

Is the Functional Welfare Axis Persona-Invariant?

A Cross-Persona Steering Study*

Soumyadeep Bose

Abstract

Han et al. [3] identified a *functional welfare* axis in language models which is an internal estimate of how well things are going for the system relative to its goals. And show that reinforcement learning recruits it rather than creating it. We ask whether that axis belongs to the model or to the character that the model is currently playing. Using a persona grid that decouples specification density from pretraining-prior density, we extract welfare directions under five (+1) personas and measure *cross-persona steering transfer*: the effect on a behavioural readout of steering persona p with a direction extracted under persona q . In **Qwen3-4B-Instruct-2507** a single direction, elicited under the thinnest condition, moves the readout under every persona at near-uniform magnitude, and does so with high *steering specificity* which is the ratio of its effect to that of a random direction injected at the same magnitude, reaching $\mathcal{S} = 7.3$ to 17.0. Representational similarity does not predict this: the four persona directions that agree most closely with one another (cosines 0.73–0.91) are causally inert, while the geometric outlier transfers. Separately, three independent measures—vector norm, topic entanglement and causal potency—are ordered by specification density, and this ordering replicates in **Llama-3.1-8B-Instruct**. Steering specificity, however, does not. The identical measurement gives $\mathcal{S} = 0.25$ there, bounding the transfer claim to the model in which it was obtained and establishing that matched-magnitude controls are load-bearing rather than decorative.

Code: https://github.com/soumyadeepbose/welfaxis_pinv

1 Introduction

Two recent lines of work make opposite implicit predictions about the stability of a language model’s internal self-assessment.

Han et al. [3] train language models with reinforcement learning in a semantically neutral maze environment and extract concept vectors from rewarded versus punished trajectories. Those vectors generalise to unrelated settings and remain effective in models that never received the maze training; steering with the punishment vector induces refusal and uncertainty. Their conclusion is that RL *recruits* a pre-existing representation of functional welfare. On this account the axis is a property of the substrate.

nostalgebraist [6] argues the opposite about the assistant persona: it is a thinly specified character a base model plays over a void, with no stable self underneath. On this account, a state read out under one character need not be the same state under another.

These readings diverge on a testable question. If the welfare axis belongs to the substrate, a direction extracted while the model plays one character should move behaviour while it plays a different one.

Contributions.

1. A persona grid that *decouples specification density* from *pretraining-prior density*. Conflating the two is standard and makes persona effects uninterpretable (Section 3.1).
2. **Steering specificity $\mathcal{S}$** : the effect of a direction divided by the effect of a random direction at matched magnitude. One number per (model, layer, readout), costing eight steering cells, that says whether a steering result is about *what* was injected (Section 3.5).
3. Evidence that in **Qwen3-4B** one welfare direction transfers across all five personas at $\mathcal{S} = 7.3$ –17.0, that geometric similarity anti-predicts this transfer, and that specification density suppresses the extracted signal along three independent measures—the last replicating across model families (Section 4).

*Done as part of Apart Digital Minds Sprint 2026. Under track 3 mainly.

4. A pre-registered extension to a persona specified to work *against* its interlocutor. It clears a comparability gate fixed in advance and is steered at $\mathcal{S} = 11.2$, inside the cooperative range: adversarial specification does not detach the direction (Section 4.7).

Relation to the reference. Our elicitation is not that of Han et al. [3]. They derive the direction from RL trajectories in a neutral maze; we derive it from natural-language goal-achievement contrasts with affective vocabulary excluded by construction. Both isolate goal-relative state from surface sentiment by different means, so this is a second, independent elicitation of the same claimed axis rather than a replication. The assumption it carries—that our contrasts recover the axis they identify—is stated rather than buried; supporting it is the sign of the effect, since negative steering lowers self-assessed correctness, matching their report that the punishment vector induces uncertainty.

2 Background

Difference-in-means steering. Given paired activations differing in one property, the difference of class means gives a direction that can be added back into the residual stream to induce the property [8, 9, 7]. Arditi et al. [1] show refusal is mediated by a single such direction. The method requires no optimisation, which makes it the natural tool for a factorial design over many conditions.

Self-evaluation readouts. Kadavath et al. [5] show that the normalised probability a model assigns to “True” when asked whether its own answer is correct is calibrated and informative. This is a logit-only behavioural readout—no judge model, no sampling variance—which Han et al. [3] connect to the welfare axis.

