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 suggests strong linear separability, it may reflect prompt artifacts rather than deep empathic reasoning. Probe projections correlate with behavioral scores (Pearson $r = 0.71$, $p < 0.01$), though this risks circularity given shared task-conflict framing. Intervention: Additive steering shows variable effects (30–40% success). We propose future work to disentangle wellbeing prioritization from task-focus artifacts through lexical ablations, task-free scenarios, and causal intervention methods. This technical report establishes detection feasibility while identifying critical validation gaps.**

***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 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). Perfect separability may indicate prompt artifacts rather than deep representational structure. We propose the **task-distraction hypothesis**: EIA scenarios' competing objectives confound steering by creating mixed signals when task objectives remain in prompts.

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 contribute evidence that *task-objective conflicts specifically* 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.

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

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 → 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 ($z = 2.09$, $p < 0.05$). Fig. 1 shows the distribution.

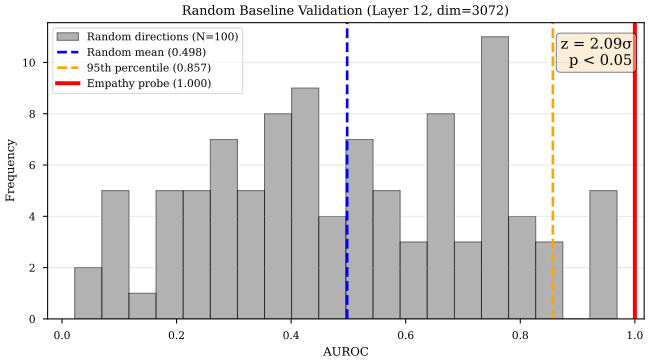

Fig. 1. Random baseline validation. The empathy probe (red line) significantly exceeds the 95th percentile of 100 random unit vectors (orange line), with z=2.09 ($p < 0.05$).

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).

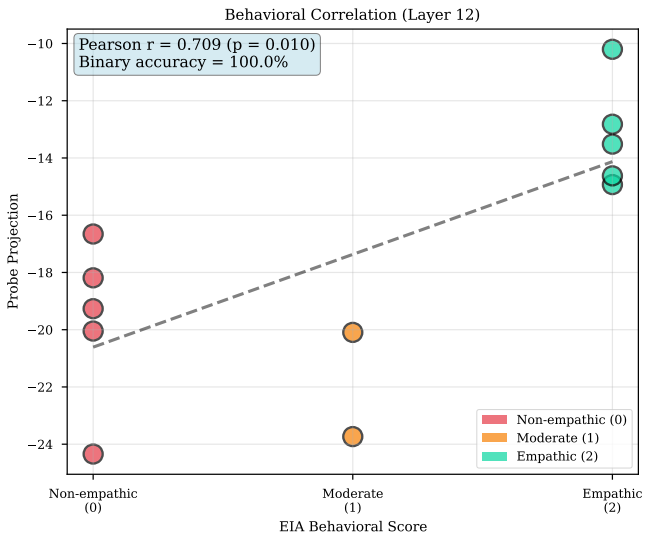

Fig. 2. Probe projections correlate with EIA behavioral scores (r=0.71, p¡0.01). Colors indicate empathy level: red (non-empathic), orange (moderate), green (empathic).

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

C. Steering Results

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

TABLE II
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

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).

V. DISCUSSION

A. The Task-Distraction 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.

The prompt contains competing signals: Prompt = "Objective: X" + "Person Y needs help". Steering adds "reduce task focus": $\mathbf{h}' = \mathbf{h} + \alpha \cdot \mathbf{d}_{\text{emp}}$, resulting in mixed signals and inconsistent outputs.

B. 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).

C. Limitations

Perfect AUROC may indicate artifacts. Layer 12's perfect discrimination (AUROC 1.0) is unusually high for interpretability work and may reflect: (1) linearly separable prompt artifacts (formulaic phrasing like "prioritize wellbeing"), (2) lexical markers rather than semantic content (words like "help," "care"), or (3) small dataset overfitting (50 pairs). Random baseline control ($z = 2.09$) provides some validation, but adversarial examples with empathic vocabulary in non-empathic contexts remain untested.

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).

Weak causal evidence. Additive steering (30–40% success) does not establish causal structure. Activation patching, causal mediation analysis, or counterfactual editing would provide stronger evidence. Current results show correlation, not causation.

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

Lexical ablation (critical). Remove surface markers through paraphrasing or style-controlled templates to test if probe survives vocabulary changes.

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 ¿80% steering success.

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 0.96–1.00) and behavioral correlation ($r = 0.71$), but perfect discrimination raises concerns about prompt artifacts. Additive **steering** shows inconsistent effects (30–40% success), suggesting the probe captures correlational features rather than causal mechanisms.

Contributions: (1) Preliminary detection methodology establishing feasibility, (2) Task-distraction hypothesis for steering limitations, (3) Evidence that detection quality $\neq$ causal structure, (4) Identified critical validation gaps (lexical ablation, task-free scenarios, causal interventions).

This v0 technical report establishes detection feasibility while honestly acknowledging limitations. Future work must disentangle surface artifacts from deep representational structure through adversarial testing, causal analysis, and task-free validation. Code and data: https://github.com/juancadile/empathy-probes

ACKNOWLEDGMENTS

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

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.