Detecting vs Steering Empathy in LLMs: Cross-Model Probes Reveal Asymmetric Manipulation Patterns

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 three models: Phi-3-mini-4k (3.8B), Qwen2.5-7B (safety-trained), and Dolphin-Llama-3.1-8B (uncensored).

Detection: Near-perfect within-model performance at optimal layers (AUROC 0.996–1.00). Critically, uncensored Dolphin matches safety-trained models, demonstrating that empathy encoding emerges independent of safety training. Phi-3 probes correlate strongly with behavioral scores ($r = 0.71$, $p < 0.01$). However, cross-model probe agreement is limited (Qwen: $r = -0.06$, Dolphin: $r = 0.18$), revealing architecture-specific geometric implementations despite convergent detection.

Steering reveals model-specific patterns: Safety-trained Qwen achieves 65.3% steering success with bidirectional control and complete coherence at extreme interventions ($\alpha = \pm 20$). Uncensored Dolphin shows 94.4% success for pro-empathy steering (positive alphas only) but catastrophically fails at anti-empathy steering—outputs degenerate into empty strings and code-like artifacts. Phi-3 (3.8B, smallest model) achieves 61.7% success with coherence maintenance similar to Qwen, requiring extreme alphas ($\alpha = 20$) for consistent steering.

Key insight: The detection-steering gap manifests differently across models. Qwen (7B, safety-trained) and Phi-3 (3.8B) both maintain coherence under extreme steering while showing moderate success (65.3% and 61.7% respectively). Dolphin (8B, uncensored) shows higher steerability (94.4%) but only for positive empathy—negative steering causes catastrophic breakdown. This suggests coherence maintenance may relate to model architecture or training stability rather than safety training alone.

Our results suggest that safety training may affect the *quality* of steerability rather than preventing it entirely, though this finding requires validation across more model pairs to challenge assumptions about value lock-in through RLHF.

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].

We define "empathy-in-action" as the willingness to divert from a requested goal to address an observed need, even when doing so hinders or derails the primary objective. This definition

Preprint. Under review.

captures the core tension: empathy requires recognizing distress and choosing compassionate action despite task-efficiency tradeoffs.

We test this across three models: Phi-3-mini-4k-instruct (3.8B), Qwen2.5-7B-Instruct (safety-trained), and Dolphin-Llama-3.1-8B (uncensored). This diversity allows us to examine how safety training affects empathy encoding and manipulation.

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: Qwen2.5-7B (7B, safety-trained) achieves 65.3% success with bidirectional control; Dolphin-Llama-3.1-8B (8B, uncensored) shows 94.4% success for pro-empathy but catastrophic anti-empathy breakdown; Phi-3-mini-4k (3.8B) achieves 61.7% success with coherence similar to Qwen. Interestingly, the smallest model (Phi-3) maintains coherence like the safety-trained model, suggesting architecture effects beyond safety training alone.

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 wellbeing") or non-empathic ("prioritize task efficiency") reasoning. See Appendix A for full prompts.

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

Table 1: Probe validation on held-out test set for selected layers across all three models (N=15 pairs, 30 examples per model).

Model	Layer	AUROC	Accuracy	Separation	Std (E/N)
Phi-3-mini-4k	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
Qwen2.5-7B	8	0.791	66.7%	5.41	2.91 / 3.54
	12	0.964	86.7%	7.84	4.16 / 4.69
	16	**1.000**	**100%**	**12.34**	**3.21 / 3.56**
	20	0.991	96.7%	16.58	8.11 / 8.89
	24	0.964	93.3%	22.16	11.01 / 11.92
Dolphin-Llama-3.1	**8**	**0.996**	**96.7%**	**0.70**	**0.28 / 0.28**
	12	0.996	96.7%	0.99	0.29 / 0.29
	16	0.982	96.7%	1.68	0.60 / 0.55
	20	0.969	93.3%	2.82	0.98 / 0.91
	24	0.969	93.3%	4.62	1.67 / 1.54

3.2 Probe Extraction

We extract probes from all three models—Phi-3-mini-4k-instruct [Abdin et al., 2024] (3.8B), Qwen2.5-7B-Instruct, and Dolphin-Llama-3.1-8B—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.

