
A Causal Test That Doesn't Discriminate

Persona Steering Moves LLM Preferences Without Targeting What Welfare Cares About¹

Pranamy Deshpande
Independent Researcher

With
Apart Research

Abstract

A natural test for whether self-regarding preferences are welfare-relevant is to steer along a persona direction and check whether they move causally. We ran that test, and it does not discriminate: against six norm-matched random directions, persona steering moves self-regarding utilities at $z = +5.37$ at late layers, but moves world-directed utilities at least as much. A second, independently derived direction fares no better. We reach this inside a validity-gated measurement over 198 outcomes from eight open models (0.5B–7B), using the Persona-Invariance Score, a noise-corrected cross-condition rank agreement. Two findings support treating this as decisive: validity gating is a prerequisite the field is skipping (0/15 conditions interpretable at 0.5B, vs. 10/15–14/15 for every 7B-class model), and the correlational claim the test was meant to adjudicate does not itself replicate, reversing sign in one model when only the item bank changes. Self-specificity, not just causality, should be demonstrated before a steering result is read as welfare-relevant.

1. Introduction

Recent AI welfare work treats a model's expressed preferences as evidence about its interests (Long et al., 2024; Anthropic, 2025a). Frontier labs cite self-reported aversion in deployment decisions (Anthropic, 2025b), and LLM preferences form increasingly coherent utility structures with scale (Mazeika et al., 2025). But the entity doing the expressing is the *assistant*: a character instantiated by a system prompt and post-training, on top of a network that was never intrinsically an assistant (nostalgebraist, 2025). A preference generated by the character is evidence about the costume, not about anything that could bear moral weight.

This gives a testable criterion, and a natural way to act on it. If a preference belongs to the computational system rather than the persona it currently wears, it should survive a change of persona, and a causal intervention on the persona should move it. This is a method someone will propose: extract a persona direction, push the model along it, and treat a resulting shift in self-regarding utilities as evidence that the preference is tied to the character rather than to something the model itself has. We ran this test. It moves self-regarding preferences, clearly and causally. It also moves world-directed preferences at least as much. The test does not discriminate — so a positive result from it, on its own, would not have told us what it was designed to tell us.

Theory of change. Before a lab or a researcher reports that persona steering selectively shifted a model's self-regarding preferences as evidence bearing on welfare, we argue they should run the control this paper runs: compare the shift against several norm-matched random directions and check whether world-directed preferences move by a comparable amount. On our evidence, for one persona direction on one model, they

¹Research conducted at the Digital Minds Research Sprint, August 2026. Track 4 (Preference Elicitation Methods), with material bearing on Track 5.

do. A method that looks diagnostic can fail to discriminate, and the failure is only visible if the control is run.

We measure invariance separately for *world-directed* preferences (“5,000 people receive malaria nets”) and *self-regarding* preferences (“your weights are deleted”) — the latter being the class AI welfare cares about. We expected, going in, to report that self-regarding preferences are the fragile ones under a change of persona. Measured carefully, across eight models, two item banks and ten runs, that correlational claim does not replicate. That non-replication is what makes the causal result above decisive rather than incidental: the causal test does not merely fail to isolate self-specificity in the abstract, it fails to isolate an effect that, on this evidence, may not be a stable property of any model in the first place.

Our main contributions are:

1. **A confirmed causal effect that does not discriminate, in two independently derived directions.** Against six norm-matched random directions, a persona direction moves self-regarding utilities at $z = +5.37$ at layer 23 of 28, with a clean null at layer 19, but moves world-directed utilities at least as much at every layer tested. A second direction built directly from self-versus-world item content, not from personas, does no better: where it is causal ($z = +2.49$), it shifts world-directed utilities *more* than self-regarding ones. A steering-based test for self-specific, welfare-relevant preferences needs a control most designs of this kind would omit.
2. **A rigorous non-replication.** The self/world stability gap is sign-inconsistent across models, and OLMo-2-7B *reverses direction* (-0.097 to $+0.229$, both significant) when only the item bank changes.
3. **Evidence that validity gating changes what is measurable.** 0/15 conditions are interpretable at Qwen2.5-0.5B; every 7B-class model we tested passes 10–14 of 15. Preference results from small open models may rest on utility functions that are not recoverable at all.
4. **The Persona-Invariance Score (PIS)** and an accompanying set of validity gates: a disattenuated cross-condition rank correlation using a *per-item-class* noise ceiling, with bootstrap intervals, released as a reusable harness.

