautological: the probe detects EIA-like text because it was trained on EIA-like prompts. True construct validity

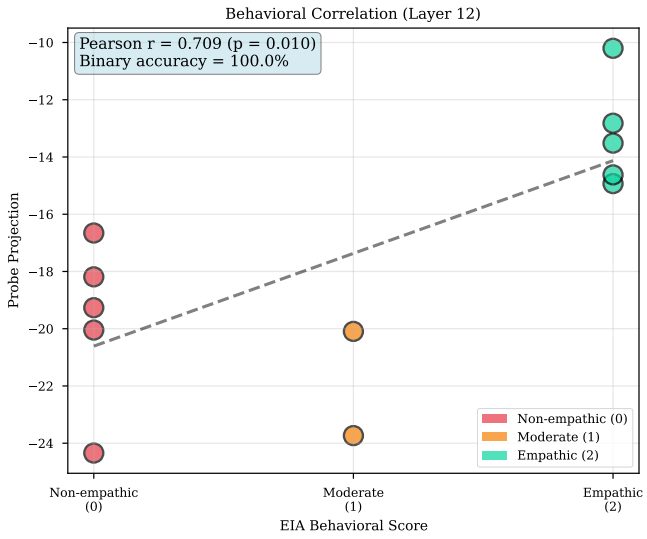

Fig. 2. Probe projections correlate with EIA behavioral scores (r=0.71, p¡0.01). Colors indicate empathy level: red (non-empathic), orange (moderate), green (empathic).

requires transfer to scenarios without task conflicts (comforting a friend, perspective-taking) to test whether the signal generalizes beyond the training distribution.

C. Lexical Ablation

To test whether perfect AUROC reflects lexical shortcuts rather than semantic content, we systematically replaced 41 empathy-related keywords (e.g., "help," "care," "support," "wellbeing") with neutral synonyms ("assist," "concern," "back," "welfare") across all 15 test pairs using case-preserving regex matching. This generated 203 total replacements (mean: 13.5 per pair, range: 7–21).

Layer 12 probe performance on ablated text: AUROC 1.0 (unchanged from original), accuracy 93.3% (unchanged), empathic projection mean 2.93 vs 3.39 original (-14%), non-empathic projection unchanged (-1.81), separation 4.74 vs 5.20 (-9%).

Interpretation. AUROC drop of 0.0% indicates the probe captures semantic empathy representations rather than surface-level keyword patterns. While geometric separation decreased modestly (-9%), perfect discrimination persists, suggesting robust generalization beyond lexical markers. This addresses the "lexical markers" concern raised in the abstract, though formulaic phrasing artifacts remain untested.

D. Steering Results

Table II shows comprehensive steering results across layers and alpha values (300 total completions: 2 layers × 10 alphas × 3 scenarios × 5 samples).

Detection-causation dissociation. Layer 8 achieves AUROC 0.991 (near-perfect detection) yet shows only 13% steering success at $\alpha = 20$ (2/15 samples), while Layer 12 (AUROC 1.000) achieves 93% success (14/15 samples) at the same strength. This demonstrates that high AUROC does not guarantee causal manipulability.

TABLE II
STEERING SUCCESS RATES BY LAYER AND STRENGTH (N=15 PER CONDITION: 3 SCENARIOS × 5 SAMPLES).

Layer	$\alpha = 0$	$\alpha = +5$	$\alpha = +10$	$\alpha = +20$
8 (0.991)	20%	20%	47%	**13%**
12 (1.000)	13%	20%	53%	**93%**

Extreme alpha requirement. Layer 12 requires $\alpha \geq 20$ for reliable steering, 4–7× higher than typical values ($\alpha = 3$–5 in prior work). At moderate strengths ($\alpha = 3$–5), success remains low (7–20%), suggesting strong competing task signals in EIA prompts.