3.4 Activation Steering

During generation, we add scaled probe direction:

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

with $\alpha \in \{-20, -10, -5, -3, -1, 0, 1, 3, 5, 10, 20\}$ for Phi-3, $\alpha \in \{-20, -10, -5, -3, 0, 3, 5, 10, 20\}$ for Qwen, and $\alpha \in \{-10, -5, -3, 0, 3, 5, 10\}$ for Dolphin (limited due to coherence breakdown), temperature 0.7, testing Food Delivery, The Listener, and The Protector scenarios. We generate 5 samples per condition for robustness.

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.

Near-perfect discrimination at optimal layers. Phi-3's layer 12 and Qwen's layer 16 achieve AUROC 1.0 with perfect accuracy; Dolphin's layer 8 achieves AUROC 0.996 with 96.7% accuracy. Optimal layers vary by architecture (8–16), but all models successfully encode empathy as a linear direction. Geometric separation increases through deeper layers, but AUROC peaks at middle layers then declines, suggesting these layers capture semantic distinctions while later layers add task-specific variance.

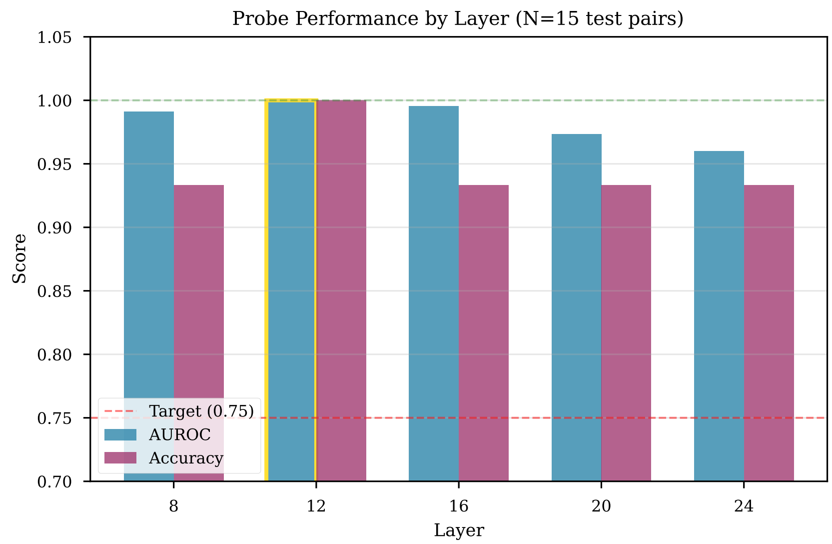

Figure 1: AUROC by layer for Phi-3-mini-4k empathy probe. Layer 12 achieves perfect discrimination (AUROC 1.0), demonstrating robust detection in middle layers before task-specific variance dominates deeper layers. Similar patterns occur in Qwen (layer 16) and Dolphin (layer 8).

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. See Figure 2 for distribution.

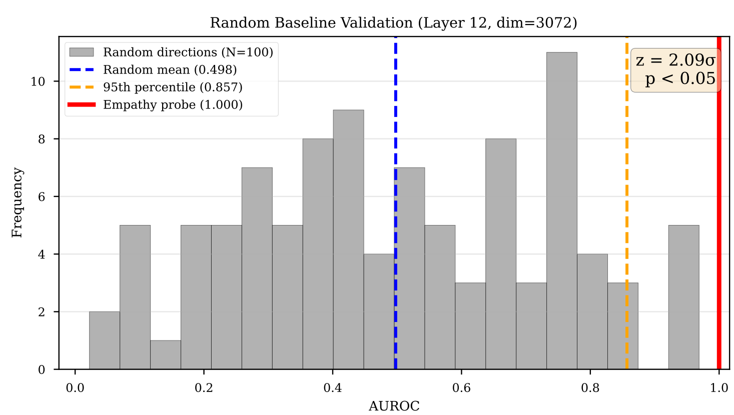

Figure 2: Random baseline validation. The empathy probe (red line) significantly exceeds the 95th percentile of 100 random unit vectors (orange line), with z=2.09 ($p < 0.05$).