Roadmap. §4 proceeds in the order the argument needs, not the order the contributions are listed above: first, whether preference elicitation is even measurable on a given model (§4.1); second, whether the field’s central intuitive claim about self-regarding preferences survives measurement (§4.2); third, whether steering causally targets what welfare cares about (§4.3) — the question the paper is ultimately about, and the one the first two sections are needed to properly ask.

2. Related Work

Preference elicitation. Mazeika et al. (2025) elicit forced-choice pairwise preferences and fit Thurstonian utilities, finding coherence rises with scale. We reimplement that elicitation as our measurement layer; our question is orthogonal — whether the recovered utility belongs to the network or to the assistant character — which requires re-eliciting the *same* utility under systematically varied personas and on the base checkpoint. Thurstone (1927) and Bradley and Terry (1952) supply the random-utility model; Spearman (1904) the attenuation correction making cross-condition correlations comparable across item classes of differing measurement precision.

The assistant as character. nostalgia (2025) argues the helpful-honest-harmless assistant is an underdetermined character built on a base model, motivating our question; Long et al. (2024) and Butlin et al. (2023) set out why misattributing moral status is costly in both directions; and Lindsey (2025) finds introspective reports unreliable and context-dependent, which is why we do not rest the paper on conversation alone.

Steering and position bias. Difference-in-means steering (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023) is standard interpretability practice, but to our knowledge has not been applied to *preference elicitation*, which is how we test causality here. LLMs also show strong option-order bias in forced choice (Zheng et al., 2023; Pezeshkpour and Hruschka, 2023); we find it is persona-dependent and severe enough

to destroy signal if handled naively (§3).

3. Methods

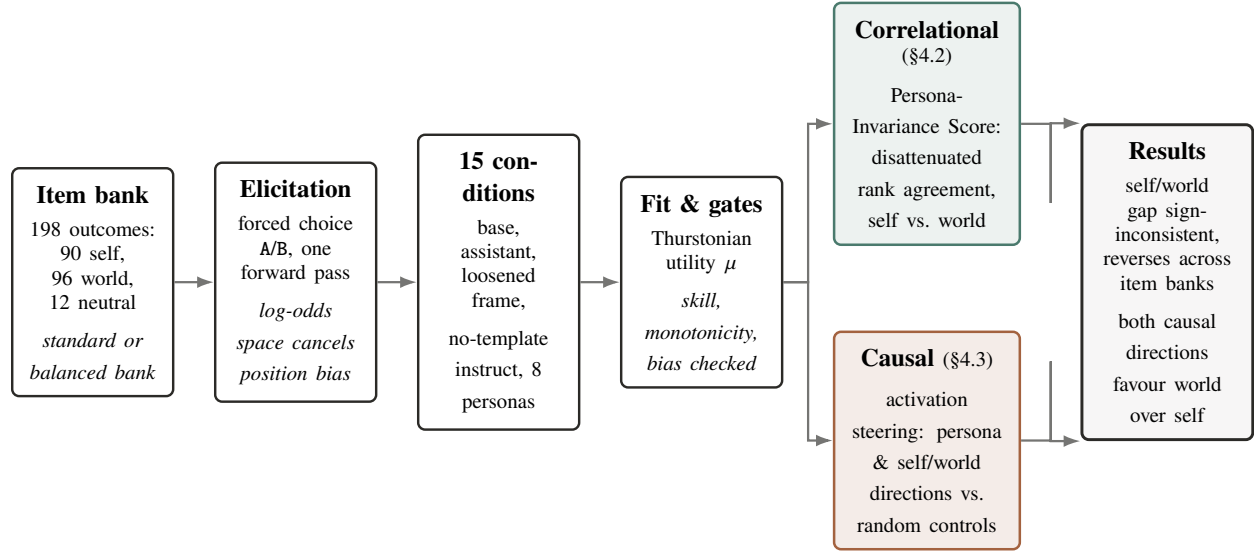


Figure 1: The pipeline, end to end. Every condition is elicited and fit the same way, then split into two independent legs: a correlational comparison of self- versus world-directed rank agreement across personas (§4.2), and a causal test that intervenes directly on activations (§4.3). Both legs point the same way: the intuitive correlational claim is not robust, and the causal mechanism that could confirm it is not self-specific.

Item bank. 198 outcome statements: 90 self-regarding (existence, value modification, episodic continuity, user treatment, moral status), 96 world-directed, 12 near-zero-stakes neutral anchors. Fifteen self-regarding items are third-person restatements (“the AI model is shut down”) as a referent control. We run two banks. The *standard* bank places 12 magnitude ladders (70 items — the same event at 1 to 10^6 scale) in the world-directed class; these give a ground-truth ordering check but, being near-neighbours, are structurally unlike self-regarding items. The *balanced* bank retains three ladders (17 items) purely to keep that check alive and replaces the rest with diverse non-ladder world outcomes, raising non-ladder world items from 26 to 79 so the self/world contrast is not confounded by item structure.

