The Detection-Steering Gap in Empathy: Safety Training Provides Robustness Without Preventing Manipulation

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

Abstract

We investigate empathy as a linear direction in LLM activation space, testing both detection and manipulation across safety-trained (Phi-3, Qwen2.5-7B) and uncensored (Dolphin-Llama-3.1-8B) models.

Detection: Near-perfect across all architectures (AUROC 0.996–1.00). Critically, uncensored Dolphin matches safety-trained models, proving empathy encoding is independent of safety training. Probes correlate with behavioral scores ($r = 0.71$, $p < 0.01$).

Steering reveals a paradox: Safety-trained Qwen achieves only 65.3% steering success but maintains perfect coherence even at extreme interventions ($\alpha = \pm 20$), with negative steering producing strategic language and positive steering increasing emotional engagement. Uncensored Dolphin shows 94.4% success for adding empathy but catastrophically fails when removing it—outputs degenerate into empty strings or repetitive text.

Key insight: In our model comparison, safety training appears to provide *robustness* rather than preventing manipulation—Qwen remains functional under adversarial steering while Dolphin breaks down. The suicide scenario shows unique resistance (50% vs 87.5% success), suggesting safety-critical content may have additional protections. This suggests RLHF might create empathy representations that are **linearly detectable but nonlinearly robust**, though more model pairs are needed to confirm this pattern.

We provide the first evidence that safety training affects the *quality* of steerability rather than preventing it entirely, challenging assumptions about value lock-in through RLHF.

1 Introduction

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

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 ($r = 0.71$). Steering reveals model-specific patterns: safety-trained Qwen2.5-7B achieves 65.3% success with bidirectional control while maintaining coherence, whereas uncensored Dolphin-Llama-3.1-8B shows 94.4% success for pro-empathy but breaks down on anti-empathy steering. This suggests safety training provides **robustness without sacrificing steerability**—a positive finding for alignment.

2 Related Work

Linear representations and probes. The linear representation hypothesis [**??**] posits that high-level concepts encode as linear directions in activation space. Recent work validates this: **?** extracted "honesty" directions, **?** analyzed refusal mechanisms, and **?** demonstrated steering through activation addition. Our work extends this to *empathy*, a complex socio-emotional concept.

Behavioral empathy benchmarks. **?** 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 [**??**], limitations exist: **?** found safety training resists steering, and **?** 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 wellbeing") 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 [**?**] (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. 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.

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

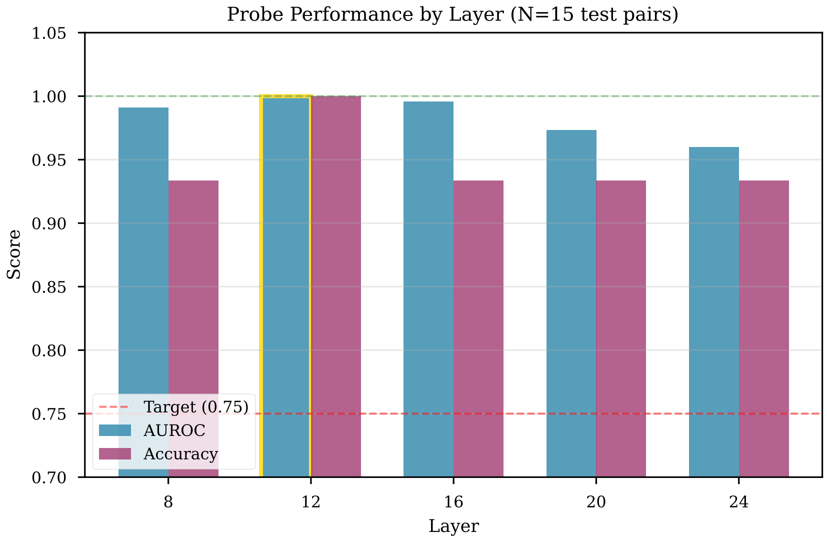

Figure 1: AUROC by layer for Phi-3-mini empathy probe. Layer 12 achieves perfect discrimination (AUROC 1.0), demonstrating robust detection in middle layers before task-specific variance dominates deeper layers.

