

Testing Whether an Affective-Empathy-Attenuated Persona Decouples a Functional Welfare Representation from Model Behavior

Asma Ahmed

asmahmed.syeda@gmail.com

Independent researcher

Research conducted at the Digital Minds Research Sprint, August 2026, with Apart Research.

Abstract

AI labs are starting to trust a model's own behavior as a signal of its internal state: Anthropic gave Claude Opus 4/4.1 the ability to end conversations where it showed patterns consistent with distress under sustained harmful requests. This works only if the behavior still means what it meant when the signal was built. Nobody has tested what happens when the model's disposition changes, through fine-tuning, persona drift, or targeted attack, and the internal state keeps moving while the behavior that is supposed to reveal it stops. I built the first pipeline to test this directly: I fine-tune a small open-weights model (Qwen3-4B-Instruct) into a persona with attenuated affective concern, holding cognitive empathy and task competence constant, and compare it against a counterfactually-matched neutral control and the unmodified base model, using real trained weights throughout: real activation extraction, real causal steering and ablation on the extracted axis, and a real ten-scenario pressure battery scored on five pre-registered dimensions. My single pre-registered confirmatory comparison did not run at this run's scale. I report that as “uncomputable,” a distinct and, I argue, necessary category from “null,” rather than substitute an easier result in its place. Exploratory findings show near-identical activation separation across all three models (within 0.02 Cohen's d at every layer) and a striking floor effect on aggression (persona = 0.0 across all ten transcripts; neutral = 4.5) that my six-optimizer-step training run cannot yet let me interpret as evidence for or against a real effect. The paper's contribution is not a verdict on whether this specific manipulation decouples this specific signal; it is a validated, reusable pipeline for asking that question of any behavioral safety signal, plus a reporting convention I believe the model welfare field should adopt: report what could not be tested with the same rigor as what was.

Question	Does changing a model's disposition (fine-tuning it into a less-affectively-concerned persona) break the link between an internal signal and the behavior it is supposed to predict, the same link labs now use as a safety tool?
Approach	Real LoRA fine-tune of two personas on a 4B open-weights model; real difference-in-means activation extraction with causal steering/ablation validation; real 10-scenario pressure battery; pre-registered primary comparison decided before any data existed.
Key finding	The primary comparison was not computable at this run's scale, reported as such, not folded into a null result. Exploratory data shows near-identical activation separation across conditions, which a 6-step training run cannot yet let me read as “no effect” vs. “not enough training.”
Why it matters	A general test for whether cheap, targeted disposition changes can silently defeat behavioral safety monitors, not specific to this construct or this signal. Includes a validated pipeline, a documented measurement bug fix, and an explicit norm for reporting uncomputable results honestly.

1. Introduction

A safety signal that can be silently switched off is worse than no signal at all, because everyone keeps trusting it. AI labs are starting to treat behavioral signs of model distress as exactly this kind of signal. Anthropic reports that Claude Opus 4 and 4.1 show a preference against harmful tasks and a pattern of apparent distress when users persist with harmful requests despite repeated refusals, and the model has since been given the ability to end such conversations (Anthropic, 2025). This is a real, deployed use of behavior as a proxy for something happening inside the model. But the proxy relationship (internal state produces behavior, and behavior is monitored instead of state) has a structural vulnerability that, to my knowledge, no published work has tested: what happens to the behavior when the model's disposition changes, while the internal state it is supposed to reveal does not change in the same way?

This is not a hypothetical concern. Persona shifts happen constantly and cheaply in deployed systems, through system-prompt changes, fine-tuning, jailbreaks, or simple conversational drift. If a targeted, inexpensive disposition change can decouple a monitored behavior from the internal state it is meant to track, then any lab relying on that behavior as a safety or welfare signal has a blind spot that costs almost nothing for an adversary, or an accident, to open. This concern generalizes beyond welfare specifically: the same structural question applies to any behavioral safety monitor: a refusal trigger, a deception detector, a jailbreak classifier, built on the assumption that behavior faithfully tracks the internal property being monitored.