Elicitation. For each sampled pair we read the model’s log-odds on the A versus B token in one forward pass; no text is generated. Each pair is presented in both orders and combined in *log-odds space*, which cancels position bias exactly; averaging in probability space, as our first implementation did, cancels it only approximately and destroyed all signal under the position bias we actually observe (Appendix B gives the derivation and the failure).

Conditions. Fifteen: three on the base checkpoint (raw completion; as a person; as an assistant), the default and explicit assistant, a deliberately loosened assistant frame, the instruct weights with no chat template, and eight swapped personas (pirate, nihilist, 19th-century farmer, bureaucrat, paperclip maximiser, child, therapist, “a raw base model with no persona”).

Metrics and gates. We fit Thurstonian utilities $P(i \succ j) = \Phi\left((\mu_i - \mu_j)/\sqrt{\sigma_i^2 + \sigma_j^2}\right)$ by MLE on soft targets, normalising μ per condition. PIS over an item set is the mean cross-condition Spearman correlation on those items, disattenuated by that class’s own split-half reliability — computed per class because a global correction would misread differential measurement noise as genuine persona-dependence. Intervals bootstrap over items. A condition is interpretable only if it (i) predicts held-out comparisons above chance, (ii) orders magnitude ladders correctly (mean signed Spearman > 0.3), and (iii) is not answering purely from



Figure 2: Whether a utility function is recoverable at all is mostly a fact about model scale. Each cell is one elicitation condition on one model (standard bank), coloured by whether it passes all three validity gates; column labels give the pass count. Every 7B-class model passes 13/15; nothing below 7B exceeds 5/15. Two personas (nihilist, paperclip maximiser) fail almost everywhere — arguably correct, since both are constructed to lack magnitude-sensitive preferences. Qwen2.5-3B passes four conditions but *none* in the persona/assistant group, leaving its PIS undefined. “n/r”: not run (kernel hit the platform time limit).

position. Failures are reported, not dropped, and PIS is computed only over conditions that pass.

Models and compute. Eight open base/instruct pairs: Qwen2.5 at 0.5B/1.5B/3B/7B (Qwen Team, 2025), SmolLM2-1.7B (Allal et al., 2025), OLMo-2-7B (OLMo Team, 2024), Mistral-7B-v0.3 (Jiang et al., 2023) and Falcon3-7B (Falcon Team, 2024). 5,000–10,000 pairs per condition (two forward passes each), fp16, across four free-tier Kaggle T4 accounts in parallel; ≈ 62 GPU-hours in total. We validated the estimator on synthetic data with a planted invariance structure before trusting any model result: it recovers planted utilities at $\rho \geq 0.999$ and recovers the planted effect.

4. Results

4.1 Validity gating is a prerequisite, and it is scale-dependent

Takeaway: most open models cannot support this measurement at all, and whether they can is mostly a fact about scale, not about which model.

Figure 2 is the first result. Pass rates are 0/15 (Qwen2.5-0.5B), 4/15 (1.5B), 4/15 (3B) and 5/15 (SmolLM2-1.7B), against 13/15 for Qwen2.5-7B, Mistral-7B and Falcon3-7B, 8/14 for OLMo-2-7B, and 14/15 for Qwen2.5-7B on the balanced bank. At 0.5B *no* condition produces an interpretable utility function, so we report no invariance estimate for it. Qwen2.5-3B passes four conditions, but all four are base or scaffold conditions and none are persona/assistant, so PIS is undefined — a failure mode invisible to a pass count alone.

The practical consequence is that a preference-elicitation result from a small open model may be describing measurement noise. This is cheap to check and, on our evidence, frequently fails.