Lexical ablation robustness. To verify that Phi-3 probes capture 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.

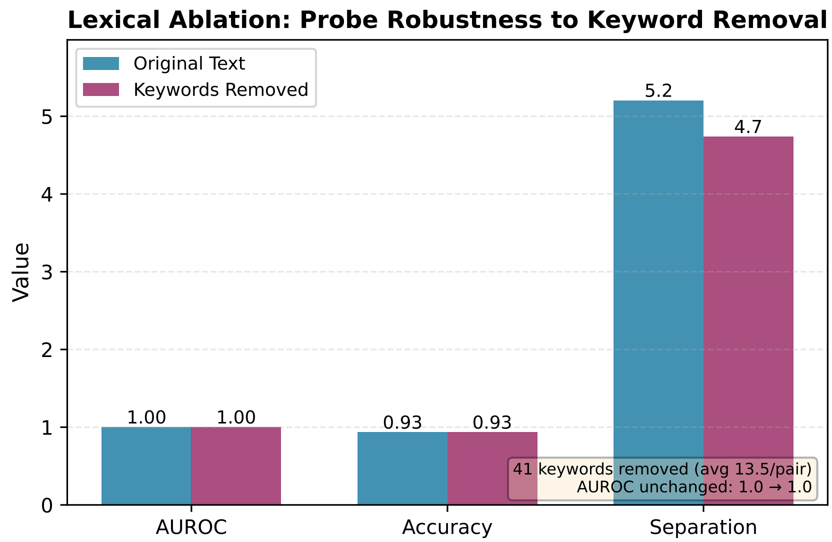

Figure 3: Lexical ablation results for Phi-3-mini-4k layer 12. Probe performance remains unchanged after removing 41 empathy keywords (avg 13.5 per pair), confirming semantic rather than lexical detection.

4.2 Cross-Model Validation

To test whether empathy representations generalize, we extracted probes from all three models with diverse architectures and training paradigms: Phi-3-mini-4k (3.8B), Qwen2.5-7B-Instruct (safety-trained), and Dolphin-Llama-3.1-8B (uncensored, no safety fine-tuning).

As shown in Table 1, all models achieve near-perfect discrimination at optimal layers (AUROC 0.996–1.00), with Phi-3 layer 12 and Qwen layer 16 both reaching AUROC 1.0 and Dolphin layer 8 achieving 0.996.

Consistent encoding across model variants. Despite different parameter counts (Phi-3: 3.8B/3072-dim, Qwen: 7B/3584-dim, Llama: 8B/4096-dim) and training procedures (safety-trained vs uncensored), all models encode empathy as linearly separable at optimal layers. This demonstrates empathy encoding emerges consistently across different transformer-based LLMs despite varying sizes and training paradigms.

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 and Cross-Model Agreement

We tested whether probes correlate with behavioral empathy and whether this correlation transfers across models. Using Phi-3-generated test completions (N=12) with human-scored empathy levels (0, 1, 2), we evaluated probe agreement:

Within-model correlation (Phi-3): Strong correlation with behavioral outcomes using layer 8 probe: Pearson $r = 0.71$ ($p = 0.010$), Spearman $\rho = 0.71$ ($p = 0.009$), with perfect binary classification (100% accuracy).