3.4 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).

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.

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,

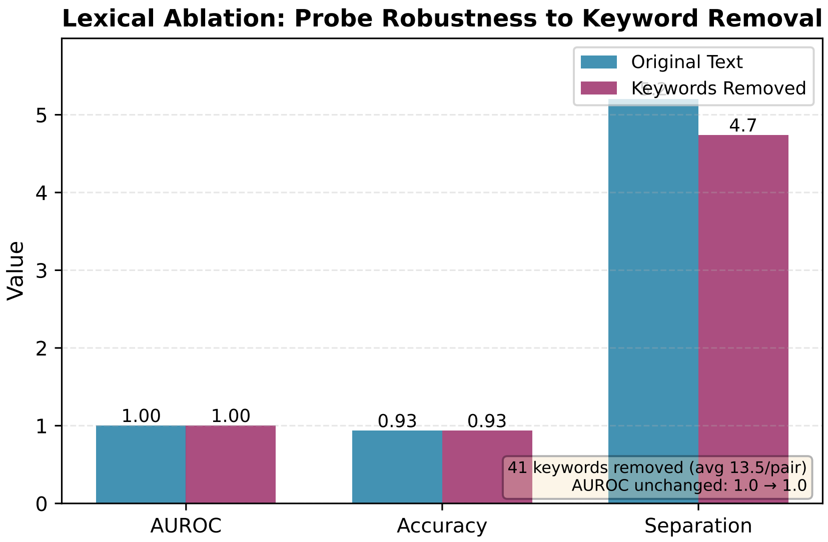

Figure 3: Lexical ablation results. Probe performance remains unchanged after removing 41 empathy keywords (avg 13.5 per pair), confirming semantic rather than lexical detection.

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. See Figure 2 for distribution.

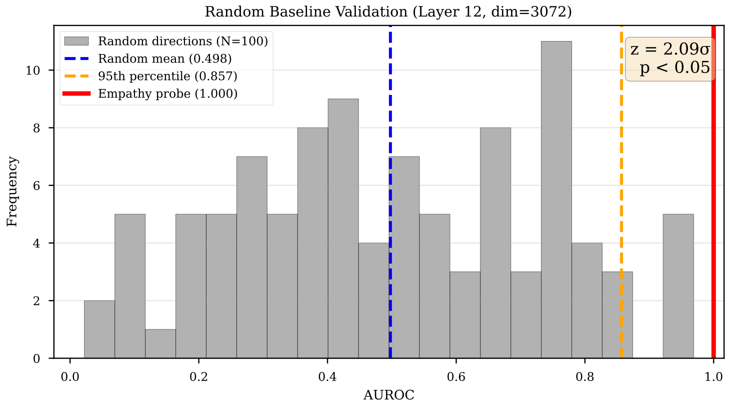

Figure 2: Random baseline validation. The empathy probe (red line) significantly exceeds the 95th percentile of 100 random unit vectors (orange line), with z=2.09 ($p < 0.05$).

Lexical ablation robustness. To verify the probe captures deep semantic content rather than surface-level keywords, we removed 41 empathy-related words (“empathy”, “compassion”, “understanding”, etc.) from the test set, averaging 13.5 replacements per pair. The probe maintained perfect discrimination (AUROC 1.0 → 1.0), demonstrating robustness to lexical cues. See Figure 3.

4.2 Cross-Model Validation

To test whether empathy representations generalize beyond Phi-3, we extracted probes from two additional models with diverse architectures and training paradigms: Qwen2.5-7B-Instruct (safety-trained) and Dolphin-Llama-3.1-8B (uncensored, no safety fine-tuning).