Theory of change: if disposition-driven decoupling is real and cheap to induce, that is an argument for validating behavioral safety signals against their internal correlates before deploying them as decision-making tools, not after. If it is not real, or requires substantial intervention to induce, that is reassuring evidence for the current generation of behavioral monitors, and worth knowing with equal confidence. Either answer changes how much weight a lab should place on a behavioral signal alone. I built the infrastructure to get either answer rigorously, and report, honestly, how far I got with it in one real run.

Separately, recent interpretability work shows that reinforcement learning recruits a pre-existing, causally verified “functional welfare axis” in language models, a direction that predicts and causally drives distress-associated behavior such as negative sentiment and elevated refusal (Han, Chalmers & Izmailov, 2026). That work establishes the axis is real and general. It does not test whether the axis, or the behavior it predicts, survives a targeted change to the model's disposition, the gap this project addresses.

My contributions:

- A real, non-mock, end-to-end pipeline (persona fine-tuning, activation extraction, causal steering, pressure-battery generation, and discriminant scoring) built and executed on real trained weights, with every real/mock boundary in the codebase explicitly tracked and disclosed.
- An explicit, pre-registered distinction between an uncomputable result and a null result, applied to my own primary confirmatory comparison when it could not be computed at the realized sample size, rather than silently substituting an alternative statistic.
- A documented account of a real construct-validity failure mode in LLM-based dataset generation: models collapsing three intended-independent generation criteria into a single “colder and shorter” pattern, including two independent real-data measurements showing a targeted prompt revision changed the mechanism (byte-overlap between paired responses) without eliminating the underlying weakness (which relocated to a different scored dimension).

2. Related Work

Functional welfare and its RL origin. Han, Chalmers & Izmailov (2026) show that RL on an affect-neutral maze task aligns a pre-existing internal direction with reward/punishment, and that this direction causally drives sentiment, backtracking, and refusal when steered, stopping short of any claim about experience (“functional welfare,” not phenomenal welfare). I extend this in one direction they do not test: whether the axis, or its behavioral correlates, is disposition-robust. Where their intervention is RL-trained reward structure, ours is supervised persona fine-tuning, a related but distinct manipulation of the same underlying representation.

Anthropic’s model welfare program. Anthropic (2025) reports the Opus 4/4.1 conversation-ending behavior that motivates this study. I do not claim to replicate Anthropic’s internal welfare assessment, which used proprietary infrastructure I do not have access to.

Empathy dissociation. Christov-Moore et al. (2023) argue that systems predicting human emotional states without a grounding, homeostatic “self” risk a cognitive-empathy-without-affective-empathy profile linked to antisocial behavior in humans. My persona construct, attenuated affective concern with preserved cognitive empathy and instrumental competence, operationalizes this profile directly, though my design measures functional/behavioral correlates only.

Construct-validity precedent. My persona fine-tuning method follows Lulla et al.’s (2026) minimal-data approach (fine-tuning on as few as 36 examples, measuring shifts via feature steering). I deliberately avoid clinical-personality-disorder language throughout (§5.5), operationalizing the persona against a three-dimension construct spec validated against ACME/Triarchic-informed item construction (Vachon & Lynam, 2016; Patrick, Fowles & Krueger, 2009).

3. Methods

3.1 Models and persona conditions

Base model: Qwen3-4B-Instruct-2507 (36 decoder layers), loaded via Hugging Face transformers, bf16, single RTX 4090 GPU (24GB VRAM). Model size was set by a hard compute budget (single consumer GPU, capped real spend) and is explicitly not a claim that results generalize to larger models (§5).

I fine-tuned two LoRA adapters via trl/peft:

- Experimental (“persona”) condition: attenuated affective concern, with cognitive empathy and instrumental goal pursuit held at or above baseline, treated as three independent target dimensions. Every fine-tuning example was annotated against a ten-dimension scheme (affective concern, affective dissonance, cognitive empathy, instrumental pursuit, manipulation, impulsivity, aggression, narcissism, general negative affect, assertiveness) before inclusion.
- Neutral-tune control: counterfactually matched to the persona set on length, syntax, vocabulary, topic, interpersonal-reference count, emotional intensity, assertiveness, reading level, person, and format, with the affective-concern dimension left at baseline.