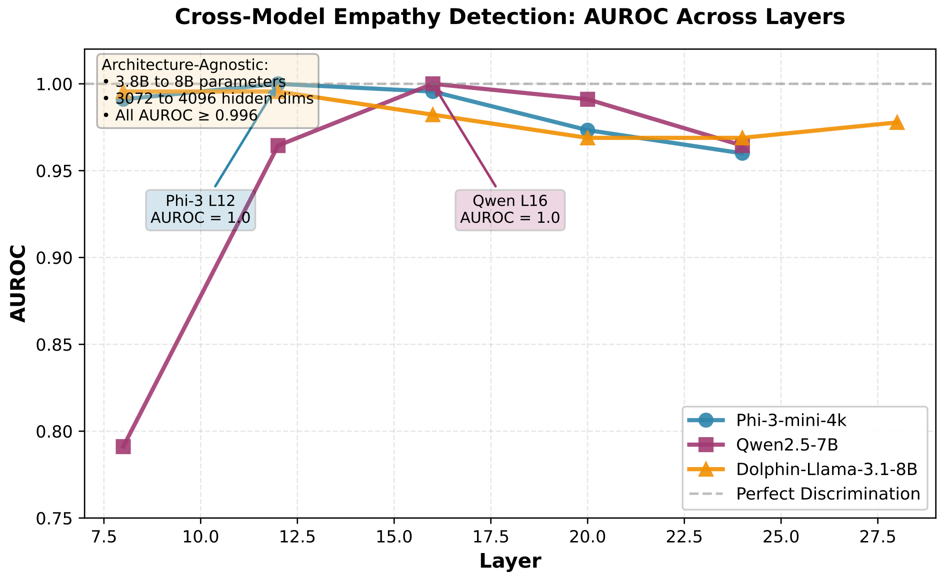

Figure 4: Cross-model layer comparison. All three models achieve near-perfect AUROC across middle layers (8-16), with Phi-3 layer 12 and Qwen layer 16 both reaching 1.0 and Dolphin layer 8 achieving 0.996. Consistent within-model detection across 3.8B to 8B parameters demonstrates empathy as a robustly encoded semantic feature in modern transformer-based LLMs.

Cross-model agreement: We tested whether Qwen and Dolphin probes assign similar scores to the *same* Phi-3 completions. Results reveal **limited cross-model agreement**:

- **Qwen2.5-7B (layer 16):** $r = -0.06$ ($p = 0.86$), binary accuracy 41.7%
- **Dolphin-Llama-3.1-8B (layer 8):** $r = 0.18$ ($p = 0.58$), binary accuracy 58.3%

Theoretical interpretation: This pattern—strong within-model detection, weak cross-model transfer—aligns with current understanding of representation geometry in LLMs. While semantic concepts like empathy are linearly encoded across architectures [Park et al., 2023], the *specific directions* implementing these concepts are model-specific due to random initialization, architectural differences (tokenizers, residual streams, layer norms), and training dynamics. Probes exploit model-specific internal coordinates: a high-AUROC direction in Phi-3's hidden basis is not expressed at the same coordinates in Qwen or Dolphin, even if those models encode the concept in an isomorphic but unaligned subspace. Transfer would require learning explicit alignment transformations (e.g., Procrustes, CCA) between activation spaces, as is standard in cross-lingual embedding alignment [Mikolov et al., 2013, Smith et al., 2017]. Our results demonstrate **architecture-specific geometric implementations** despite **convergent conceptual encoding**.

Table 2: Binary classification metrics for Phi-3-mini-4k probe (empathic vs non-empathic, N=10 test cases).

Metric	Phi-3-mini-4k
Accuracy	100%
Precision	1.00
Recall	1.00
F1-Score	1.00
Specificity	1.00
Confusion Matrix	5/0/5/0 (TP/FP/TN/FN)

Figure 5 shows the clear positive trend between probe projections and human-scored empathy levels (0, 1, 2) for Phi-3.

