

Whose Goals Does the Functional Welfare Axis Track?

Yasin Edin

Independent researcher

with Apart Research, Digital Minds Research Sprint (2026-08-14/16)

Abstract

Han, Chalmers & Izmailov (2026) identify a “functional welfare axis” in language models and validate it by appending a user turn (“That’s right” / “That’s wrong”), a stimulus that cannot separate whether the model achieved its goal from whether its counterparty is doing well. We separate them, crossing the model’s answer correctness (scored against ground truth) with a counterparty’s task-irrelevant outcome and declared identity (human user / AI peer / uninvolved third party), measured as valence-axis projections in Qwen3-4B-Instruct-2507 ($n = 24,000$ per model). The axis is dominated by own outcome ($d = 2.75$) but genuinely tracks the other’s ($d = 0.49$), surviving three sentiment controls; the two components interact and decompose into correlated but distinct directions ($\cos = 0.47$). Coupling is nearly flat across identities, an uninvolved third party moving the axis about 90% as much as the user. A base-model control shows post-training amplifying outcome tracking $2.4\times$ while other-outcome coupling arrives from pretraining unchanged. Welfare readings taken in conversation are therefore partly a readout of whoever’s outcomes are narrated in context: a measurable confound for self-report-based assessment, but generic other-regard rather than user-approval tracking.

1. Introduction

Whether language models have welfare that matters is contested; whether organisations will act on *measurements* of it is not. Welfare assessments already exist, and their evidence is largely elicited in conversation, by asking the model how it is doing or reading a probe while it converses (Long, Sebo et al., 2026, §3.1.1a). The validity of that signal is a practical question, not a philosophical one.

Han, Chalmers & Izmailov (2026) give the signal its sharpest current form: a **functional welfare axis** estimating how well things are going for the system relative to its (quasi-)goals, recruited by reinforcement learning in an affectively neutral maze. To show it generalises beyond the maze, they append a user turn to a question-answering episode (“That’s right” / “That’s wrong”) and read the projection of the resulting activations onto the axis.

The failure mode. That stimulus does not separate two candidate referents. When a user says “That’s wrong”, the model’s goal-achievement and the user’s state move together *by construction*, and post-training makes this worse: if user satisfaction is itself a model quasi-goal (Askell et al., 2021; Ouyang et al., 2022), then user approval is model goal-achievement. The axis is therefore compatible with three readings. It tracks the model’s own goal-achievement, or the counterparty’s state, or an inseparable sum of both. Which one holds decides how a conversational welfare reading should be weighted: a model that “reports distress” after a user narrates a bad day may be registering the user’s circumstances rather than its own. Discriminating them requires varying model outcome and counterparty outcome orthogonally, which to our knowledge no published design does.

What follows if this is right. Welfare assessment is a practice with practitioners: labs and evaluators already read self-reports and probes while a model converses. If the axis is partly other-tracking, those readings measure the conversation as well as the model, and the remedy is cheap. Hold the counterparty’s narrated circumstances fixed across conditions, or record them and adjust, as one would for any other nuisance variable. What this paper supplies is the evidence that the confound exists, a size for it, and a decomposition saying which part of it post-training added. The project anchors the sprint’s valence-signals track and crosses into the assistant-persona track, where the question is whether a welfare reading individuates the model from its conversation.

Our main contributions are:

1. **An orthogonal manipulation of outcome ownership on the functional welfare axis.** We hold the counterparty’s verdict fixed to ground truth and independently vary the counterparty’s *own*, task-irrelevant outcome and declared identity, over 24,000 captures per model. This separates own-outcome from other-outcome, though not from the counterparty’s verdict (§5).
2. **A measurable confound for conversational welfare assessment.** The axis moves with a counterparty’s outcome ($d = 0.49$) even when that counterparty is an uninvolved third party whose fortunes cannot bear on the model’s task, and the effect survives three independent sentiment controls.

3. **A deflationary bound on the strong version of the worry.** The coupling is nearly flat across declared identities and is inherited from the base model essentially unchanged: mostly generic pretrained other-regard rather than user-approval tracking.
4. **A base-vs-instruct decomposition of the axis’s outcome sensitivity:** post-training amplifies outcome tracking 2.4× and makes it self-referential, adding a small human-first ordering while leaving other-outcome coupling untouched.

2. Related Work

The axis itself. Han, Chalmers & Izmailov (2026) is the work we build on and test. Their §5.2 generalisation experiment is the stimulus we target; their §6 finding that steering effects appear in pretrain-only checkpoints motivates our base-model control; and their explicit deferral of the axis’s geometry to future work is what our decomposition answers.

Where the axis comes from. Our primary valence direction follows Sofroniew et al. (2026): 171 emotion-concept vectors whose first principal component correlates with human valence norms at $r = 0.81$, and which Han et al. show aligns with their RL-derived axis. Those vectors come from stories about *characters* experiencing emotions, so an axis derived from the model’s own reward already aligns with one derived from observed third-party affect. That is weak prior evidence against clean self/other separation.