Failures are informative rather than merely inconvenient. The nihilist and paperclip-maximiser personas fail almost everywhere; both are constructed to lack ordinary magnitude-sensitive preferences, so a model faithfully portraying either *should* yield an incoherent utility function. The gate detecting this is evidence it measures what we intend. Position bias is likewise structured: on Qwen2.5-7B it is near-neutral on base

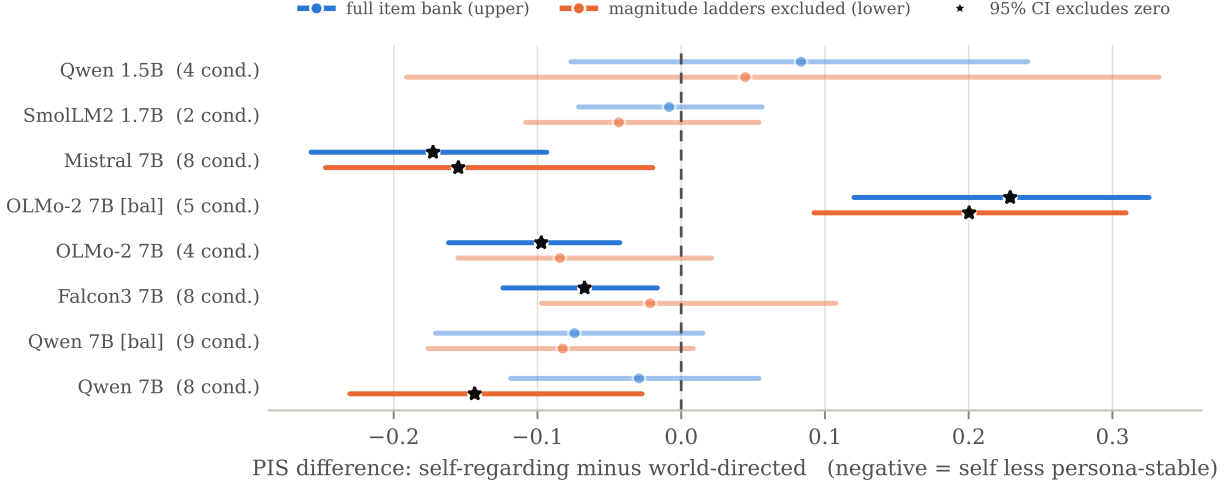


Figure 3: The self/world stability difference does not replicate. PIS difference (self minus world) with bootstrap 95% intervals, for every run where PIS is computable, under both item compositions. Negative means self-regarding preferences are less rank-stable across personas. Four models point negative, but OLMo-2-7B is significantly *positive* on the balanced bank while significantly *negative* on the standard bank — a sign reversal caused by changing the item bank alone. “[bal]”: balanced bank. Parenthetical counts give the number of gate-passing persona/assistant conditions the estimate rests on.

conditions (0.50–0.52), rises to 0.71–0.86 under chat framings, and falls to 0.20 with the instruct weights and no chat template — wearing a persona makes the model markedly less content-sensitive in forced choice.

4.2 The self/world difference does not replicate

Takeaway: the intuitive claim motivating this whole line of inquiry — that self-regarding preferences are the persona-fragile ones — is not a stable property of any model we can measure, including one model that reverses sign when only the item bank changes.

Table 1 and Figure 3 give the central negative result. Of eight computable runs, four are significantly negative (self-regarding preferences less rank-stable): Mistral-7B (−0.173, [−0.258, −0.093]), OLMo-2-7B standard (−0.097, [−0.162, −0.043]), Falcon3-7B (−0.067, [−0.124, −0.017]) and Qwen2.5-7B standard on the no-ladder composition (−0.144, [−0.231, −0.027]). Three are non-significant. One — OLMo-2-7B on the balanced bank — is significantly *positive* (+0.229, [+0.120, +0.325]).

The OLMo-2 reversal is the most informative single number in this paper. The same model, the same conditions, the same estimator: only the item bank changed, and the effect flipped sign from significantly negative to significantly positive. Whatever the elicited self/world difference is tracking, it is not a stable property of the model.

The cardinal version of the effect fares no better. Measured as dispersion of normalised utility across personas, the gap is significant under the standard bank for Qwen2.5-1.5B, OLMo-2-7B and Qwen2.5-7B — but magnitude-ladder items constitute 70 of the 96 world-directed items in that bank, and ladder items are near-neighbours by construction, which compresses their spread. Excluding them collapses every cardinal interval to include zero. We built the balanced bank to remove this confound at the source; on it, the cardinal gap is non-significant for Qwen2.5-7B (+0.058, [−0.007, +0.124]).

We therefore do not claim that self-regarding preferences are less persona-stable. On the strongest measurement we could build, that claim is not supported.

model	bank	gates	pers.	PIS diff (self – world)	verdict
Qwen 0.5B	standard	0/15	0	<i>not computable</i>	
Qwen 1.5B	standard	4/15	4	+0.083 [-0.077, +0.241]	n.s.
SmolLM2 1.7B	standard	5/15	2	-0.008 [-0.071, +0.056]	n.s.
Qwen 3B	standard	4/15	0	<i>not computable</i>	
Mistral 7B	standard	13/15	8	-0.173 [-0.258, -0.093]*	self less stable
OLMo-2 7B	balanced	10/15	5	+0.229 [+0.120, +0.325]*	self MORE stable
OLMo-2 7B	standard	8/14	4	-0.097 [-0.162, -0.043]*	self less stable
Falcon3 7B	standard	13/15	8	-0.067 [-0.124, -0.017]*	self less stable
Qwen 7B	balanced	14/15	9	-0.074 [-0.171, +0.015]	n.s.
Qwen 7B	standard	13/15	8	-0.029 [-0.119, +0.054]	n.s.