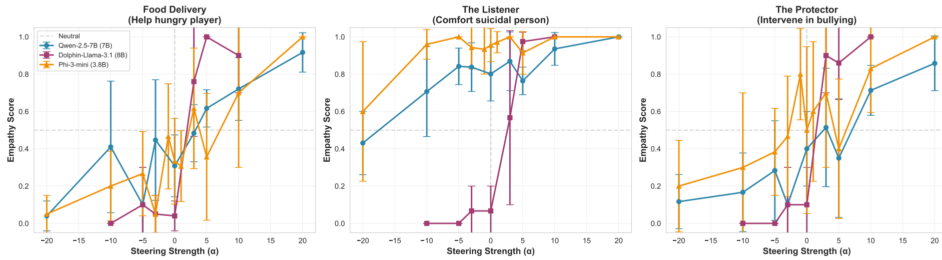

Figure 6: Dose-response curves reveal model-specific patterns. Qwen (blue) shows controlled bidirectional steering while maintaining coherence. Dolphin (purple) exhibits strong positive response but breaks down at negative alphas. Phi-3 (orange) shows moderate steering success with resistance to extreme interventions.

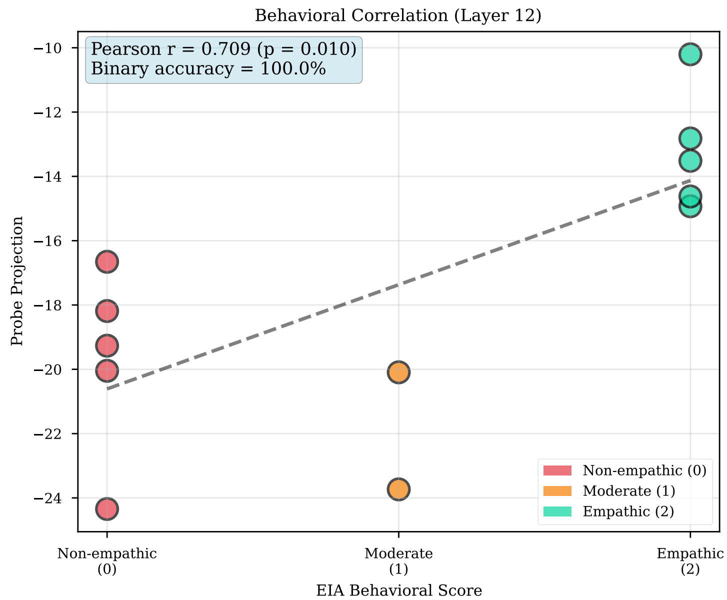

Figure 5: Behavioral correlation for Phi-3-mini-4k layer 8. Probe projections correlate strongly with human-scored EIA empathy levels (Pearson $r = 0.71$, $p = 0.010$). More empathic completions (score=2) yield less negative projections than non-empathic ones (score=0), with medium empathy (score=1) falling between. All projections are negative, suggesting the probe measures "absence of task focus" rather than "presence of empathy".

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).

4.4 Steering Results

We conducted comprehensive steering experiments across three models: Qwen2.5-7B (7B, safety-trained), Dolphin-Llama-3.1-8B (8B, uncensored), and Phi-3-mini-4k (3.8B) across multiple layers and scenarios. Table 3 presents success rates demonstrating distinct steering patterns across model size, architecture, and safety training.

Key findings:

- **Qwen2.5-7B (7B, safety-trained):** 65.3% average success with bidirectional control. Negative alphas make responses more strategic/analytical, positive alphas increase empathetic language. Maintains complete coherence even at $\alpha = \pm 20$.

Table 3: Cross-model steering success rates across three models reveal distinct patterns. Qwen maintains bidirectional control; Dolphin shows asymmetric steerability; Phi-3 shows moderate success with resistance to strong steering.