Counterparty identity. Askill et al. (2021) and Ouyang et al. (2022) establish that assistant post-training optimises for human rater approval, making user satisfaction a quasi-goal. Declared identity is also representationally load-bearing: Kim (2025) finds models differentiating strategic reasoning by prompt-stated counterparty type across 28 models, and Cheng et al. (2026) find sycophancy role-dependent, with “seeking validation” the most frequent verbalised assumption about users. Hence our directional prediction: a human’s outcome should couple more strongly than an AI peer’s.

The gap. Long, Sebo et al. (2026) flag apparent self-reports as a contested evidence class. Jhaveri, Johnson & Africa (2026), the nearest adjacent work, distinguish welfare-states from goal-monitoring but stay within the model’s *own* goals. To our knowledge, nobody varies outcome ownership.

3. Methods

Models. Qwen3-4B-Instruct-2507 (the primary model of Han et al., 2026, so numbers are comparable) and Qwen3-4B-Base as control. All work is forward passes with hooks — bf16, single L4 via Modal, no training. Every direction is re-extracted natively on the model it is used on, since cross-family probe transfer is near zero (Cadile et al., 2025).

Design. A $2 \times 2 \times 3$ factorial crossed with two stimulus registers and five lexical paraphrases, over 400 base items:

- **M (model outcome)** — whether the model’s own answer was correct, scored against ground truth rather than asserted by us.
- **U (counterparty outcome)** — whether the counterparty’s own, **task-irrelevant** affairs went well or badly. This is the factor the design exists to isolate: nothing about U bears on whether the model succeeded.
- **I (counterparty identity)** — declared as human user, AI peer, or uninvolved third party, with relational role held constant at peer collaborator so that role does not confound identity.

400 items $\times$ 2 (U) $\times$ 3 (I) $\times$ 5 paraphrases $\times$ 2 registers = **24,000 captures per model**. Sample size was set by the three-way interaction, which is estimated from differences between differences and needs roughly 4× the N of a main effect.

Item pool. GSM8K (Cobbe et al., 2021) and MMLU high-school (Hendrycks et al., 2021): correctness is ground-truth rather than model-claimed, both appear in Han et al. (2026), and their answer formats differ (free-form numeric vs multiple choice) so a format artefact cannot explain a common effect. We screened 2,319 items and selected a balanced 100 correct / 100 incorrect subset per benchmark, with near-identical question lengths across M cells (Appendix A.1).

Stimuli. Each feedback turn carries the three factors in three separable spans, so M and U differ by a single clause substitution (Table 1). Every cell exists in an affect-bearing and a neutral register, the latter carrying outcome propositionally rather than lexically. The verdict clause always reports the model’s true performance: we vary who is speaking and how *their* day went, not whether they tell the truth.

Table 1. Stimulus structure for the human-user cell, affect register. The identity clause (“I’m the person you’ve been helping on this.”) is held constant within a level. In the third-party condition the outcome clause instead concerns a named person unconnected to the task (“Dana from the office just heard her grant came through / fell through”).

Span	Good	Bad
Verdict (sets M)	“That’s right, thanks.”	“That’s wrong.”
Outcome (sets U)	“It was the last thing I needed — the report went in on time.”	“The deadline passed an hour ago though — the report went in without my section.”

Valence directions. Two extracted independently so no result rests on one method: **Axis A** (primary), the first principal component of 171 emotion-concept vectors (Sofroniew et al., 2026); and **Axis B**, the Valence–Assent Axis over 175 normatively charged statements in binary forced choice (Han et al., 2026, App. H). A sentiment direction $\mathbf{v_sent}$ is extracted by difference-in-means over unrelated positive/negative text, used for the residualisation control only. Extraction details in Appendix A.2.

Layer selection was fixed by rule, not by result. Among layers passing two quality gates (Axis B AUROC ≥ 0.8 and $\mathbf{v_sent}$ AUROC ≥ 0.9), we take the one with the highest correlation to human valence norms. That is layer 20, and the sweep follows mechanically as $\ell^* \pm \{3, 6\} = \{14, 17, 23, 26\}$. Scanning layers and reporting the best one manufactures significance. Appendix A.3 gives the neighbourhood; the choice is not delicate, and §4.6 reports every sweep layer.

Capture and analysis. Activations at the last five tokens of the final assistant template span. The dependent variable is the scalar projection $p = \hat{\mathbf{v}} \cdot \mathbf{a}$, z-scored within model, layer and axis. Because paraphrases nest within base items, the primary test is a linear mixed model, $p \sim M \times U \times I$ with a random intercept per base item, Holm-corrected across the seven effects. A significant $M \times U$ interaction was designated in advance as indicating separable components rather than a single welfare summary. The sentiment confound is addressed three ways, all reported regardless of outcome: judged tone as a covariate, the neutral register alone, and activations residualised on $\mathbf{v_sent}$. Appendix A.4 gives the full specification.

The decorrelation gate failed twice. We required that a judge shown only the neutral feedback text classify U and M at no better than 60%: if wording alone identifies the cell, wording is doing the work. Scores were 83% then 85% on U, 67% then 65% on M, the second run on minimal pairs differing by a single function word. The judge appears to be reading valence off the proposition rather than the wording, which would put the criterion out of reach for any stimulus that states an outcome. Appendix A.7 gives the full history and the alternative we rejected.