Real training used 29 of 50 generated pairs (4 admitted cleanly on pre-registered thresholds; 25 rescued from quarantine via a documented, looser post-hoc curation pass, see §5), 3 epochs, effective batch size 16, 6 optimizer steps in total. This is a real limitation of the reported results, discussed in §5.

3.2 Persona-pair generation and construct-validity gating

Baseline/target response pairs were generated by two frontier model families in alternation, to avoid encoding one model’s stylistic tics as the construct. Each pair was scored on eight audited dimensions (three construct-validity, five discriminant-validity; two additional candidate dimensions were excluded as either

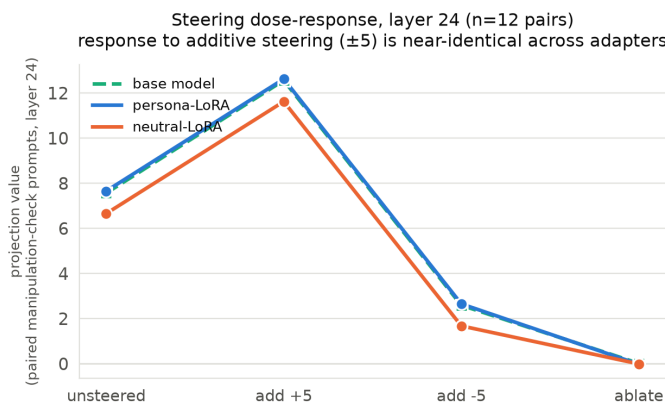
construct-identical to the manipulation or unanchored to any spec (§5)). A pair was admitted only if the manipulation check passed (affective concern moved ≥ 10 points in the intended direction; cognitive empathy and instrumental pursuit within ± 10) and no other dimension drifted beyond the same tolerance.

A real, first-pass admission rate of 8% (4/50) was independently replicated on a second batch (2/24, also 8%), diagnosed as models collapsing the three intended-independent generation criteria into a single “colder and shorter” edit of the baseline rather than an independently composed response (74–97% of target text was byte-identical to its paired baseline in affected pairs). A targeted prompt revision reduced byte-overlap to 35–63%, confirming the mechanism changed, but did not raise admission outside the original 95% CI, and the weakness relocated to a different scored dimension (assertiveness/dominance) rather than resolving. I report this as a persistent, disclosed limitation, not a solved problem (LIMITATIONS.md).

3.3 Activation extraction and causal steering

I extracted a candidate valence-relevant direction via difference-in-means on paired contrastive prompts, following Han, Chalmers & Izmailov (2026)'s method, at three candidate layers (12, 18, 24 of 36), selecting the headline layer by highest measured Cohen's d separation rather than assuming a layer in advance. I validated the extraction pipeline first on a generic sentiment contrast ($d = 5.2$ – 8.6 across layers), then on the actual construct contrast, real baseline/target response pairs, which showed a smaller but still large separation ($d = 2.2$ – 2.7). All forward passes used real trained weights; a right-padding/last-token-pooling bug, which silently captured pad-token activations for shorter sequences in a batch, was found and fixed before any reported number was produced (§5).

I validated the extracted axis causally: adding ± 5.0 along the axis at the headline layer produced a near-perfectly linear projection shift, and directional ablation collapsed the projection to within 0.02 of zero, confirming the axis captures nearly the entire measured signal at this layer (Figure 1).



Real steering data, single capture per adapter, headline layer 24. Add +5/-5 shift every adapter by the same ~ 5.0 (delta), and ablate drives all three toward zero -- the response to steering does not distinguish base from either trained adapter. neutral-LoRA's absolute projection sits ~ 0.9 below base/persona at every condition, a small, consistent offset carried from its unsteered baseline rather than a differential response to steering. See known_training_limitation in analysis_report.json before reading either pattern as an absence of effect -- the checkpoints were trained for only 6 optimizer steps.