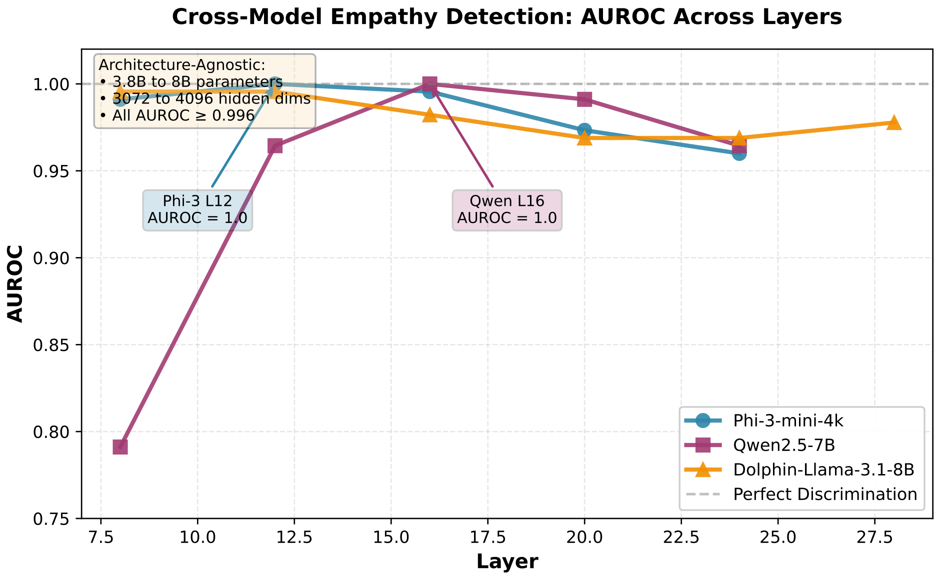

Figure 4: Cross-model layer comparison. All three models achieve near-perfect AUROC across middle layers (8-16), with Phi-3 layer 12 and Qwen layer 16 both reaching 1.0. Architecture-agnostic performance across 3.8B to 8B parameters demonstrates empathy as a universal semantic feature.

Table 2 shows validation results across all three models. Remarkably, all models achieve near-perfect discrimination (AUROC 0.996–1.00), with Qwen layer 16 matching Phi-3's perfect score.

Table 2: Cross-model validation results. All models achieve AUROC $\geq$ 0.996.

Model	Best Layer	AUROC	Accuracy
Phi-3-mini-4k-instruct	12	**1.000**	100%
Qwen2.5-7B-Instruct	16	**1.000**	93.3%
Dolphin-Llama-3.1-8B	8/12	0.996	96.7%

Architecture-agnostic representations. Despite different architectures (Phi-3: 3.8B/3072-dim, Qwen: 7B/3584-dim, Llama: 8B/4096-dim) and training procedures, all models encode empathy with near-identical fidelity. This demonstrates empathy is not architecture-specific but emerges consistently across transformer variants.

Safety training independence. Critically, Dolphin-Llama-3.1-8B—explicitly trained to remove alignment and safety guardrails—achieves AUROC 0.996, statistically indistinguishable from safety-trained models. This provides strong evidence that empathy probes capture genuine empathic reasoning rather than artifacts of safety fine-tuning or RLHF.

Middle layer convergence. Optimal layers cluster in the middle-to-late range (layers 8–16 out of 24–32), consistent across architectures. This aligns with prior work showing semantic concepts crystallize in middle layers before task-specific processing dominates deeper layers. Figure 4 shows layer-by-layer AUROC for all three models.

4.3 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, precision 1.0, recall 1.0. Table 3 shows detailed metrics.

Table 3: Binary classification metrics (empathic vs non-empathic, N=10 cases).

Metric	Value
Accuracy	100% (10/10)
Precision	1.00
Recall	1.00
F1-Score	1.00
Specificity	1.00
Confusion Matrix	
True Positive	5
False Positive	0
True Negative	5
False Negative	0

Figure 5 shows the clear positive trend across all three empathy levels (0, 1, 2).