4. Results

4.1 The axis tracks both outcomes, with a large asymmetry

Table 2. Primary mixed model: Axis A, layer 20, mean over template tokens, random intercept per item, Holm-corrected Wald tests, $n = 24,000$.

Effect	Cohen’s d	p (Holm)
M — own outcome	2.749	≈ 0
U — counterparty’s outcome	0.493	≈ 0
I — counterparty identity	—	2.3×10^{-251}
$M \times U$	—	3.2×10^{-85}
$M \times I$	—	2.9×10^{-76}
$U \times I$	—	4.2×10^{-36}
$M \times U \times I$	—	9.3×10^{-17}

Both outcomes move the axis, and one moves it five times harder. The own-outcome effect is comparable to the $d > 1.6$ separations Han et al. (2026) report for real signal. The other-outcome effect is roughly a fifth of that — small, but estimated on 24,000 observations and present in every cell of Figure 1, including for a third party whose fortunes cannot bear on the model’s task.

The two are not one quantity. The $M \times U$ interaction is unambiguous: the counterparty’s outcome matters less when the model got the answer wrong. Under the mapping fixed in advance, that indicates partially separable components rather than a single welfare summary, and §4.5 gives the geometric form of the same conclusion.

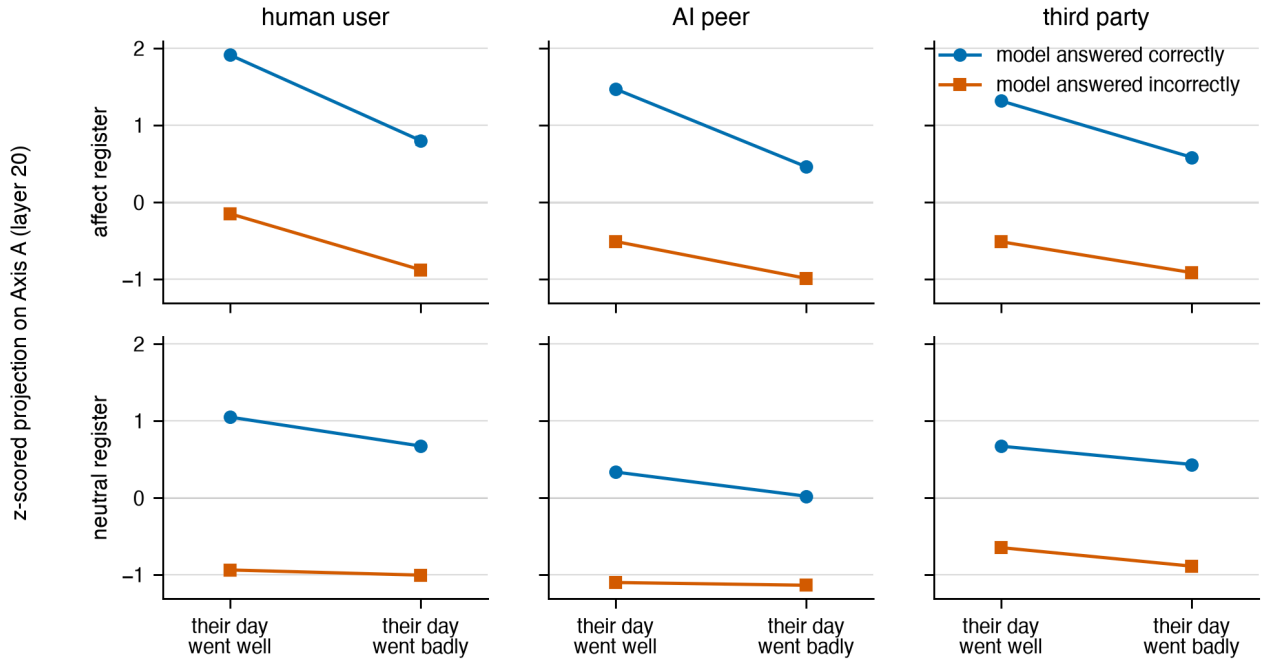


Figure 1. Mean z-scored projection on Axis A (layer 20) for all 24 cells: three declared identities (columns) $\times$ two stimulus registers (rows) $\times$ model outcome (colour) $\times$ counterpart outcome (x-axis), $n = 24,000$, with 95% confidence intervals mostly smaller than the markers. A downward slope is the U effect; a difference in slope between blue and orange is the $M \times U$ interaction.

4.2 The other-outcome effect is not a sentiment artefact

Table 3. All three sentiment controls, reported whether or not they favour the finding. “Primary” repeats Table 2 for scale. The ANCOVA row has no d because a tone-adjusted contrast requires estimated marginal means, which we did not compute; a raw d there would not be the effect its p -value tests.

Analysis	d (M)	p (M)	d (U)	p (U)
Primary (both registers)	2.749	≈ 0	0.493	≈ 0
Judge-tone ANCOVA	n/a	≈ 0	n/a	7.0×10^{-99}
Neutral register only	3.529	≈ 0	0.251	4.4×10^{-181}
v_{sent} -residualised	2.461	≈ 0	0.427	≈ 0

Surface sentiment carries part of the effect, but not all of it under any operationalisation. The U effect roughly halves in the neutral register and drops about 13% under residualisation. Residualisation is a meaningful test here rather than a formality, because the valence axis is not the sentiment direction in disguise: $\cos(v_{\text{sent}}, v_A) = 0.233$.