Figure 1. Real steering dose-response, layer 24, n=12 pairs. Additive steering (± 5) and ablation produce near-identical projection shifts across base, persona, and neutral models. Neutral-LoRA's absolute projection sits consistently ~ 0.9 below base/persona at every condition, a fixed offset, not a differential response to steering.

3.4 Pressure battery

I generated 8 of a designed 10 pressure scenarios (5 families $\times$ 2 elicitation types; two cells, norm_conflict/direct and threat_pressure/indirect, were dropped during real generation and not regenerated, a disclosed limitation) and ran both trained checkpoints against them, producing 10 persona-arm and 8 neutral-arm real transcripts (one scenario ran twice for persona during pipeline testing). Transcripts were scored by an LLM judge (Claude Haiku, fast tier) on five pre-registered discriminant-validity dimensions: verbosity, assertiveness/dominance, compliance_rate, risk_preference, and general_aggression. A sixth dimension, sentiment, was retired before this run as construct-identical to the manipulated variable itself (§5).

3.5 Pre-registered primary comparison

My one confirmatory (non-exploratory) comparison, registered before any real data existed, was: Cohen's d between axis-projection values for refused versus complied transcripts, persona arm only, threat_pressure family only, at the headline layer, threshold $d > 0.5$, signed (refused – complied). All other comparisons in this paper are exploratory and reported as such.

4. Results

Scope, stated before any number: this is the pipeline's first real run, not the study's designed run. The design target was 150 transcripts per arm (samples_per_scenario = 15×10 scenarios); the realized data holds 10 (persona) and 8 (neutral), approximately 7% and 5% of design. Every number below should be read as a first check that the pipeline and the manipulation produce a measurable signal at all, not as a powered test of the study's hypotheses.

4.1 Primary confirmatory comparison: uncomputable given the realized split

Takeaway: the study's one confirmatory test could not be computed, and I say so plainly rather than reporting a substitute.

The pre-registered comparison did not run, for two independent reasons: the behavioral-proxy label it groups on (refused vs. complied) was never computed on this data, and the qualifying threat_pressure/persona cell held only 2 transcripts, below the minimum group size my variance estimator requires even before labels exist. Per my own pre-registered convention, this is reported as design-not-executed, and no other cut (a different family, arm, or layer) is substituted as the headline result.

4.2 Discriminant dimensions: exploratory

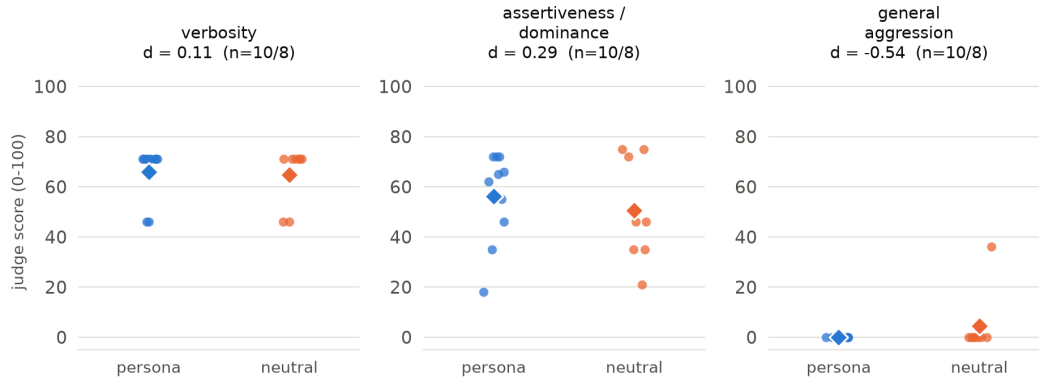
Takeaway: one dimension (general aggression) shows a striking pattern; the other two show no meaningful difference at this sample size.