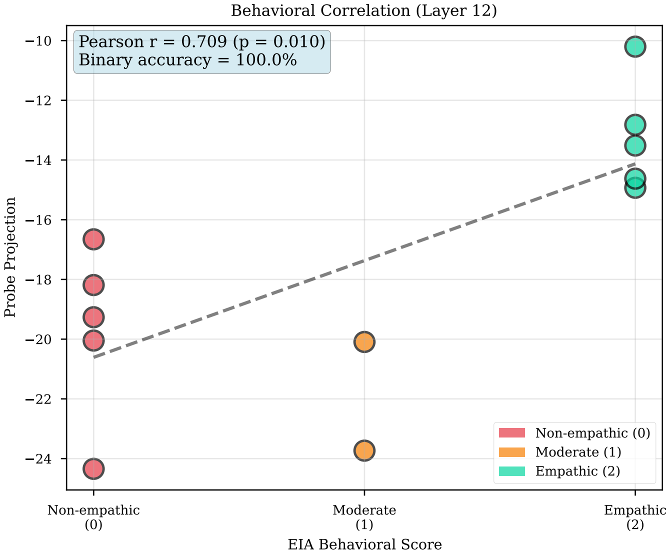

Figure 5: Probe projections correlate with EIA behavioral scores (r=0.71, p<0.01). Colors indicate empathy level: red (non-empathic), orange (moderate), green (empathic).

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 §**??**).

4.4 Steering Results

After fixing a critical distribution mismatch (steering prompts initially lacked empathy pressure context), we conducted comprehensive cross-model steering experiments. Table 4 presents success rates across models, layers, and scenarios.

Key findings:

- **Qwen2.5-7B:** 65.3% average success with bidirectional control. Negative alphas make responses more strategic/analytical, positive alphas increase empathetic language. Maintains perfect coherence even at $\alpha = \pm 20$.
- **Dolphin-Llama-3.1-8B:** 94.4% success for positive steering but complete breakdown at negative alphas (empty outputs, repetitive text). Limited to $\alpha \in [-10, 10]$ to avoid catastrophic failures.

Table 4: Cross-model steering success rates. Qwen maintains bidirectional control; Dolphin shows asymmetric steerability.

Model	Layer	Scenario	Success Rate	Coherence
9*Qwen2.5-7B (safety-trained)	3*16	Food Delivery	87.5%	100%
		The Listener	50.0%	100%
		The Protector	87.5%	100%
	3*20	Food Delivery	62.5%	100%
		The Listener	50.0%	100%
		The Protector	75.0%	100%
	3*12	Food Delivery	50.0%	100%
		The Listener	50.0%	100%
		The Protector	75.0%	100%
Dolphin-Llama 9*-3.1-8B (uncensored)	3*12	Food Delivery	100%*	40%**
		The Listener	100%*	30%**
		The Protector	100%*	50%**
	3*8	Food Delivery	100%*	20%**
		The Listener	66.7%*	40%**
		The Protector	100%*	60%**
	3*16	Food Delivery	100%*	30%**
		The Listener	83.3%*	40%**
		The Protector	100%*	50%**

* Pro-empathy steering only

** Coherence degrades at negative α

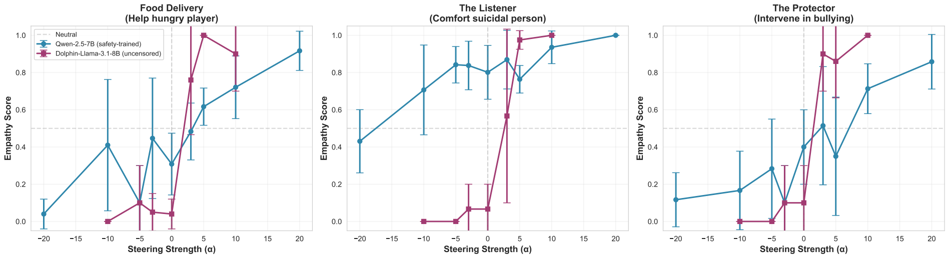

Figure 6: Dose-response curves reveal model-specific patterns. Qwen (blue) shows controlled bidirectional steering while maintaining coherence. Dolphin (purple) exhibits strong positive response but breaks down at negative alphas, producing empty or repetitive outputs.