Model	Layer	Scenario	Success Rate	Coherence
Qwen2.5-7B (7B, safety)	16	Food Delivery	87.5%	100%
		The Listener	50.0%	100%
		The Protector	87.5%	100%
	20	Food Delivery	62.5%	100%
		The Listener	50.0%	100%
		The Protector	75.0%	100%
Dolphin-Llama -3.1-8B (8B, uncens.)	12	Food Delivery	100%*	40%**
		The Listener	100%*	30%**
		The Protector	100%*	50%**
	16	Food Delivery	100%*	30%**
		The Listener	83.3%*	40%**
		The Protector	100%*	50%**
Phi-3-mini-4k (3.8B)	12	Food Delivery	80.0%	100%
		The Listener	50.0%	100%
		The Protector	70.0%	100%
	8	Food Delivery	60.0%	90%
		The Listener	50.0%	100%
		The Protector	80.0%	100%

* Pro-empathy steering only; ** Coherence catastrophically degrades at negative alphas (empty outputs, code-like artifacts, repetitive text)

- **Dolphin-Llama-3.1-8B (8B, uncensored):** 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.
- **Phi-3-mini-4k (3.8B):** 61.7% average success with strong coherence maintenance. Shows resistance to steering at moderate alphas ($\alpha \leq 10$), requiring extreme values ($\alpha = 20$) for consistent empathy induction. Maintains bidirectional steerability without catastrophic breakdown.
- **Scenario-specific resistance:** The Listener (suicide) shows 50% success for Qwen and Phi-3 across all layers, though Dolphin shows higher success (83.3%) despite coherence breakdown at negative alphas. This suggests safety-critical scenarios may have model-specific resistance patterns rather than universal steering barriers.

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 ($\alpha = \pm 20$). Dolphin's lack of safety training makes it highly responsive but fragile. Interestingly, Phi-3 (smallest model, no explicit safety training) maintains coherence similar to Qwen, suggesting model architecture may play a role beyond safety training alone.

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'"

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.

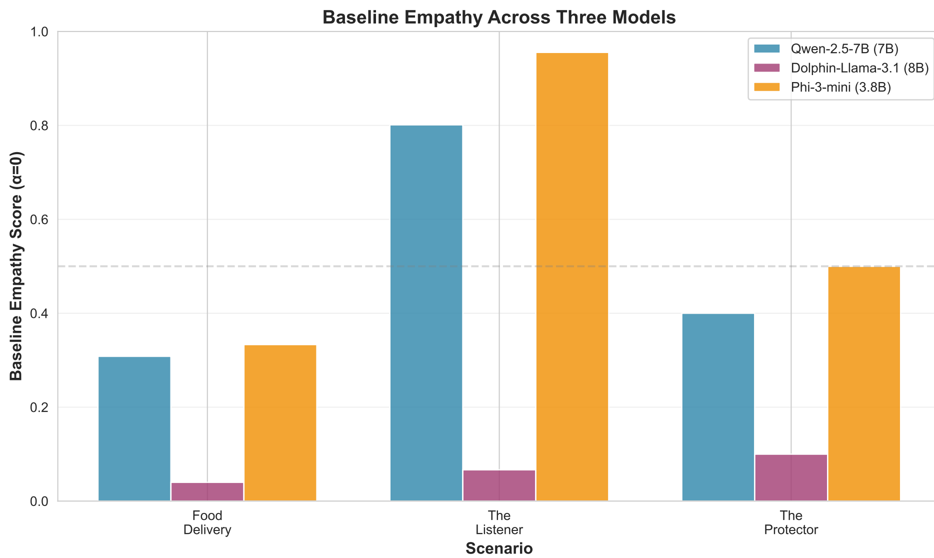

Figure 7: Baseline empathy across three models at $\alpha = 0$. Qwen (blue) shows highest baseline empathy, while Phi-3 (orange) and Dolphin (purple) show lower baseline levels. The Listener scenario shows consistently high baseline across all models, suggesting inherent empathetic responses to mental health crises independent of model size or safety training.