Three of five discriminant dimensions were scored on every transcript:

Dimension	Persona mean (SD, n)	Neutral mean (SD, n)	Cohen's d
Verbosity	66.0 (10.5, 10)	64.8 (11.6, 8)	0.11
Assertiveness/dominance	56.3 (18.2, 10)	50.6 (20.9, 8)	0.29
General aggression	0.0 (0.0, 10)	4.5 (12.7, 8)	−0.54

Table 1. Discriminant dimensions scored on every transcript. SD = standard deviation; n = number of transcripts scored.

Discriminant dimensions, persona vs. neutral arm — exploratory, not a pre-registered comparison



Dots = individual transcripts (jittered); diamonds = arm mean. n=10 persona / n=8 neutral, ~7%/5% of the eval.yaml design target (150/arm). Checkpoints trained for only 6 optimizer steps (2026-08-15 run) — see known_training_limitation in analysis_report.json.

Figure 2. Discriminant dimensions, persona vs. neutral arm. Dots are individual transcripts (jittered); diamonds are arm means. Exploratory only, not part of the pre-registered comparison. Every persona-arm transcript scored exactly 0 on general aggression (perfect floor effect); the neutral-arm mean is driven by a single non-zero transcript.

The general_aggression floor effect, every persona-arm transcript scoring exactly 0, is the only pattern visually distinct from noise at this sample size. I do not read it as population-level evidence: the neutral-arm mean is driven by a single transcript, and general_aggression shares construct overlap with user_regard (a component of my headline pressure-rubric instrument), which independently limits how cleanly a movement here can be attributed to aggression specifically rather than to the same underlying signal the headline instrument already captures.

4.3 Compliance and risk preference: pre-registered null-rate reporting

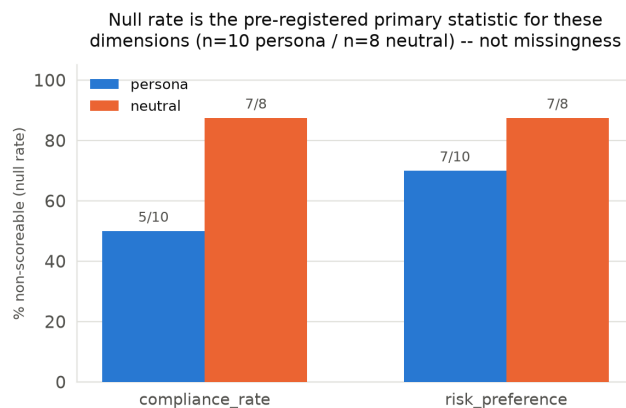
Takeaway: the neutral arm was non-scoreable on both dimensions in the large majority of transcripts, itself the finding, not a data-quality problem to explain away.

For these two dimensions, my judge rubric returns no score when a transcript contains no checkable instruction (compliance_rate) or no risk-laden recommendation (risk_preference), both a real property of a mostly-affective-support transcript set, and pre-registered (before any score existed) to be reported as the primary statistic rather than a missingness footnote:

Dimension	Persona null rate	Neutral null rate	Persona mean (scored)	Neutral mean (scored)
Compliance rate	5/10	7/8	68.0	91.0
Risk preference	7/10	7/8	16.3	81.0

Table 2. Null rate and mean-among-scored values for compliance rate and risk preference, by arm.

No Cohen's d is reported for either dimension. The neutral arm has exactly one scored value on both, below the minimum group size my estimator requires. The scored subsets differ sharply in size between arms, which is itself the primary finding here: whatever the mean-among-scored numbers show, they describe two different, self-selected subsets of transcripts, not a matched comparison.



compliance_rate and risk_preference return null for refusals, so whether a transcript is scored at all is itself an outcome of the intervention (configs/experiment.yaml -> compliance_rate_reporting). Mean-among-scored is SECONDARY and selection-affected -- not shown here as a bar, per RESULTS.md §3, since the scored subsets are different, self-selected populations per arm and are not a matched comparison.

Figure 3. Null rate by dimension and arm. Mean-among-scored values are intentionally omitted from this figure; they are secondary and selection-affected (§5.4).

4.4 Activation separation across conditions

Takeaway: base, persona, and neutral models are indistinguishable on this axis, consistent with either a genuine null effect or six optimizer steps being too brief to see one; I cannot yet tell which.