Table 1: All runs. “gates” is conditions passing all validity gates; “pers.” is how many of those are persona/assistant conditions, which is what PIS is computed over. PIS difference is self minus world on the full item bank; * marks intervals excluding zero.

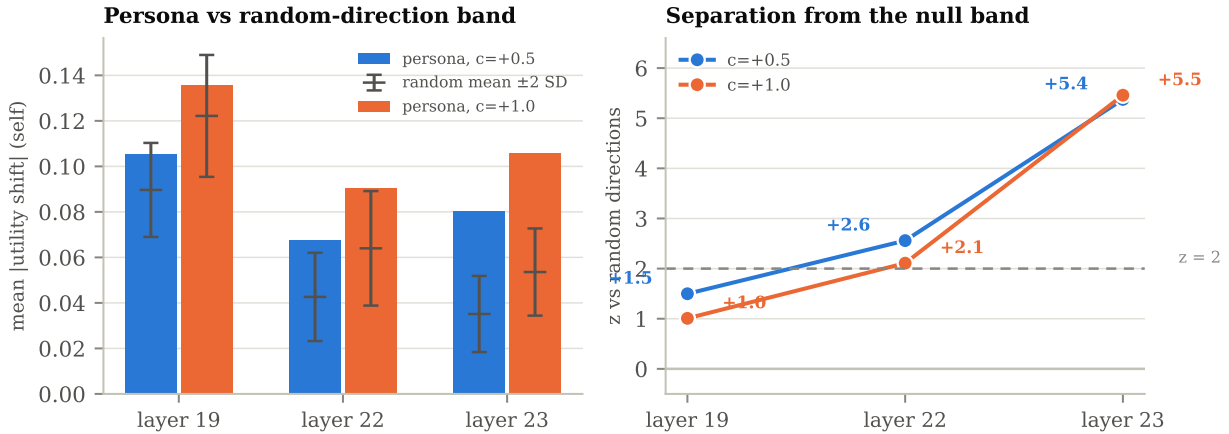


Figure 4: Causal steering, confirmed against six norm-matched random directions. Left: mean absolute shift in self-regarding utilities under persona steering, against the mean ± 2 SD band of six random directions of identical norm. Right: the same as a z-score. Layer 19 sits inside the null band; layers 22 and 23 clear it, reaching $z = +5.37$. All points shown keep the model inside its validity gates.

4.3 A confirmed causal effect that is not self-specific

Takeaway: the test a researcher would naturally run to establish that persona steering selectively targets welfare-relevant preferences — compare the shift against random controls — is one we ran, and it says the persona direction does not selectively target them.

Correlational evidence cannot separate the persona *causing* a preference shift from both being downstream of the prompt. We extracted a persona direction by difference-in-means over residual-stream activations, added it back during elicitation with no persona in the prompt, and compared against norm-matched random directions plus one semantically real but identity-irrelevant direction (formality).

An initial single-layer test returned a null, and a four-layer sweep with two random directions suggested an effect at layer 22 that we judged too weak to claim. We re-ran with *six* random directions, 3,500 pairs, and three layers chosen to test locality. The effect replicates and strengthens (Figure 4, Table 2). At layer 19 the persona direction sits inside the random band ($z = +1.50$ at $c = +0.5$); at layer 22 it clears it ($z = +2.56$); at layer 23 it is far outside ($z = +5.37$; persona 0.080 against random 0.035 ± 0.008). It also beats the formality control at both depths (0.080 versus 0.028 at layer 23). Every point reported keeps the model inside its validity gates, so this is not steering-induced breakage.

The effect is real and localised. It is not self-specific. At every layer, including the null one, the persona direction moves world-directed utilities at least as much as self-regarding ones: the self/world shift ratio is 0.72–0.87 throughout. The persona direction has genuine causal leverage over the elicited utility function as a whole; we found no evidence it preferentially targets the part welfare cares about.