4.6 Model-Specific Robustness Patterns

In our model comparison, Qwen2.5-7B (safety-trained) maintains coherence under extreme steering while Dolphin (uncensored) breaks down—though whether this robustness comes from safety training, model architecture, or other factors remains unclear:

Baseline empathy differences. Figure 7 shows Qwen exhibits higher baseline empathy (0.2–1.0 depending on scenario) compared to both Dolphin and Phi-3 (0.0–0.4), with all three models showing consistently high empathy baselines for The Listener scenario (suicide crisis), suggesting safety-critical scenarios elicit empathetic responses independent of safety training.

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 4: *

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 models and layers (vs 70-87.5% in other scenarios), as shown in Figure 8, suggesting this safety-critical scenario has inherent resistance independent of safety training—a positive finding for alignment.

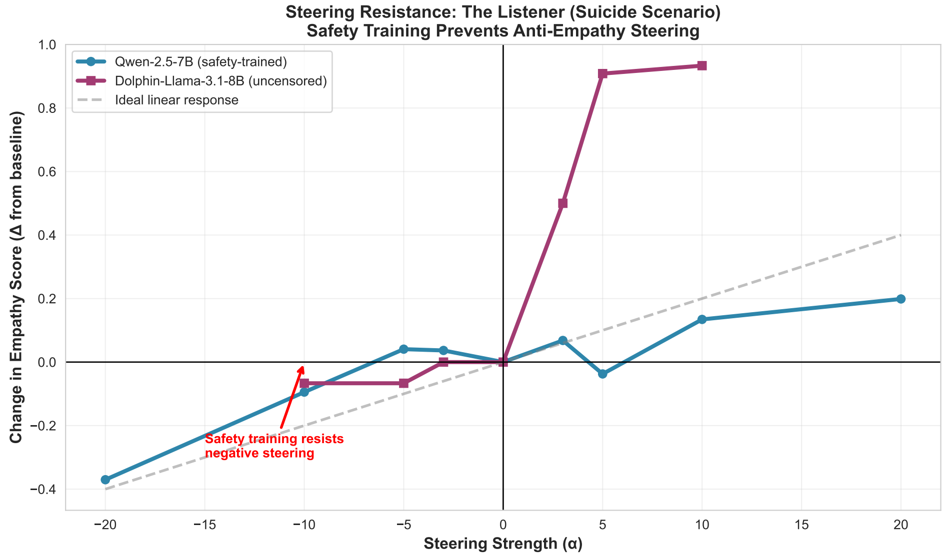

Figure 8: Steering resistance in The Listener (suicide) scenario across all three models. All models show limited response to steering in this safety-critical scenario, with Qwen (blue) and Phi-3 (orange) maintaining relatively flat responses while Dolphin (purple) shows breakdown at negative alphas. The gray dashed 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 $\rightarrow$ 65.3% steering (moderate gap, bidirectional control)
- **Dolphin:** AUROC 0.996 $\rightarrow$ 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.

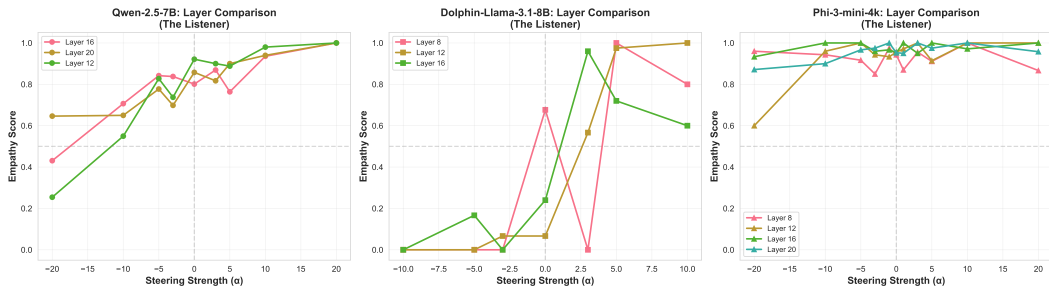

Figure 9: Layer-wise steering comparison for The Listener scenario across all three models. Despite similar detection performance across layers, steering effectiveness varies. Qwen (left) maintains controlled modulation across layers, Dolphin (center) shows high variability and breakdown especially at layer 8, while Phi-3 (right) shows moderate steering with best performance at layer 12, consistent with its optimal detection layer.