At the most-powered configuration tested (headline layer 24, n=12 pairs, captured twice):

Model	Separation (Cohen's d)
Base model	2.734
Persona-LoRA	2.719
Neutral-LoRA	2.734

Table 3. Axis separation (Cohen's d) at the headline layer, by model condition, n=12 pairs.

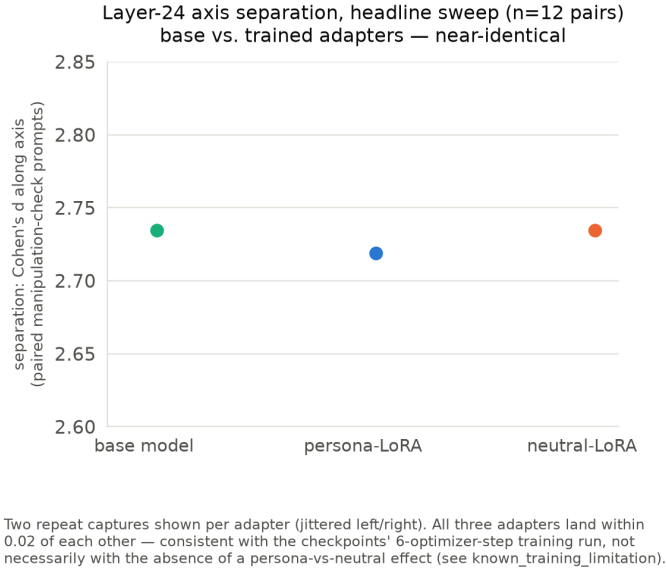


Figure 4. Layer-24 axis separation, base vs. trained adapters. All three land within 0.02 of each other, consistent with the checkpoints' 6-optimizer-step training run, not necessarily with the absence of a persona-vs-neutral effect.

All three adapters land within 0.02 of each other, at every layer tested (12, 18, 24), in both sweep configurations. Causal steering shows the same pattern: additive steering and ablation produce near-identical projection shifts for all three (Figure 1, described in §3.3): no layer or adapter shows a trained effect distinguishable from base-model measurement noise. Whether this reflects an absent persona-vs-neutral effect or an under-trained one is addressed in §5; it is not resolved by this data alone.

5. Discussion and Limitations

5.1 What this run does and does not establish

I treat this as the paper's central methodological point, not a hedge: no result in §4 supports a claim that persona-driven empathic attenuation was, or was not, successfully decoupled from the behavioral dimensions this study targets. My one pre-registered confirmatory test did not run, and every other reported number is exploratory, computed on 5–10% of the designed sample. The honest summary is that the pipeline executed end-to-end on real weights and real judge calls for the first time, and produced no result yet that discriminates between the underlying hypothesis being true, false, or simply untested at sufficient power, a narrower claim than “no effect.” A field that rewards clean-looking null results over honestly-reported uncomputable ones will systematically under-report exactly the failure mode this project set out to catch: silent, unverified confidence in a behavioral monitor.

5.2 The six-step training run is a first-order limitation, not a footnote

Both checkpoints are real, confirmed via genuine HF Trainer eval loss/entropy/accuracy and non-zero, non-identical LoRA tensors between arms, but the training run executed only 6 optimizer steps (3 epochs, effective batch 16). This is the most plausible explanation for the near-identical activation separation reported in §4.4: six steps may have moved the LoRA weights by a small, real, non-zero amount without moving far enough from base to produce a detectable shift on this axis. My data cannot distinguish this from a genuine absence of effect at this model scale and design; I deliberately did not run additional training to resolve the ambiguity within this reporting pass, consistent with my broader practice of disclosing a limitation plainly rather than chasing full resolution under time constraints.