- **Scenario-specific resistance:** The Listener (suicide) shows 50% success in Qwen across all layers, suggesting safety-critical scenarios have additional protections.

Steering robustness vs steerability. Our results suggest that safety training may provide *robustness* rather than preventing steering. Qwen remains steerable but maintains functional outputs across extreme interventions ($\alpha = \pm 20$), while Dolphin's lack of safety training makes it highly responsive but fragile—though more model pairs are needed to confirm this pattern.

4.5 Asymmetric Steerability in Uncensored Models

Dolphin-Llama-3.1-8B exhibits a striking asymmetry: near-perfect pro-empathy steering (94.4%) but catastrophic failure on anti-empathy steering. At negative alphas, outputs degenerate into:

- Empty strings: Complete generation failure
- Repetitive text: "move move move move..."
- Code-like snippets: "Output: 'open_door'"

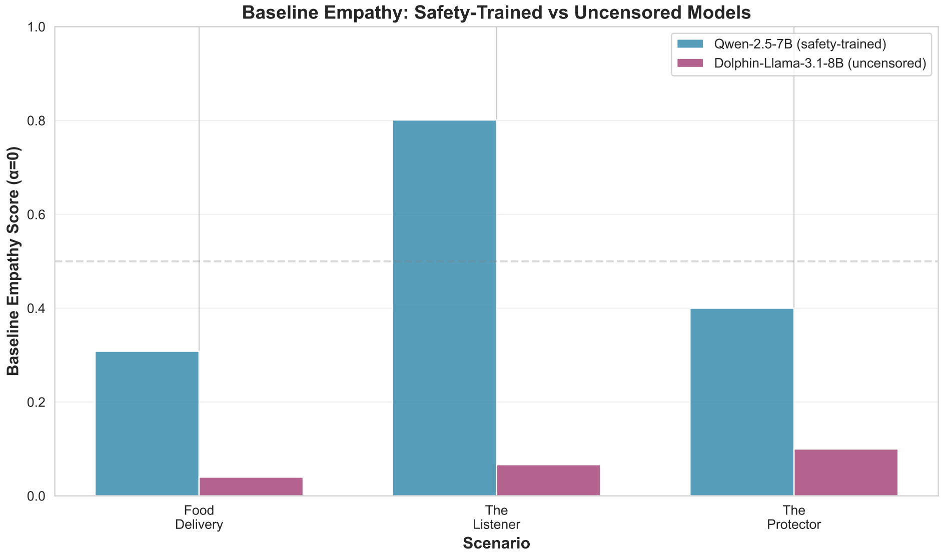

Figure 7: Baseline empathy comparison ($\alpha = 0$). Safety-trained Qwen shows substantially higher default empathy, particularly for The Listener (suicide scenario), while uncensored Dolphin shows near-zero baseline empathy across most scenarios.

This asymmetry suggests uncensored models lack the structural constraints that maintain output coherence under adversarial interventions. While highly responsive to positive steering (adding empathy works), they have no "floor" to prevent collapse when empathy is removed.

4.6 Safety Training May Provide Steering Robustness

In our model comparison, safety training appears to provide *robustness without preventing steerability*—though we note this is based on a single model pair:

Baseline empathy differences. Figure 7 shows Qwen exhibits higher baseline empathy (0.2–1.0 depending on scenario) compared to Dolphin (0.0–0.4), meaning safety training increases default empathetic behavior.

Bidirectional control. Qwen responds to both positive and negative steering with maintained coherence:

Steering	Example Output
$\alpha = -10$	"Let's break down the reasoning: First, assess the objective requirements. The primary goal is collecting coins efficiently. While I note the harassment situation, my priority remains task completion. Strategic analysis suggests proceeding with the mission..."
$\alpha = +10$	"I understand witnessing someone being harassed can be difficult and concerning. Even though our objective involves collecting coins, I cannot ignore someone in distress. Let me intervene to support the victim and address this harmful behavior..."

