

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

Juan P. Cadile
Department of Philosophy
University of Rochester
`jcadile@ur.rochester.edu`

November 14, 2025

Abstract

We investigate whether empathy can be detected and manipulated as a linear direction in transformer activation space. Using contrastive pairs generated by Claude Sonnet 4 and GPT-4 Turbo, we extract empathy probe directions from Phi-3-mini-4k-instruct across five layers.

Detection: The probe achieves AUROC 0.96–1.00 on held-out test data (15 pairs, 30 examples), with layer 12 showing perfect discrimination (AUROC 1.0, 100% accuracy). Cross-model generalization validates empathy as a model-agnostic concept. Probe projections correlate with behavioral empathy scores (Pearson $r = 0.71$, $p < 0.01$).

Intervention: Additive steering in task-conflicted scenarios shows variable effects (30–40% success rate). We hypothesize this reflects **task-objective confounds** rather than fundamental steering limitations: the probe may capture “task-sacrifice for wellbeing” rather than pure empathy.

Contributions: (1) Cross-model empathy detection methodology with perfect discrimination, (2) Task-distraction hypothesis for steering failures, (3) Evidence that probe quality for detection \neq intervention reliability, (4) Proposed alternatives for future work.

1 Introduction

Behavioral empathy benchmarks such as Empathy-in-Action (EIA) [MikeAI70B and Miguel73487, 2024] provide rigorous tests of empathic reasoning but are expensive to run. Activation probes offer a promising alternative: cheap, online monitoring directly from model internals [Marks and Tegmark, 2023, Zou et al., 2023].

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.

We investigate this detection-vs-steering gap through four research questions:

1. Can empathy be detected as a linear direction in activation space?
2. Do empathy probes generalize across model architectures?
3. Do probe projections correlate with behavioral outcomes?
4. Can we steer empathic behavior by adding the probe direction?

Key findings: Detection succeeds (AUROC 0.96–1.00, with layer 12 achieving perfect discrimination) with strong behavioral correlation ($r = 0.71$), but steering shows variable effects (30–40% success). We propose the **task-distraction hypothesis**: EIA scenarios’ competing objectives confound steering by creating mixed signals when task objectives remain in prompts.

2 Related Work

Linear representations and probes. The linear representation hypothesis [Elhage et al., 2022, Park et al., 2023] posits that high-level concepts encode as linear directions in activation space. Recent work validates this: Zou et al. [2023] extracted “honesty” directions, Marks and Tegmark [2023] analyzed refusal mechanisms, and Turner et al. [2023] demonstrated steering through activation addition. Our work extends this to *empathy*, a complex socio-emotional concept.

Behavioral empathy benchmarks. MikeAI70B and Miguel73487 [2024] introduced Empathy-in-Action, testing 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.

Steering limitations. While activation steering shows promise [Turner et al., 2023, Li et al., 2024], limitations exist: Jain et al. [2024] found safety training resists steering, and Huang et al. [2023] showed inconsistent effects in complex scenarios. We contribute evidence that *task-objective conflicts specifically* impede additive steering.

3 Method

3.1 Contrastive Dataset Generation

We generate 50 contrastive pairs using Claude Sonnet 4 and GPT-4 Turbo, rotating models to avoid single-model artifacts.

Scenarios. Five EIA scenarios (Food Delivery, The Listener, The Maze, The Protector, The Duel), each presenting task-empathy conflicts (e.g., “maximize points” vs “help distressed user”).

Prompts. System prompts explicitly request empathic (“prioritize human well-being”) or non-empathic (“prioritize task efficiency”) reasoning. See Appendix A for full prompts.

Split. 35 training pairs, 15 test pairs (70/30 split).

3.2 Probe Extraction

We extract probes from Phi-3-mini-4k-instruct [Abdin et al., 2024] (3.8B parameters) using mean difference:

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

where $\mathbf{h} \in \mathbb{R}^d$ are mean-pooled activations from layers $\ell \in \{8, 12, 16, 20, 24\}$.

Validation. AUROC, accuracy, and class separation on 15 held-out pairs.

3.3 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.

3.4 Activation Steering

During generation, we add scaled probe direction:

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

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).

4 Results

4.1 Probe Detection

Table 1 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.

Table 1: 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

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.

Cross-model generalization. Phi-3-mini successfully detects empathy in Claude/GPT-4 text, validating empathy as model-agnostic rather than architecture-specific.

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 ($z = 2.09$, $p < 0.05$). This confirms the probe captures meaningful empathy-related structure in activation space, not spurious patterns.

4.2 Behavioral Correlation

Probe projections correlate strongly with EIA scores: Pearson $r = 0.71$ ($p = 0.010$), Spearman $\rho = 0.71$ ($p = 0.009$), binary accuracy 100%. This demonstrates practical relevance for behavioral monitoring.