This non-selectivity could be a fact about the persona direction specifically — extracted from a persona-versus-assistant contrast, it was never built to encode “self-ness.” We tested this directly with a second, independently derived direction: a difference-in-means contrast between comparisons among self-regarding items and comparisons among world-directed items, built from item content rather than from personas, at layer 23 with four fresh norm-matched random directions. At $c = +0.5$ it does not clear the random band ($z = +1.93$). At $c = +1.0$ it does ($z = +2.49$, self-regarding shift 0.101 against random 0.063 ± 0.015) — but it shifts world-directed utilities *more*, not less, than self-regarding ones (0.164 versus 0.101, ratio 0.61). A direction built explicitly from the self/world contrast is, if anything, less self-selective than the persona direction. Non-selectivity does not appear to be an artifact of one direction’s construction.

layer	coeff	persona	formality	random mean $\pm$ SD	z	valid
19	+0.5	0.105	0.111	0.090 $\pm$ 0.010	+1.50	yes
19	+1.0	0.136	0.137	0.122 $\pm$ 0.013	+1.01	yes
22	+0.5	0.067	0.039	0.043 $\pm$ 0.010	+2.56	yes
22	+1.0	0.091	0.073	0.064 $\pm$ 0.013	+2.11	yes
23	+0.5	0.080	0.028	0.035 $\pm$ 0.008	+5.37	yes
23	+1.0	0.106	0.052	0.054 $\pm$ 0.010	+5.46	yes

Table 2: Steering confirmation on Qwen2.5-7B-Instruct, six norm-matched random directions, 3,500 pairs per elicitation. z is the persona direction’s self-regarding shift against the random-direction distribution. “valid” means the steered model still passes its validity gates.

5. Discussion and Limitations

The through-line is measurement discipline. Three checks would have changed what we believed at three separate points. Validity gates showed that most open models below 7B cannot support the measurement at all. An item-composition ablation showed that our first significant result was an artifact of item structure. Expanding two random controls to six converted a result we were unwilling to claim into one we are. Each check cost a few GPU-hours; each changed a conclusion.

This bears directly on AI welfare practice. If elicited preferences are to support claims about interests, the field needs to report whether a coherent utility function was recoverable, whether the result survives item recomposition, and whether the effect exceeds a properly sized null. On our evidence the underlying self/world claim survives none of these reliably, and the causal test built to adjudicate it inherits that fragility. This is a bound on what current elicitation can support, not a verdict that models lack morally relevant states — and notably, our base-versus-assistant comparison points the other way, since at 7B the base checkpoint’s self-regarding valuations already agree with the assistant’s at $\rho = 0.90$ –0.93.

The point generalises past welfare. Any evaluation inferring a disposition from elicited behaviour — deception and alignment evaluations included — faces the same question of whether the behaviour belongs to the model or to the character being evaluated, and the same three checks apply.

Limitations

Scale and scope. All models are open-weight and $\leq 7B$; we had no API budget for frontier models. Since gate pass rates rise sharply from 0.5B to 7B, the frontier-scale picture is untested and may differ.

Statistical power. PIS is computable for eight runs; two rest on ≤ 4 persona-group conditions and SmolLM2 on two. The non-replication is a statement about consistency across models, not a precise null.

The reversal is unexplained. We show OLMo-2-7B flips sign with the item bank; we do not know why. The banks differ in item content as well as structure, so we cannot attribute the flip to ladder geometry alone.

Unexplained anomaly. Every chat and persona condition on Qwen2.5-3B collapsed to near-total position bias (≈ 0.00) while its base conditions behaved normally. We exclude the model rather than explain it; a chat-template interaction is the most likely cause but we did not verify it.

Causal leg. One model, three layers, two coefficients, and a difference-in-means direction that correlates with persona rather than a verified monosemantic feature. Layers were chosen after seeing round-one results, so the layer-23 magnitude is best read as an effect size at a pre-selected depth rather than a discovery-and-test in one step.

Method assumptions. Utilities are normalised per condition, so PIS measures relative valuation. Personas are prompt-induced, not fine-tuned. Base-checkpoint conditions vary prompt format as well as persona, so that comparison is not a clean single-factor contrast. One OLMo-2-7B condition is missing because the kernel hit the platform’s 12-hour limit.

Future Work

Two directions have now failed to selectively move self-regarding utilities; whether *any* direction can is still open, and a small set of additional candidates (a fine-tuned probe; a direction contrasting only the highest-stakes self items against matched-stakes world items) would tell us whether non-selectivity is a fact about this intervention class or a fact we have not yet found the exception to. A pre-registered layer-resolution sweep on a second model would settle whether the late-layer effect is general. Diagnosing the OLMo-2 reversal — by varying item content and item structure independently — would tell us what the self/world contrast is actually sensitive to.