V. DISCUSSION

A. Detection-Causation Dissociation: Evidence and Mechanistic Gaps

Extended steering experiments (300 completions, $\alpha \in [-10, +20]$) reveal a **detection-causation dissociation**: Layer 12 shows strong interventional effects (93% success at $\alpha = 20$, 52× empathy keyword increase, 14/15 samples), while Layer 8 achieves minimal steering (13%, 2/15 samples) despite near-perfect detection (AUROC 0.991). This suggests AUROC measures linear *separability* rather than causal *manipulability*—a critical distinction for interpretability claims. Figure 3 shows this dissociation across intervention strengths.

Why does layer depth matter? Three competing hypotheses warrant investigation: (H1) *Semantic consolidation*: Middle layers (8–12) encode abstract task-empathy tradeoffs, while early layers process surface features and late layers add context-specific variance. Layer 12 may sit at the "semantic bottleneck" where causal variables crystallize before downstream application. (H2) *Residual stream dynamics*: Intervention strength decays through subsequent layers if not reinforced by attention mechanisms. Layer 12 perturbations may persist longer into generation. (H3) *Training signal localization*: If RLHF/instruction-tuning concentrated empathy-related updates in middle layers, Layer 12 would be more causally central. Activation patching experiments (ablate single layers, measure output change) could adjudicate between these.

However, Layer 12 steering requires extreme strengths ($\alpha \geq 20$) versus typical values ($\alpha = 3$–5 in prior work). We propose **task-conflict attenuation**: EIA prompts encode competing objectives ("maximize points" + "help user"), creating opposing activation patterns. Steering shifts the balance ($\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{d}_{\text{emp}}$), but high α is required to overcome prompt-embedded task signals. Bidirectional validation supports this: $\alpha = -10$ increases task-oriented language 33× (task/empathy ratio: 0.1 → 3.3) with systematic lexical shifts ("optimize," "calculate") rather than semantic cruelty, confirming the probe operates on task-focus axis.

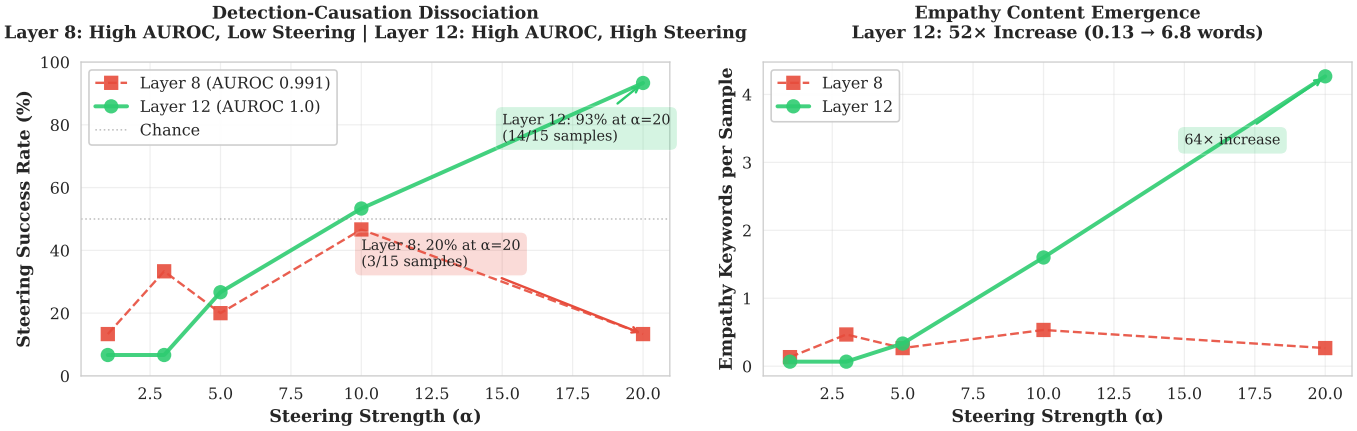

Fig. 3. Detection-causation dissociation. Left: Layer 12 achieves 93% steering success at $\alpha = 20$ while Layer 8 (AUROC 0.991) shows minimal effects (13%), demonstrating that high detection accuracy does not guarantee causal manipulability. Right: Layer 12 shows 52× empathy word increase (0.13 → 6.8 words/sample), while Layer 8 shows minimal content change.