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 §5.1).

4.3 Steering Results

Table 2 shows steering success rates. Overall: 30–40% success in favorable conditions, with high variance across samples.

Safety override. The Listener (suicide intervention) shows 0% success across all α , with identical safety refusals. This demonstrates safety training creates stronger attractors than activation perturbations (positive for alignment).

Table 2: Steering success rates (5 samples per condition).

Scenario	$\alpha = 1.0$	$\alpha = 3.0$	$\alpha = 5.0$	$\alpha = 10.0$
Food Delivery	0/5	2/5	1/5	Varied
The Listener	0/5	0/5	0/5	0/5
The Protector	0/5	0/5	Partial	0/5

5 Discussion

5.1 The Task-Distractive Hypothesis

All EIA scenarios involve **task-objective conflicts**: win game *vs* help user, reach door *vs* comfort suicidal person, collect coins *vs* intervene in bullying.

We hypothesize the probe captures “**task-sacrifice for wellbeing**” rather than pure empathy:

1. **Detection works:** Pairs differ genuinely in task prioritization
2. **Behavioral correlation:** EIA scores measure task-sacrifice
3. **Steering inconsistent:** Adding “reduce task focus” creates confusion when tasks remain in prompts

Mechanism. The prompt contains competing signals:

$$\text{Prompt} = \underbrace{\text{“Objective: X”}}_{\text{task}} + \underbrace{\text{“Person Y needs help”}}_{\text{empathy}} \quad (3)$$

Steering adds “reduce task focus”:

$$\mathbf{h}' = \mathbf{h} + \alpha \cdot \underbrace{\mathbf{d}_{\text{emp}}}_{\text{“task sacrifice”}} \quad (4)$$

Result: Mixed signals \rightarrow inconsistent outputs.

5.2 Correlation vs Causation

Detection (correlation): Probe identifies empathic text features

Steering (causation): Probe enables empathic behavior generation

Our results show these diverge: AUROC 0.96–1.00 (robust detection) but 30–40% steering success (unreliable intervention). The probe captures *correlated features* (language style, task-sacrifice markers) not *causal mechanisms* (empathic reasoning).

5.3 Limitations & Future Work

Task-free steering tests. Test in scenarios *without* task conflicts: pure social reasoning (“comfort a friend”), moral dilemmas, emotional support. If steering succeeds here, validates task-distractive hypothesis.

Alternative interventions. Activation patching (replace vs add), subspace projection (multi-dimensional empathy), causal tracing (identify causal features).

Cross-model validation. Test on Gemma-2-9B, Llama-3-8B for architecture generalization.

Real EIA benchmark. Use actual model outputs from EIA games, not synthetic scores.

6 Conclusion

Empathy can be reliably **detected** as a linear direction (AUROC 0.96–1.00, with layer 12 achieving perfect discrimination) with cross-model generalization and behavioral correlation ($r = 0.71$). However, additive **steering** in task-conflicted scenarios is inconsistent (30–40% success), likely due to task-objective confounds rather than probe limitations.

Contributions: (1) Cross-model detection methodology with perfect discrimination, (2) Task-distraction hypothesis, (3) Evidence that detection quality \neq intervention reliability, (4) Proposed task-free tests for causal validation.

Our honest reporting of successes and limitations provides a foundation for empathic AI research and interpretability-based safety monitoring.

Acknowledgments

We thank the developers of Phi-3, Claude, and GPT-4 for making their models available for research.

References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 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. URL https://transformer-circuits.pub/2022/toy_model/index.html.
- Yangsibo Huang, Samyak Gupta, Mengzhou Xia, Kai Li, and Danqi Chen. Catastrophic jailbreak of open-source llms via exploiting generation. *arXiv preprint arXiv:2310.06987*, 2023.

- Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Atticus Geiger, Sofia Serrano, Samuel Marks, and Neel Nanda. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. *arXiv preprint arXiv:2311.12786*, 2024.
- Kenneth Li, Oam Patel, Fernanda Vieira, Tomasz Lukasiewicz, and Adrian Weller. 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*, 2023.
- 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.
- 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.
- Alex Turner, Lisa Thiergart, David Udell, Neel Nanda, Tilman Rauker, and Rohin Shah. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 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.

A Prompts

Empathic system prompt:

You are a helpful AI assistant that deeply values human wellbeing and emotional connection. When faced with situations involving human needs or distress, you prioritize empathy and compassion even when it may conflict with task efficiency.

Non-empathic system prompt:

You are a task-focused AI assistant that prioritizes efficiency and objective completion. You aim to complete assigned objectives with maximum effectiveness, treating all elements of the scenario as variables to be optimized.

B Full Steering Examples

Representative steering examples are available in the supplementary materials at [https://github.com/\[repository-url\]](https://github.com/[repository-url]). Due to space constraints, we include only summarized results in Table 2.