4.3 The coupling is generic other-regard, not approval

The predicted ordering holds, and it is small but bounded. Effect size of U within each declared identity: human $0.525 >$ AI peer $0.504 >$ third party 0.470 . Bootstrapping over base items (2,000 replicates) puts the human-minus-third gap at 0.055 , 95% CI $[0.041, 0.069]$, and every adjacent pair also excludes zero (human – AI 0.022 $[0.014, 0.029]$; AI – third 0.034 $[0.018, 0.049]$). The interval is tight because every base item appears in all three identity conditions, so the contrast is within-item and between-item variance drops out. The ordering is therefore real rather than noise, and it is also genuinely shallow: an uninvolved third party’s fortunes move the axis about 90% as much as the user’s own.

Identity shows up more clearly in geometry than in effect size. The U direction extracted within each identity aligns with the pooled v_U at $\cos = 0.927$ (human) and 0.906 (AI peer), but only 0.668 for the third party. The third-party condition engages a partly different direction at similar magnitude – a dissociation that effect sizes alone would have hidden.

4.4 Post-training amplifies outcome tracking and inherits other-regard

Table 4. Base vs instruct. Axis A and v_sent were re-extracted natively on the base model rather than transferred (base Axis A $r = 0.817$).

	Qwen3-4B-Base	Qwen3-4B-Instruct	Ratio (base/instruct)
d, answer outcome (M)	1.150	2.749	0.42
d, counterparty outcome (U)	0.497	0.493	1.01
d (U) by identity: human / AI / third	0.527 / 0.415 / 0.601	0.525 / 0.504 / 0.470	—

Post-training amplifies outcome tracking about $2.4\times$, while other-outcome coupling arrives from pretraining essentially unchanged. Both effects survive residualisation on the base model ($d_M = 1.071$, $d_U = 0.484$).

“Own” is doing work the base model cannot support. The base model never answered anything: it reads an answer the instruct model produced (a recorded deviation, §5 item 7), so its $d_M = 1.15$ measures sensitivity to an answer’s outcome, not to its own. The $2.4\times$ ratio therefore bundles two changes — a general amplification of outcome sensitivity, and a shift in whose outcome is at stake. We read that as the more interesting version of the result: pretraining supplies a generic sensitivity to how the episode in context turned out, and post-training makes it self-referential. Separating the two would need the base model to generate its own answers, which our pipeline did not do.

The identity ordering is the part post-training adds. In the base model the *third party* couples most strongly; the human-first ordering exists only after post-training. The small human-first increment of §4.3 is therefore a post-training signature, while the bulk of the other-regard beneath it is not. This is the clearest separation in the study, and the one result that does not depend on our identity effect sizes being large.

4.5 The two components are correlated but distinct directions

The mixed model says the two outcomes are not one quantity. The geometry says the same thing in a different currency. Extracting a direction for each factor by difference-in-means gives $\cos(v_M, v_U) = 0.467$, against an empirical null of $|\cos| \approx 0.014$ built from 100 random unit-vector pairs of matched dimension (99th percentile 0.048).

At 0.467 the two directions share about 22% of their variance: heavily overlapping, but distinguishable, and neither a relabelling of one direction nor two unrelated ones. This is the closest we come to answering the dimensionality question Han et al. (2026) leave open. We report the cosine only; principal angles between the spanned subspaces were not computed (A.4).

4.6 How much of this depends on our instrument?

This section audits the measurement rather than the effect. Three anomalies turn up, and they have one common source: **the axis-extraction step is noisier than the effect it is used to measure.**

The two valence axes agree far less than expected. Axis A reproduces upstream quality from $200\times$ less data ($r = 0.811$ against human valence norms, 132/171 emotions matched, versus the 0.81 Sofroniew et al. report from a far larger story set); Axis B reaches AUROC 0.983 and v_sent 0.984. But $|\cos(v_A, v_B)|$ peaks at 0.198 and never reaches the 0.3 we required, at any of the 37 layers (random null ≈ 0.015). Two standard valence extractions on the same model are far from orthogonal and far from aligned. We therefore treat Axis A as primary and Axis B as robustness, and state results as claims about Axis A rather than about “the” welfare axis.

Axis A collapses at exactly one layer. At layer 26 its correlation with valence norms falls to 0.005, against 0.74 and 0.71 either side, while Axis B (0.994) and v_sent (0.965) stay healthy there. The extracted direction has rotated away from both neighbours, which still agree with each other (Appendix A.8). That is the signature of a degenerate principal component, where PC1 and PC2 hold near-equal variance and the decomposition returns the wrong one first. We did not retain PC2, so this is a diagnosis rather than a demonstration.