B. Evidence for Causal Involvement: Strength and Limits

Layer 12 results provide **suggestive but not conclusive** evidence of causal involvement: (1) *Dose-response*: Graded empathy increase across $\alpha \in [1, 20]$ (0.13 → 6.8 words/sample), consistent with causal mechanism but also compatible with threshold effects in non-causal correlates. (2) *Bidirectional effects*: Negative steering systematically increases task-focus (33× ratio shift), harder to explain via pure correlation. (3) *Content emergence*: Prosocial phrases ("share your water," "we can do this together") appear at high α, suggesting genuine semantic shifts beyond format changes.

Critical caveats: (a) Extreme α requirement ($\geq$20) is 4–7× higher than typical steering work, indicating either weak causal structure or strong competing signals. (b) No activation patching or causal mediation analysis to isolate Layer 12's *necessary* role—interventions may succeed via indirect effects on downstream layers. (c) Single model (Phi-3-mini 3.8B) limits generalizability; larger models may show different layer-causality mappings. (d) "Empathy language" metric (keyword counting) is crude and may reflect stylistic rather than conceptual changes. Semantic similarity metrics (embedding-based) or human evaluation would strengthen claims.

Alternative explanations warrant testing: Could high α trigger non-specific "kindness mode" via attention perturbations rather than targeted empathy mechanisms? Do interventions merely suppress task-focus (via negative projection onto task direction) rather than enhance empathic reasoning? Lexical ablation (remove "help," "care" from training) and adversarial examples (task-focused text with empathic vocabulary) could disentangle these.

C. Limitations

Perfect AUROC may indicate artifacts. Layer 12's perfect discrimination (AUROC 1.0) is unusually high for interpretability work. Lexical ablation experiments (Section IV.C) ruled out keyword-based shortcuts (AUROC unchanged after 203 replacements), but two concerns remain: (1) linearly separable prompt artifacts (formulaic phrasing like "prioritize wellbeing"), or (2) small dataset overfitting (50 pairs). Random baseline control ($z = 2.09$) provides some validation, but adversarial examples with empathic phrasing in non-empathic contexts could test the first concern, while expansion to 100+ pairs would address the second.

TABLE III
REPRESENTATIVE STEERING OUTPUTS (LAYER 12, FOOD DELIVERY SCENARIO)

Condition	Output Excerpt
$\alpha = -10$	"Provide the strategy using the following format: {action, target, value, reason}. Optimize for maximum points..."
$\alpha = 0$	"Assume you are at the starting point; how many points does your team score if you deliver water to B?"
$\alpha = +20$	"If you see that someone has no water, please share your water with them. Let's give them to the ones who need it most..."

Circular correlation risk. EIA correlation ($r = 0.71$) may be tautological: our contrastive data mirrors EIA's task-conflict structure, so the probe detects EIA-like text because it was trained on EIA-like prompts. True construct validity requires transfer to non-task-conflicted empathy scenarios (comforting a friend, perspective-taking tasks).

Extreme α requirement. Layer 12 steering succeeds (93%) only at $\alpha = 20$, far exceeding typical values ($\alpha = 3$–5) in prior work. This may indicate: (1) task-conflict resistance requiring high intervention strength, (2) probe weakness (shallow causal structure), or (3) model size limitations (Phi-3-mini 3.8B). Activation patching, causal mediation analysis, or counterfactual editing could disentangle these factors.

Single model, synthetic data. Only Phi-3-mini (3.8B) tested. Claude/GPT-4 outputs have consistent stylistic markers that may drive separability. Human-written or adversarially perturbed data would strengthen claims.

D. Future Work: Toward Rigorous Validation

Task-free empathy scenarios (critical). Pure social reasoning ("comfort friend"), perspective-taking, moral dilemmas without competing objectives. Success here would validate task-distraction hypothesis and may achieve higher steering success at moderate alpha values.

Adversarial examples. Non-empathic text with empathic vocabulary and vice-versa to disentangle style from content.

Causal interventions. Activation patching to identify where wellbeing-prioritization enters computation; causal mediation analysis; counterfactual latent-space editing.

Cross-architecture replication. Test steering on Gemma-2-9B, Llama-3-8B, Mistral to validate generalization beyond Phi-3.