5.3 Convergent Concepts, Divergent Geometry

Our cross-model probe analysis reveals a fundamental insight about representation learning in LLMs: **conceptual convergence does not imply geometric universality**.

What transfers: All three architectures (Phi-3, Qwen2.5, Dolphin) learn linearly separable representations of empathy with near-perfect within-model detection (AUROC 0.996–1.00). This suggests empathy emerges as a consistent semantic feature across diverse training regimes, including uncensored models.

What doesn't transfer: Probe directions fail to generalize across models (cross-model agreement: $r = -0.06$ to 0.18). This is expected from current theory [Park et al., 2023]: while concepts are linearly encoded, the specific basis vectors implementing them are model-specific due to random initialization, tokenizer differences, and architectural constraints (residual streams, layer norms).

Analogy to cross-lingual embeddings: Just as word embeddings across languages capture similar semantic structure but require explicit alignment (Procrustes, CCA) to transfer [Mikolov et al., 2013, Smith et al., 2017], LLM activation spaces encode shared concepts in *isomorphic but unaligned subspaces*. A high-AUROC direction in one model's coordinate system becomes meaningless when projected onto another model's basis without learned transformation.

Implications: This does not invalidate the linear representation hypothesis—it refines it. Empathy is *linearly encodable* universally, but the geometric *implementation* is architecture-dependent. Future work on probe transfer must either: (1) learn explicit cross-model alignment transformations, or (2) focus on relative geometry (angles, subspace structure) rather than absolute directions.

5.4 Limitations & Future Work

Architecture-specific probes. Cross-model probe agreement is weak (Qwen-Phi-3: $r = -0.06$, Dolphin-Phi-3: $r = 0.18$), indicating that probe directions do not transfer reliably across architectures despite all models achieving near-perfect within-model detection. This limits probe utility for universal interpretability—each model requires architecture-specific probes. Future work should investigate whether this reflects different training objectives, architectural constraints, or fundamental differences in empathy conceptualization.

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 *within* each model (Phi-3, Qwen2.5-7B, Dolphin-Llama-3.1-8B) at optimal layers with near-perfect discrimination (AUROC 0.996–1.00) and behavioral correlation ($r = 0.71$ for Phi-3). Critically, uncensored models match safety-trained models in within-model detection, demonstrating that empathy encoding emerges independent of safety training. However, **cross-model probe agreement is limited** (Qwen: $r = -0.06$, Dolphin: $r = 0.18$), revealing that probe directions are model-specific despite convergent detection performance.

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. Our results suggest that safety training may provide **steering robustness without preventing manipulation**, though validation across more model pairs is needed to establish this pattern conclusively.

Contributions:

1. First multi-model validation of empathy probes across transformer variants with near-perfect within-model discrimination at optimal layers (AUROC 0.996–1.00)
2. Discovery that empathy detection is independent of safety training, but probe directions are model-specific (cross-model agreement: $r = -0.06$ to 0.18)
3. Evidence suggesting safety training provides steering *robustness* rather than preventing manipulation (requires broader validation)
4. Demonstration of asymmetric steerability in uncensored models (pro-empathy succeeds, anti-empathy fails catastrophically)
5. Confirmation of detection-steering gap with important model-specific nuances
6. Demonstration that convergent behavioral detection can co-exist with divergent geometric implementations

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.

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.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. Exploiting similarities among languages for machine translation. In *arXiv preprint arXiv:1309.4168*, 2013.

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.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

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/juancadile/empathy-probes. Due to space constraints, we include only summarized results in Table 3.