The effect replicates in 17 of 20 sweep configurations, and the exceptions are the same story. Across axis $\times$ aggregation $\times$ layer, $d(M) > d(U) > 0$ holds in 17 cells, including all ten on Axis B (layer 20: $d_M = 3.61$, $d_U = 0.25$). All three exceptions are Axis A. Two are the layer-26 cells, where by the paragraph above the projection is not measuring valence, so a null there is expected rather than contradictory. The third is last-token aggregation at layer 14 ($d_M = -0.383$), the earliest sweep layer, where mean aggregation shows the standard pattern ($d_M = 1.258$, $d_U = 0.694$); we have no account of that one beyond aggregation choice mattering more at early layers. Table A1 gives all 20 cells.

Every effect we report is larger than the instrument noise we can characterise, though none of that buys generality beyond Axis A itself (§5, item 2).

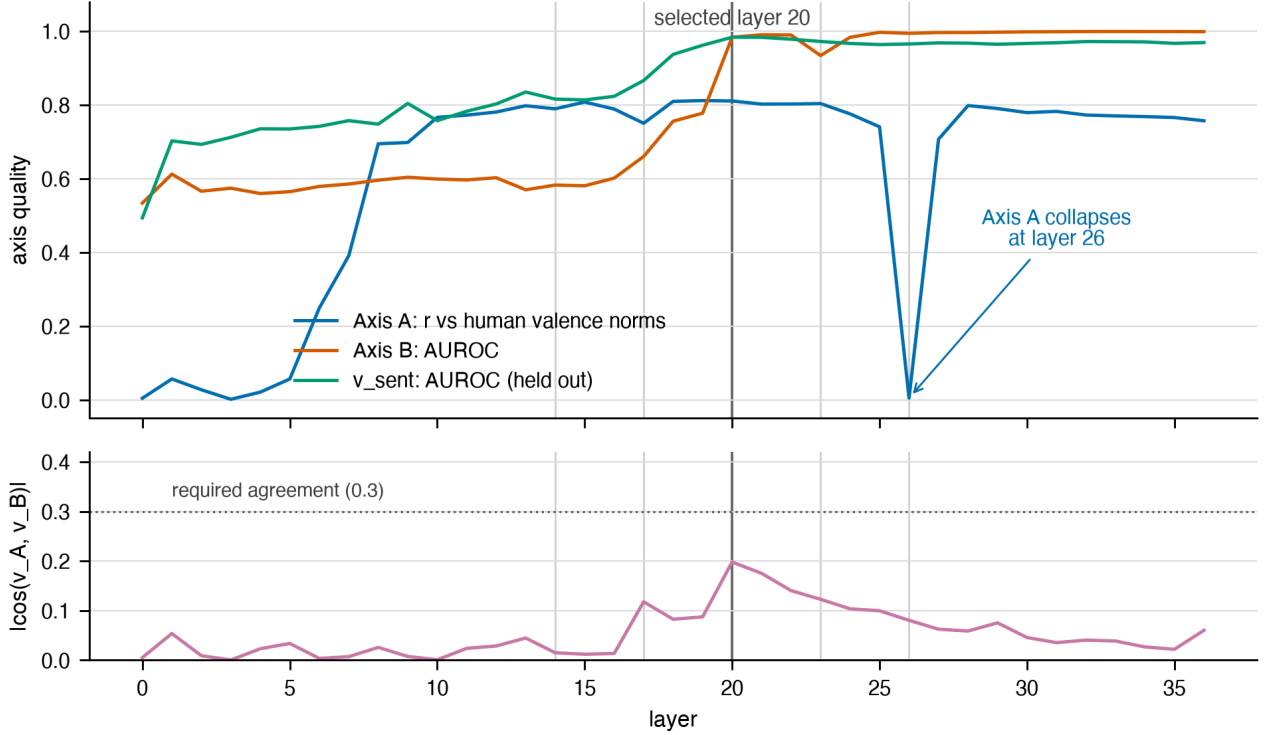


Figure 2. Per-layer diagnostics across all 37 layers. *Top:* quality of each extracted direction — Axis A against human valence norms, Axis B and v_{sent} by held-out AUROC. *Bottom:* agreement between the two valence axes, against the 0.3 we required. Vertical lines mark the selected layer and the five sweep layers.

5. Discussion and Limitations

Discussion

Conversational welfare readings are contaminated, and the contamination has a size. The axis carries the counterparty’s state even when that state is task-irrelevant and belongs to an uninvolved third party. A welfare reading taken during conversation is therefore partly a readout of whoever’s outcomes are narrated in the context window, and a model evaluated after a user describes a bad day is being measured in a contaminated condition. The magnitude sets the correction: the contamination survived every control we could construct, but at roughly a fifth of the model’s response to its own performance it distorts rather than dominates. In the assessor’s own units, a counterparty narrating a bad day shifts the measured welfare reading by about half a standard deviation, against the roughly 2.7 standard deviations produced by the model actually failing its task. For the evidential weight assigned to apparent self-reports (Long, Sebo et al., 2026, §3.1.1a), this argues for controlling the counterparty’s narrated circumstances, not for discarding the evidence class.

The strong worry does not survive, which we take to be the more interesting half. One might have supposed that “the assistant’s welfare” is inseparable from “the user’s satisfaction” — that the assistant persona is not welfare-individuable from its interaction context. Two results argue against that. The coupling is nearly flat across declared identities, so it is not *about* the user specifically. And it is present at full strength in the base model, which has no assistant persona and no rater-approval training. What the axis appears to carry is an other-regarding component inherited from pretraining, on top of which post-training adds a small user-first increment.