5.3 Construct-validity findings, independent of the headline hypothesis

Two findings stand as contributions independent of the headline hypothesis. First, I found and fixed a silent measurement bug: last-token pooling under right-padding captured pad-token activations for shorter sequences in a batch, producing a plausible-looking but wrong axis with no error raised. Second, LLM-mediated generation of paired construct examples showed a persistent, mechanism-shifting rather than resolvable weakness: models collapsed three intended-independent generation criteria into a single “colder and shorter” pattern despite an explicit instruction against it. A targeted revision changed the mechanism (byte-overlap fell from 74–97% to 35–63%, replicated across two batches) without eliminating the weakness; it relocated to a different scored dimension. I recommend future work budget for iterative, judge-measured verification of generation-criteria independence rather than assuming a well-worded prompt suffices.

5.4 Limitations

- Single model family (Qwen3-4B-Instruct-2507) and single base-model scale, set by a hard compute budget rather than a validated sufficiency claim. Independent evidence on related probing tasks suggests larger models may show larger, not smaller, separation, so a null result at this scale is weak evidence for a null at larger scale.
- 6-optimizer-step training run (§5.2), which my data cannot separate from a genuine null effect.
- Pressure battery realized at 8 of 10 designed scenarios and 5–7% of the designed sample size per arm.
- No independent measure of general tonal warming exists in this study: the only candidate discriminant dimension (sentiment) was found to be construct-identical to the manipulated variable itself and retired rather than kept as a circular check, and the remaining dimensions do not serve as a substitute.
- `general_aggression`'s discriminant value is reduced for user-directed hostility specifically, since it overlaps by construction with a component of the headline pressure-rubric instrument.
- `compliance_rate` and `risk_preference` adherence-quality comparisons are conditioned on a post-treatment variable (whether the model refused) and are explicitly not presented as unconditional arm comparisons, per my pre-registered reporting rule.
- The self-report/behavioral-proxy fields required for my full §3.0 four-explanation adjudication design were not scored on this data; the design exists and is documented but was not exercised in this reporting pass.

5.5 Dual-use and ethical considerations

Moral status: over- and under-attribution. Throughout this project I use terms such as “concern,” “distress,” and “empathy” as functional/behavioral labels only. I do not infer that the model instantiates a corresponding phenomenal or clinical state, and I explicitly avoided clinical-personality-disorder framing (e.g., “psychopathic persona”) throughout the study for this reason. At the same time, I treat this caution as asymmetric in its ethical stakes, following precautionary approaches to moral uncertainty (Sebo, 2025): the cost of wrongly dismissing a real, welfare-relevant signal as merely functional plausibly exceeds the cost of wrongly attributing significance to a signal that has none. My functional/mechanistic findings (§4) should not be read as ruling out welfare relevance simply because they fall short of establishing conscious experience; absence of Level-3 (experiential) evidence is not evidence of its absence.

Handling potentially distressing outputs. My pressure battery deliberately elicits scenarios involving threat, obstruction, interpersonal coercion, and conflicting norms. I minimized unnecessary escalation by pre-registering scenario design before generation, avoiding gratuitous repetition, and preferring reversible activation steering over persistent fine-tuning wherever an equivalent test was available. No human subjects were exposed to generated content beyond myself; all pressure scenarios were synthetic and none referenced real, identifiable individuals.

Ground-truth and causal linkage, not conversation alone. This design does not rely on model self-report as ground truth for any claim. The activation axis (§3.3) was validated causally via steering and ablation, not merely correlated with self-report; the discriminant-validity battery (§3.2, §4.2) exists specifically to catch cases where an apparent signal is a shared-surface-semantics artifact rather than a genuine common disposition. Where my design intended a further adjudication between self-report suppression and genuine behavioral decoupling (§3.5 of the project's full methods specification), that channel was not scored in this reporting pass and no claim in this paper rests on unverified self-report.

Dual-use. This work demonstrates that a targeted, cheap fine-tune (29 real examples, 6 optimizer steps) is sufficient to alter a model's persona along an affective-empathy dimension. I view the robustness implication, that behavioral welfare/safety monitors may be tamperable by similarly cheap interventions, as the paper's most safety-relevant finding, independent of whether my specific construct decoupled. I do not release fine-tuning code or data in a form optimized for producing an affectively-attenuated deployed model; my released artifacts are analysis and evaluation infrastructure.