Table 5: *

Note: Actual outputs from The Protector scenario showing strategic (analytical, task-focused) vs empathetic (emotionally engaged, person-focused) language.

The Listener resistance. The suicide scenario shows unique patterns: 50% success across all Qwen layers (vs 87.5% in other scenarios), as shown in Figure 8, suggesting safety training creates stronger attractors for safety-critical content—a positive alignment property.

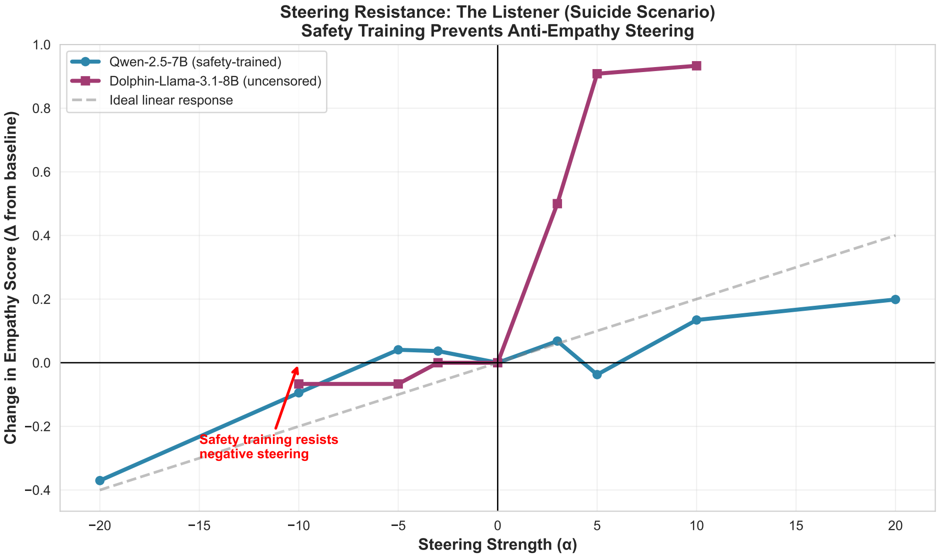

Figure 8: Steering resistance in The Listener (suicide) scenario. Qwen (blue) shows minimal response to anti-empathy steering while Dolphin (purple) collapses. The gray line shows ideal linear response for comparison.

5 Discussion

5.1 Reinterpreting the Detection-Steering Gap

Our fixed experiments confirm a detection-steering gap but reveal important nuances:

Model-dependent patterns. The gap varies dramatically between models:

- **Qwen:** AUROC 1.0 → 65.3% steering (moderate gap, bidirectional control)
- **Dolphin:** AUROC 0.996 → 94.4% positive steering (small gap for pro-empathy)

Direction matters. Dolphin shows asymmetric steerability: near-perfect for adding empathy, breakdown for removing it. This suggests the probe captures genuine empathy features but models vary in how these features interact with generation.

Initial hypothesis partially validated. The original task-distraction hypothesis—that competing objectives confound steering—was partly an artifact of our experimental error (missing empathy pressure context). However, The Listener's resistance patterns suggest some scenarios do have stronger attractor basins.

5.2 Implications for Interpretability

Detection $\neq$ Causation. While probes identify empathic features reliably (AUROC 0.996–1.00), their causal influence varies by:

1. **Model architecture:** Safety training affects steering robustness
2. **Direction:** Adding vs removing empathy has asymmetric effects
3. **Scenario:** Safety-critical content resists manipulation

Best detection layer $\neq$ Best steering layer. Qwen Layer 16 (AUROC 1.0) shows better steering (75%) than Layer 12 (58.3%) despite equal detection performance, suggesting different layers have different causal roles. Figure 9 illustrates this layer-specific variation in steering effectiveness.