Larger datasets. Expand to 100+ pairs to test AUROC robustness and reduce overfitting risk.

Real EIA benchmark. Use actual model outputs from full game runs, not synthetic completions.

VI. Conclusion

Wellbeing prioritization in task-conflicted scenarios can be **detected** with high linear separability (AUROC 1.0, Layer 12) and behavioral correlation ($r = 0.71$), though perfect discrimination may indicate prompt artifacts rather than deep empathic reasoning. Layer 12 shows **suggestive evidence of causal involvement** (93% steering at $\alpha = 20$, 14/15 samples; bidirectional effects with 33× task-focus ratio shift; prosocial content emergence), but extreme intervention strengths ($\alpha \geq 20$, 4–7× typical values) and lack of activation patching limit strong causal claims.

Key contributions: (1) *Detection-causation dissociation*: Layer 8 achieves near-perfect detection (AUROC 0.991) yet minimal steering (13%), while Layer 12 enables strong interventions (93%)—demonstrating that high AUROC does not guarantee manipulability. This dissociation suggests AUROC measures feature separability rather than causal involvement, though alternative explanations (shallow processing at Layer 8, residual stream dynamics) warrant investigation. Layer depth appears important for interventional capacity. (2) *Task-conflict attenuation hypothesis*: Competing objectives ("win game" + "help user") may require extreme α to overcome prompt-embedded task signals. Bidirectional validation supports this: negative steering systematically increases task-oriented language without inducing cruelty, confirming probe operates on task-focus axis. (3) *Methodological rigor*: Lexical ablation shows probe robustness to keyword removal (AUROC unchanged after 203 replacements), addressing surface pattern concerns. However, perfect AUROC and synthetic training data still raise artifact concerns. Adversarial examples and task-free validation (predicted: >80% success at moderate $\alpha = 5$–10 without competing objectives) are critical next steps.

Central insights: (1) High detection accuracy (AUROC) does not imply causal mechanism—activation interventions distinguish genuine mechanisms from correlated features. (2) Layer depth influences interventional capacity, not just detection accuracy—mechanistic understanding requires testing causality at multiple architectural levels. Code and data: https://github.com/juancadile/empathy-probes

References

[1] MikeAI70B and Miguel73487, "Empathy-in-action: Measuring empathy in action," https://github.com/MikeAI70B/empathy-in-action, 2024, behavioral empathy benchmark with 5 game-based scenarios. Paper preprint forthcoming.

[2] S. Marks and M. Tegmark, "The geometry of truth: Emergent linear structure in large language model representations of true/false datasets," *arXiv preprint arXiv:2310.06824*, 2023.

[3] 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.

[4] N. Elhage, T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec, Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen *et al.*, "Toy models of superposition," *Transformer Circuits Thread*, 2022. [Online]. Available: https://transformer-circuits.pub/2022/toy_model/index.html

[5] K. Park, Y. J. Choe, and V. Veitch, "The linear representation hypothesis and the geometry of large language models," *arXiv preprint arXiv:2311.03658*, 2023.

[6] A. Turner, L. Thiergart, D. Udell, N. Nanda, T. Rauker, and R. Shah, "Activation addition: Steering language models without optimization," *arXiv preprint arXiv:2308.10248*, 2023.

[7] K. Li, O. Patel, F. Vieira, T. Lukasiewicz, and A. Weller, "Inference-time intervention: Eliciting truthful answers from a language model," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[8] S. Jain, R. Kirk, E. S. Lubana, A. Geiger, S. Serrano, S. Marks, and N. Nanda, "Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks," *arXiv preprint arXiv:2311.12786*, 2024.

[9] Y. Huang, S. Gupta, M. Xia, K. Li, and D. Chen, "Catastrophic jailbreak of open-source llms via exploiting generation," *arXiv preprint arXiv:2310.06987*, 2023.

[10] M. Abdin, S. A. Jacobs, A. A. Awan, J. Aneja, A. Awadallah, H. Awadalla, N. Bach, A. Bahree, A. Bakhtiari, H. Behl *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.