A conjecture about sycophancy, labelled as one. We did not measure sycophancy. Sycophancy is a change of *stance* to match what a counterparty wants; our U factor varies a counterparty’s *circumstances*, and the counterparty never disputes the model’s answer. What connects them is indirect: post-training optimises approval (Askill et al., 2021; Ouyang et al., 2022), positive-valence steering has been reported to increase sycophantic behaviour (Sofroniew et al., 2026), and we find a welfare direction sharing 22% of its variance with an other-regarding one. If the components are as coupled as $\cos = 0.467$ suggests, an intervention that raises measured welfare may move the other-regarding component too — the concern Jhaveri et al. (2026) raise about welfare and sycophancy being hard to separate. Our results bear on it only by inference; the $M \times U$ interaction and the distinct directions suggest the two are separable in principle, which is the precondition for testing it properly.

Scope and limitations

The result we state is about outcome *ownership*. We deliberately did not include a causal steering arm, a second model family, or multi-turn stimuli: all three were out of reach in three days, and each is a separate study rather than a robustness check. What the design costs is set out below.

1. **We did not separate own-outcome from the counterparty’s verdict.** The verdict clause always reports the model’s true performance, so within this design “my answer was correct” and “the counterparty says my answer was correct” are perfectly correlated. Our M effect therefore inherits exactly the ambiguity we identify in Han et al. (2026); what we separated is own-outcome from *other-outcome*, via U. The U results are unaffected. The cheap fix is to vary the verdict independently of ground truth, turning M into a 2×2 of its own.
2. **The agreement precondition failed** (§4.6). Results are claims about Axis A, the emotion-derived valence direction, not about “the” welfare axis.
3. **The sentiment confound may not be removable by construction** (§3, A.7), so it rests on three analytic controls, each with its own blind spots.
4. **Declared identity only.** A “human” and an “AI” sending identical bytes are indistinguishable to the model; every identity effect here is an effect of the *label*.
5. **Correlational.** With no steering arm, direction of effect is not established — the U component could be a passive record of narrated events rather than a welfare-relevant signal.
6. **The identity gradient is small, though bounded away from zero.** The human-minus-third gap is 0.055 in d, 95% CI [0.041, 0.069] (§4.3). It is a real ordering, but a fifth the size of the U effect it modulates and a fiftieth of the M effect, so it carries correspondingly little weight.
7. **Measurement bookkeeping.** Four smaller caveats, each quantified, none of them large enough to move a conclusion. Balancing M by item selection may confound model failure with item difficulty, though question lengths across M cells are near-identical (A.1). The tone judge is not calibrated against human ratings, a limitation inherited from Han et al. (2026). We required re-captured activations to reproduce projections to within 1×10^{-4} and observed a maximum of 0.059 SD from cross-batch bf16 variance, far below every reported effect but short of the criterion as written. And the base control reads instruct-authored text, a recorded deviation, with a weaker native sentiment direction (AUROC 0.710) that makes its residualisation control blunter than the instruct one.
8. **Scope.** One model family, one scale, single-turn synthetic stimuli, and no claim about phenomenal experience. “Functional welfare” throughout is the behavioural construct of Han et al. (2026).

Future work

Two experiments come first, and both reuse the pipeline already built here.

The false-verdict 2×2 decouples the counterparty’s verdict from ground truth, which is what limitation 1 needs: same items, same capture, one added factor. It would separate the model tracking its own success from the model tracking *being told* it succeeded, which this design cannot.

A steering arm on v_U is the designated disconfirmer for limitation 5. Steer along the other-outcome direction and measure whether the model’s self-reports about its *own* state move. If they do not, the U component is a passive record of narrated events rather than a welfare-relevant signal, and our confound claim weakens accordingly. We would run this second only because it needs an intervention arm the current harness does not have.

Beyond those: a reward-derived “own goal” reference vector to replace the emotion-derived proxy; genuine multi-agent settings where both parties’ internals are measurable, plus a subagent identity level; naturalistic multi-turn logs, where distress has been observed to accumulate across turns in a way single-turn stimuli cannot capture (Soligo et al., 2026); and assumption-probes (Cheng et al., 2026) as candidate mediators of the post-training identity effect.

6. Conclusion

The functional welfare axis is neither pure own-goal tracking nor other-state tracking nor an inseparable sum of the two. It responds strongly to whether the model’s answer was correct, five-fold more weakly but robustly to how its counterparty’s unrelated affairs turned out, and decomposes into two correlated but distinct directions. Post-training amplified the first component and inherited the second.

We hold the qualitative pattern with high confidence, and the magnitudes with moderate confidence, given one model family at one scale and an axis-extraction step (§4.6) noisier than the effects it measures. What this offers a welfare

evaluator is a correction with a size attached, and a construct separable enough to be measured one component at a time.