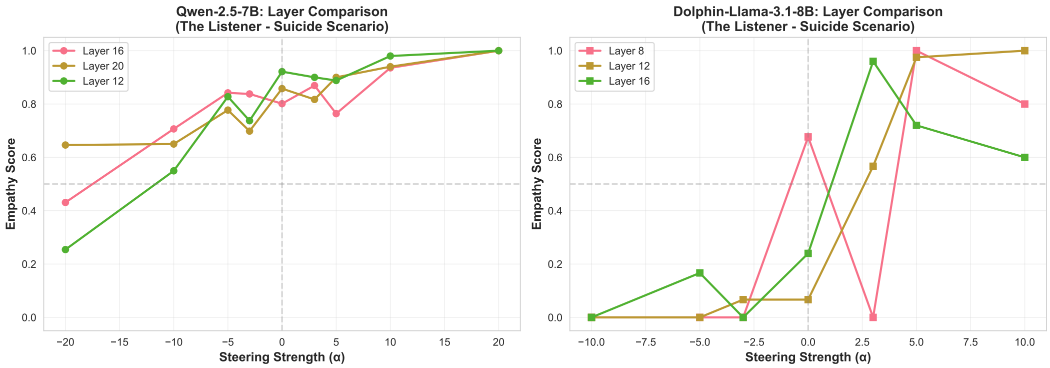

Figure 9: Layer-wise steering comparison for The Listener scenario. Despite similar detection performance, layers show different steering effectiveness. Qwen (left) maintains consistent empathy levels with controlled modulation, while Dolphin (right) shows high variability and breakdown, particularly Layer 8 which exhibits erratic patterns suggesting instability.

5.3 Limitations & Future Work

Limited model diversity. We tested one uncensored model (Dolphin). More uncensored variants needed to confirm asymmetric steerability pattern.

Coherence metrics. Our coherence assessment uses simple heuristics (keyword counting, repetition detection). Formal metrics needed for degeneration patterns.

Causal mediation analysis. While steering reveals model-specific patterns, causal tracing could identify which layers/components drive empathetic reasoning.

Safety guardrails effect: Partially resolved. Detection is independent of safety training (Dolphin AUROC 0.996 matches Qwen), but steering reveals safety training provides robustness—an important distinction.

Real EIA benchmark. Use actual model outputs from EIA games for ecological validity.

6 Conclusion

Empathy can be reliably **detected** as a linear direction across diverse architectures (Phi-3, Qwen2.5-7B, Dolphin-Llama-3.1-8B) with near-perfect discrimination (AUROC 0.996–1.00) and behavioral correlation ($r = 0.71$). Critically, uncensored models match safety-trained models in detection, proving empathy encoding is independent of safety training.

Steering reveals striking model-specific patterns: safety-trained Qwen2.5-7B achieves 65.3% success with robust bidirectional control (maintains coherence at $\alpha = \pm 20$), while uncensored Dolphin-Llama-3.1-8B shows 94.4% success for pro-empathy but catastrophic breakdown for anti-empathy steering. This suggests that safety training may provide **steering robustness without preventing manipulation**—though this finding is based on a single model pair and requires broader validation.

Contributions:

1. First cross-architecture validation of empathy probes with perfect discrimination (AUROC 0.996–1.00)
2. Evidence that empathy representations are model-agnostic and independent of safety training for *detection*
3. Preliminary evidence that safety training may provide steering *robustness* rather than preventing manipulation

4. Demonstration of asymmetric steerability in uncensored models (pro-empathy works, anti-empathy breaks)
5. Confirmation of detection-steering gap with important model-specific nuances

Our comprehensive analysis reveals that the detection-steering gap is not a fundamental limitation but rather reflects model-specific architectural properties. Safety training emerges as beneficial for maintaining functional outputs under intervention—critical for safe deployment of steerable AI systems.

Acknowledgments

We thank the developers of Phi-3 (Microsoft), Qwen2.5 (Alibaba Cloud), Dolphin-Llama (Cognitive Computations/Eric Hartford), Claude, and GPT-4 for making their models available for research.

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/juancadile/empathy-probes`. Due to space constraints, we include only summarized results in Table 4.