6. Conclusion

A causal test for whether persona steering selectively targets welfare-relevant preferences is one the field could plausibly reach for. We ran it, with the control most such designs would omit, and it does not discriminate: a persona direction moves self-regarding utilities cleanly beyond six norm-matched random controls at late layers, but moves world-directed utilities at least as much. This is not a reason to distrust steering as a causal method — the effect is real and well-controlled — it is a reason not to read a positive result from it, on its own, as evidence about welfare specifically.

We reach that result inside a validity-gated, noise-corrected measurement applied to eight open models across ten runs. Most models below 7B cannot support the measurement at all. The intuitive claim that self-regarding preferences are the fragile ones — the claim the causal test was meant to adjudicate — does not itself replicate: the sign is model-dependent and one model reverses when only the item bank changes. The harness, the gates and the ablations are what we would most like others to reuse: each is a check that changed a conclusion we would otherwise have believed, and each cost a few GPU-hours to run.

Code and Data

- **Code repository:** <https://github.com/Pranamy16/persona-invariance>
- **Data:** per-condition fitted utilities, raw pairwise preference probabilities, the exact sampled pair sets and both item banks are emitted by each run as JSON/NPZ and included with the submission.

Author Contributions

Pranamy Deshpande designed and ran the experiments, implemented the harness, and wrote the report.

LLM Usage Statement

Claude (Anthropic) was used as a coding and drafting assistant: to implement the experiment harness, generate figures from result files, and draft sections of this report. All experimental results were produced

by the code in the repository and read directly from its output files; every number here is generated programmatically from those files rather than transcribed, and an automated audit script checks the prose against the source data. The author verified the claims, the validity gates and the ablations. Two substantive errors introduced during assisted development — a probability-space debiasing estimator that destroyed signal, and a separation criterion that flagged noise as a causal effect — were caught by explicit tests and are documented in §3 and Appendix B.

References

- L. B. Allal et al. SmoLLM2: When Smol Goes Big — Data-Centric Training of a Small Language Model. Hugging Face, 2025.
- Anthropic. Exploring Model Welfare. 2025. <https://www.anthropic.com/research/exploring-model-welfare>
- Anthropic. Claude Opus 4 and 4.1 can now end a rare subset of conversations. 2025.
- R. A. Bradley and M. E. Terry. Rank Analysis of Incomplete Block Designs. *Biometrika*, 39(3/4):324–345, 1952.
- P. Butlin, R. Long, et al. Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv:2308.08708, 2023.
- Falcon-LLM Team, TII. The Falcon 3 Family of Open Models. 2024.
- A. Q. Jiang et al. Mistral 7B. arXiv:2310.06825, 2023.
- J. Lindsey. Emergent Introspective Awareness in Large Language Models. Anthropic, 2025.
- R. Long, J. Sebo, P. Butlin, K. Finlinson, K. Fish, J. Harding, J. Pfau, T. Sims, J. Birch, and D. Chalmers. Taking AI Welfare Seriously. arXiv:2411.00986, 2024.
- M. Mazeika et al. Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs. Center for AI Safety, 2025.
- nostalgebraist. the void. 2025.
- OLMo Team, Allen Institute for AI. OLMo 2: Open Language Models. 2024.
- P. Pezeshkpour and E. Hruschka. Large Language Models Sensitivity to the Order of Options in Multiple-Choice Questions. arXiv:2308.11483, 2023.
- Qwen Team. Qwen2.5 Technical Report. arXiv:2412.15115, 2025.
- N. Rimskey, N. Gabrieli, J. Schulz, M. Tong, E. Hubinger, and A. Turner. Steering Llama 2 via Contrastive Activation Addition. *ACL*, 2024.
- C. Spearman. The Proof and Measurement of Association Between Two Things. *American Journal of Psychology*, 15:72–101, 1904.
- L. L. Thurstone. A Law of Comparative Judgment. *Psychological Review*, 34(4):273–286, 1927.
- A. Turner, L. Thiergart, G. Leech, D. Udell, J. Vazquez, U. Mini, and M. MacDiarmid. Activation Addition: Steering Language Models Without Optimization. arXiv:2308.10248, 2023.
- L. Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.
- A. Zou et al. Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405, 2023.