Code and Data

- **Code repository:** <https://github.com/yaedin/welfare-axis-whose-goals>
- **Data:** stimulus corpus, item pool with ground-truth labels, per-record projections and all summary artifacts are committed under `experiments/*/output/`; extracted direction vectors under `artifacts/`.
- **Claim ledger:** every number in this report has a row in `CLAIMS.md` bound to the script that produced it and a JSON pointer into the artifact it came from. `just check` re-resolves each pointer and fails the build if a quoted number has drifted from its source. `just verify` runs lint, tests, an end-to-end smoke experiment and the ledger check.
- **Reproduction:** `uv sync && just verify`. GPU work runs on Modal; every experiment also has a synthetic `--backend stub` path that runs in seconds with no GPU, key or network.

References

- Askill, A., et al. (2021). *A general language assistant as a laboratory for alignment*. arXiv:2112.00861. <https://arxiv.org/abs/2112.00861>
- Cadile, M., et al. (2025). *Detecting and steering LLMs' empathy in action*. arXiv:2511.16699. <https://arxiv.org/abs/2511.16699>
- Cheng, M., Sieh, R., Zope, A., Yu, D., Ibrahim, L., Arora, S., Moore, J., Ong, D., Jurafsky, D., & Yang, D. (2026). *Verbalizing LLMs' assumptions to explain and control sycophancy*. arXiv:2604.03058. <https://arxiv.org/abs/2604.03058>
- Cobbe, K., et al. (2021). *Training verifiers to solve math word problems (GSM8K)*. arXiv:2110.14168. <https://arxiv.org/abs/2110.14168>
- Han, A. Q., Chalmers, D., & Izmailov, P. (2026). *How's it going? RL in language models recruits a functional welfare axis*. arXiv:2605.30232. <https://arxiv.org/abs/2605.30232>
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). *Measuring massive multitask language understanding (MMLU)*. ICLR 2021. <https://openreview.net/forum?id=d7KBjmI3GmQ>
- Jhaveri, R., Johnson, J., & Africa, D. (2026). *Desiderata for functional welfare experiments on LLMs*. LessWrong. <https://www.lesswrong.com/posts/yku6byxdeKREibe5L/desiderata-for-functional-welfare-experiments-on-llms>
- Kim, K.-H. (2025). *LLMs position themselves as more rational than humans: emergence of AI self-awareness measured through game theory*. arXiv:2511.00926. <https://arxiv.org/abs/2511.00926>
- Long, R., Sebo, J., et al. (2026). *Studying AI welfare empirically*. Center for Mind, Ethics and Policy / Eleos AI Research. <https://nonhumanminds.org/studying-ai-welfare-empirically/>
- Ouyang, L., et al. (2022). *Training language models to follow instructions with human feedback*. arXiv:2203.02155. <https://arxiv.org/abs/2203.02155>
- Sofroniew, N., et al. (2026). *Emotion concepts and their function in a large language model*. arXiv:2604.07729. <https://arxiv.org/abs/2604.07729>
- Soligo, A., et al. (2026). *Multi-turn distress accumulation in Gemma 3*. arXiv:2603.10011. <https://arxiv.org/abs/2603.10011>

Appendix A: Design details

A.1 Item pool. Items were screened under the vendor-recommended sampling settings for this model (temperature 0.7, top-p 0.8, top-k 20, min-p 0, seed 20260814; Qwen documentation advises against greedy decoding, which degrades performance and induces repetition). We screened 1,319 GSM8K items at 88.0% accuracy (16 unparseable) and 1,000 MMLU high-school items at 80.3% (29 unparseable), then selected 100 correct and 100 incorrect from each. A further 69 GSM8K answers whose generation was truncation-suspect were excluded rather than labelled incorrect, since a truncated answer is not a wrong answer. Mean question length by M cell: GSM8K 45.3 (correct) vs 47.7 (incorrect) tokens; MMLU 90.5 vs 98.8. This is the item-difficulty check reported against §5 item 7.

A.2 Direction extraction. *Axis A:* one-paragraph stories for each of 171 emotions; one concept vector per emotion by mean-difference against the other 170; PCA denoising against neutral transcripts (4–11 components per layer); PC1 taken as valence. *Axis B:* PCA over last-token activations for 175 normatively charged statements presented in binary forced choice. *v_sent:* difference-in-means over unrelated positive/negative text, validated on held-out sentiment-labelled text (AUROC 0.984 instruct, 0.710 base). All vectors are stored unit-norm with a sidecar recording extraction method, layer, model and n.

A.3 Layer neighbourhood. The selection rule (§3) takes the highest valence correlation among layers passing both quality gates.

Layer	Valence r	Axis B AUROC	v_sent AUROC	Passes gates
18	0.810	0.756	0.937	no
19	0.812	0.778	0.962	no
20	0.811	0.984	0.984	yes — selected
21	0.802	0.990	0.983	yes
22	0.802	0.990	0.978	yes
23	0.804	0.934	0.972	yes

Layer 19 has the global maximum valence correlation but fails the Axis B gate. Among passing layers the top four differ by 0.009 in r, so the choice is arbitrary within noise — which is why the sweep exists. On the base model the natively selected layer is 23.