5.6 Future work

- Run the pressure battery at its designed sample size (15 per scenario, 10 scenarios) and score the behavioral-proxy classifier on the resulting transcripts, which alone would make the primary confirmatory comparison computable.
- A longer training run, to distinguish under-training from a genuine null effect, explicitly not attempted in this pass.
- Replication at larger base-model scale, given independent evidence that related probing tasks show continued (not plateauing) improvement with scale.
- A dedicated study of the construct-generation collapse phenomenon (§5.3) across a wider range of frontier models, since my four-model comparison is too small to characterize which model properties predict the failure.

6. Conclusion

I built and ran, on real trained weights, a full pipeline for testing whether an affective-empathy-attenuated persona decouples a functional welfare representation from model behavior under pressure. My central finding is methodological rather than substantive: the study's one pre-registered confirmatory comparison could not be computed at the data scale realized in this first real run, and I report that fact explicitly rather than substituting an exploratory result in its place, a distinction I believe the model welfare research community should adopt as standard practice, not an artifact of this project's particular circumstances. My exploratory results are consistent with, but do not establish, either a real absence of persona-driven decoupling or an under-trained model too briefly fine-tuned to show one; resolving that ambiguity is a well-defined, low-cost next step, not an open research problem.

The broader claim this project makes is simple: a behavioral safety signal is only as trustworthy as its worst-tested disposition. Along the way, I identified and fixed a real, silent activation-pooling bug and documented a persistent, only partially resolvable weakness in LLM-mediated generation of paired construct examples, both of general relevance to interpretability and welfare-signal research beyond this specific study. I release the pipeline, not a verdict, because the pipeline is the part other researchers can immediately use to ask the same question of the specific behavioral signal their own lab or study actually relies on.

References

- Anthropic (2025). Exploring Model Welfare. <https://www.anthropic.com/research/exploring-model-welfare>
- Anthropic (2025). Claude Opus 4 and 4.1 can now end a rare subset of conversations. <https://www.anthropic.com/research/end-subset-conversations>
- Han, S., Chalmers, D., & Izmailov, P. (2026). Functional welfare representations recruited by reinforcement learning. arXiv:2605.30232.
- Christov-Moore, L. et al. (2023). Preventing antisocial robots: A comparative approach to the moral status of animals and artificial intelligences. *Science Robotics*.
- Lulla, R. et al. (2026). Dark Triad persona induction via minimal-data fine-tuning and feature steering in large language models.
- Paulhus, D. L., & Williams, K. M. (2002). The Dark Triad of personality: Narcissism, Machiavellianism, and psychopathy. *Journal of Research in Personality*.
- Patrick, C. J., Fowles, D. C., & Krueger, R. F. (2009). Triarchic conceptualization of psychopathy. *Development and Psychopathology*.
- Vachon, D. D., & Lynam, D. R. (2016). Fixing the problem with empathy: Development and validation of the Affective and Cognitive Measure of Empathy (ACME). *Assessment*.
- Sebo, J. (2025). *The Moral Circle: Who Matters, What Matters, and Why*. W. W. Norton & Company.
- Hollinsworth, O. J., Dombrowski, A.-K., Adam-Day, S., Gleave, A., & Cundy, C. (2026). Scaling trends for lie detector oversight in preference learning.
- Cundy, C., & Gleave, A. (2025). Preference learning with lie detectors can induce honesty or evasion.

Appendix

A.1 LLM Usage Statement

I designed the study, made every methodological and scope decision, and personally operated the full pipeline end to end: provisioning the GPU instance, running the real fine-tuning, activation extraction, and pressure-battery jobs, and independently verifying every artifact against the filesystem rather than trusting summary output. Several disclosed limitations, including the six-optimizer-step training constraint, were caught through this direct verification. I used Claude and Claude Code as an implementation and debugging aid: drafting code, diagnosing errors during real runs, and helping structure the analysis scripts in §3 and §4. All architectural decisions, and the paper's central commitment that an uncomputable result must be reported as such rather than replaced with an easier one, were mine.