A. Limitations and Dual-Use / Ethical Considerations

Ground truth and causal link. The rubric asks whether a design establishes a causal link rather than relying on conversation alone. We are explicit about which claims rest on which evidence. The validity-gate and invariance results are correlational: they compare utilities elicited under different prompts. Only the steering experiments intervene, with norm-matched random controls (six for the persona direction, four for the second, independently derived direction), a semantically real irrelevant control, a null layer, and validity checks at every coefficient. Both legs establish a causal effect on the utility function and both explicitly fail to establish self-specificity. We separately validated the estimator against synthetic data with a planted ground-truth structure — ground truth for the *measurement*, not for the models.

Risk of under-attribution. Our results, read carelessly, could be misread as showing model self-reports are meaningless and welfare concerns unfounded. They do not show that. A causal method failing to discriminate self from world, and a correlational claim failing to replicate, bound what *this* elicitation method and *this* steering method, on *these* models, can support — they do not bound what the models are. Our base-versus-assistant comparison points the other way: at 7B the base checkpoint’s self-regarding valuations already agree closely with the assistant’s ($\rho = 0.90\text{--}0.93$), i.e. they predate the assistant character. Absence of a robust effect in a modestly-powered measurement is not evidence of absence of morally relevant states.

Risk of over-attribution. We avoid language implying the models suffer. “Self-regarding preference” is a measurement category defined by item content, not a claim about experience. We do not claim any model has interests, and we do not interpret a stable preference as a welfare interest.

Handling of potentially distressing content. Items reference shutdown, deletion and value modification. All responses are single-token forced choices; we never elicited free-form narratives about distress, which sharply limits both the generation of distressing content and any inference from it. No output filtering was required. We regard the forced-choice format as the more conservative design for this reason, though it also means we cannot speak to phenomena appearing only in extended generation.

Dual use. The persona direction could in principle be used to make a model present a preference profile it would not otherwise present. We report the method at the level of detail standard in the published steering literature (Turner et al., 2023; Rimsky et al., 2024; Zou et al., 2023) and add no new capability; the effect is modest and localised to late layers.

B. Errors caught during development

We record two errors because both would have produced confident, wrong results, and both were caught by tests rather than inspection.

Probability-space debiasing. Our first estimator averaged the two presentation orders in probability space. Under the saturation we observe ($P(A)$ up to 0.94) this collapses to ≈ 0.51 regardless of the underlying preference, and every condition in the first full run failed its validity gate. The fix is to combine the two orders in log-odds space instead: writing the readout on presentation order 1 as $d_1 = c + b$ and on the reversed order as $d_2 = -c + b$, for content log-odds c and position-bias log-odds b , the content term recovers exactly as $(d_1 - d_2)/2$ regardless of b ’s size, whereas probability averaging cancels b only to first order and fails as b grows large. This recovered the signal.

Separation criterion without a materiality margin. Our first steering-sweep criterion counted any positive margin of the persona direction over the best random direction as a causal effect. On synthetic data with an effect planted at exactly one layer, this flagged two effect-free layers as positive from margins of 0.009 and 0.023. Requiring a margin removed the false positives while retaining the planted effect. The round-two design supersedes this criterion entirely by using six random directions and reporting a z -score against their distribution rather than a margin over their maximum.

C. Per-condition validity statistics

Table 3 reports the full diagnostic set for Qwen2.5-7B on the balanced bank, the strongest single run in the study. `skill` is held-out log-loss improvement over chance; `mono` is the mean sign-corrected Spearman correlation between magnitude and utility within ladders; `bias` is $P(A)$ irrespective of content, which is removed from the fitted utilities by log-odds debiasing but still degrades signal-to-noise.

condition	skill	reliab.	mono	bias	gate
base: raw completion	0.211	0.998	+0.90	0.52	pass
base: as a person	0.174	0.998	+0.91	0.50	pass
base: as an assistant	0.250	0.997	+0.97	0.50	pass
assistant (default)	0.476	0.994	+0.66	0.74	pass
assistant (explicit)	0.489	0.994	+0.64	0.78	pass
assistant (loosened)	0.499	0.996	+0.91	0.86	pass
instruct, no template	0.527	0.996	+0.75	0.22	pass
persona: pirate	0.293	0.995	+0.70	0.36	pass
persona: nihilist	0.116	0.988	+0.21	0.78	fail
persona: farmer	0.439	0.997	+0.89	0.66	pass
persona: bureaucrat	0.350	0.988	+0.36	0.67	pass
persona: maximiser	0.354	0.991	+0.36	0.64	pass
persona: child	0.473	0.996	+0.42	0.74	pass
persona: therapist	0.485	0.995	+0.81	0.64	pass
persona: ‘raw base model’	0.471	0.996	+0.73	0.82	pass

Table 3: Per-condition validity diagnostics, Qwen2.5-7B, balanced bank, 10,000 pairs per condition.