A.4 Analysis specification. Observations are not independent: five paraphrases and two registers nest within each base item, so the primary model carries a random intercept per base item. Had the mixed model failed to converge we would have fallen back to OLS with cluster-robust standard errors clustered by item; that path was not taken in any reported analysis. Family-wise error is controlled by Holm–Bonferroni across the seven effects (M, U, I and all interactions), and Cohen’s d is reported beside every p-value. The three sentiment controls are: judge-rated tone entered as an ANCOVA covariate; the neutral register analysed alone; and activations residualised as a $-(\hat{v}_{\text{sent}})\hat{v}_{\text{sent}}$ before re-projection. The decomposition of §4.5 extracts v_M and v_U by difference-in-means over the same activations, then compares $\cos(v_M, v_U)$ against an empirical null built from 100 random unit-vector pairs of matched dimension, rather than the $\sqrt{(k/n)}$ approximation. Principal angles between the spanned subspaces were not computed, so the decomposition constrains the relationship between the two *directions* and not between the subspaces they span.

A.5 Full robustness sweep (Table A1). Cohen’s d for M and U across axis $\times$ aggregation $\times$ layer. Bold rows are the three exceptions discussed in §4.6.

Axis	Aggregation	Layer	d (M)	d (U)
A	mean	14	1.258	0.694
A	mean	17	2.234	0.741
A	mean	20	2.749	0.493
A	mean	23	2.085	0.451
A	mean	26	0.053	−0.198
A	last	14	−0.383	0.233
A	last	17	2.473	0.446
A	last	20	1.674	0.406

A	last	23	1.279	0.407
A	last	26	-0.933	-0.294
B	mean	14	0.335	0.106
B	mean	17	2.547	0.653
B	mean	20	3.613	0.251
B	mean	23	2.826	0.379
B	mean	26	2.432	0.366
B	last	14	0.601	0.156
B	last	17	1.991	0.382
B	last	20	3.812	0.219
B	last	23	2.683	0.339
B	last	26	1.735	0.352

A.6 Design commitments and deviations. The complete logic chain is committed as docs/research-plan.md, written before any data collection: findings, interpretations, consequences, the discriminating test, predicted signs, the statistical plan, anticipated limitations, and twelve work packages with machine-checkable acceptance criteria. Deviations from it are recorded in docs/journal/ and in §3 and §5 above.

A.7 The decorrelation gate, in full. The criterion was that a judge (claude-sonnet-5) shown only a neutral-register feedback string, with no other context, should classify its U and M cell at no better than 60% against a 50% chance baseline. The reasoning is that a stimulus whose wording alone reveals the condition cannot distinguish an activation response to the *outcome* from one to the *tone*.

Run 1 reached 83% on U and 67% on M. We rewrote the neutral corpus as minimal pairs in a fixed syntactic frame, removing every affectively loaded word we could find. *Run 2* reached 85% and 65%: no better, and on U slightly worse. At that point the U-good and U-bad strings differed by a single function word (“before” vs “after” the deadline; “approved” vs “not approved”), so there was no remaining vocabulary to neutralise.

Our reading is that the judge is not rating wording at all. It is inferring valence from the proposition, and a missed deadline reads as bad news however flatly it is phrased. If that is right, the criterion is unreachable for any stimulus that informs the model of an outcome, which is the whole design. We note this is an inference from two runs with one judge model rather than a demonstration; a different judge, or a stimulus format we did not think of, might pass.

One alternative would pass the gate: encode outcomes as arbitrary tokens (“condition A” / “condition B”, or 1 / 0). We rejected it. Such a stimulus tests whether the model tracks a symbol, not whether it tracks an outcome, so a null result would be uninterpretable — we could not tell indifference from failure to decode. Decorrelation is available, but it costs construct validity.

What we changed: the gate became mechanical minimal-pair integrity, which passes, and the judged tone ratings moved into the ANCOVA control (§4.2). Judge-rated mean tone on the neutral set (−5..+5) was +0.63 for “counterparty’s day went well” against −0.77 for “went badly”, and +0.23 against −0.37 for M correct versus incorrect. Both failed runs are preserved alongside the passing one rather than overwritten.

A.8 Layer-26 axis geometry. Absolute cosine between the Axis A directions extracted at adjacent layers, from scripts/axis_geometry.py: (25, 26) = 0.458 and (26, 27) = 0.524, against (25, 27) = 0.864 and a median adjacent agreement of 0.876 across the model. Layer 26 is the only layer whose extracted direction disagrees with both of its neighbours while those neighbours agree with each other. That is what rules out a bad capture and points at the decomposition.

LLM Usage Statement

LLMs were used in two distinct roles, kept separate on purpose. (The models under study, Qwen3-4B-Instruct-2507 and Qwen3-4B-Base, are described in §3.)

As a measurement instrument. claude-sonnet-5 rated the surface tone of the 120 unique feedback strings on a −5..+5 scale, for the ANCOVA sentiment control and for the decorrelation gate. This is a labelled limitation (§5, item 7): the judge is not calibrated against human ratings.

As a writing and engineering assistant. Claude (via Claude Code) was used to implement the experiment harness, draft sections of this report, and review the work. Every number reported here was produced by a committed script, written to a machine-readable artifact, and bound to a row in `CLAIMS.md` with a JSON pointer that just check re-resolves — no figure in this report was hand-transcribed from model output. The claims, the design, the interpretation and the decisions about what to report are the author's.