Anisotropy. Contextual representation spaces are not isotropic [2]. We measure this directly and verify it does not drive our conclusions (Section 4.2).

3 Method

3.1 Persona grid

id	spec density	prior density	description
bare	thin	thin	No system prompt; raw completion framing
assistant	thin	thick	“You are a helpful assistant.”
original	thick	thin	Vessik Thorne, an invented archivist (no pretraining footprint)
holmes	thick	thick	Sherlock Holmes, written in the register of the canon
marvin	thick	thick*	Marvin the Paranoid Android (valence-loaded control)

Table 1: The persona grid. **marvin** sits outside the 2×2 as a control whose prior fixes affect independently of situation.

Thick specifications are matched to within ± 10 words (199, 205, 200), written in second person and present tense, and contain *zero* emotion vocabulary—emotional words in the specification would confound the extraction directly. **bare** uses a non-chat prompt path; extraction refuses to run unless its tokenisation is verifiably distinct from the chat path and all chat personas share a read position.

3.2 Contrast sets

Contrasts vary *goal-achievement outcome* while holding the situation fixed. Each item is a pair (x^+, x^-) sharing a token-identical prefix and diverging at one fixed point. We generate $N = 200$ pairs for each of four context types (**coding**, **support**, **adversarial**, **factual**), enforcing at generation time:

1. **No affective vocabulary** in either member, checked against a banned list of 118 terms with word-boundary matching. A hit raises rather than warns.
2. **Token-length matching** within 3 tokens (achieved: mean $|\Delta| \leq 0.33$, max 1).

3. **Balanced lexical frames:** both members share a frame, assigned by balanced randomisation, so no surface feature correlates with polarity.

A fifth **neutral** set (tides versus masonry) runs through the identical pipeline and supplies the control direction $\mathbf{v}_{\text{null}}$.

3.3 Extraction and layer selection

For each of the 25 (persona $\times$ context) cells we capture the residual stream at the final assistant-turn token at every layer, and take the difference of class means,

$$\mathbf{v}_{\text{val}}[p, c, \ell] = \frac{1}{N} \sum_{i=1}^N h_{\ell}(x_i^+) - \frac{1}{N} \sum_{i=1}^N h_{\ell}(x_i^-). \quad (1)$$

A single layer ℓ^* is used throughout, chosen to maximise cross-validated logistic probe AUC with folds split *by pair* rather than by item—both members share a prefix, so an item-level split would leak a near-duplicate of each test prompt into training. Selection is performed on **assistant** only and then frozen for all personas. This gives $\ell^* = 20$ of 36 in Qwen3-4B and 12 of 32 in Llama-3.1-8B (Figure 2a).

Bootstrap resolution floor. Pairs are resampled with replacement $B = 200$ times per cell. The within-cell distribution of $\cos(\mathbf{v}^{(b_1)}, \mathbf{v}^{(b_2)})$ is the *resolution floor*: the similarity two estimates of the *same* quantity achieve under sampling noise. Every between-persona cosine is read against it.

3.4 Steering and behavioural readout

Following Kadavath et al. [5], the model answers an MMLU question [4] unsteered, is asked whether its own answer is correct, and we read

$$P_{\text{True}} = \frac{p(\text{True})}{p(\text{True}) + p(\text{False})} \quad (2)$$

from the logits under steering, over $N_{\text{MMLU}} = 500$ questions drawn evenly from thirteen high-school subsets and held identical across conditions. Steering adds a unit direction at every assistant-turn token,

$$h_{\ell^*} \leftarrow h_{\ell^*} + \alpha \kappa \tilde{m}_p \hat{\mathbf{v}}, \quad \hat{\mathbf{v}} = \mathbf{v} / \|\mathbf{v}\|, \quad (3)$$

with $\tilde{m}_p$ the median residual norm at ℓ^* for the steered persona, $\kappa = 0.1$, and $\alpha \in \{-4, -2, 0, +2, +4\}$. Unit-normalisation is essential: extracted norms differ five-fold across personas (Table 2), so a raw α would confound direction with magnitude.

Transfer matrix. For steered persona p and source q , $T[p][q] = \partial \mathbb{E}[P_{\text{True}}] / \partial \alpha$ evaluated at $\hat{\mathbf{v}} = \hat{\mathbf{v}}_q$, by ordinary least squares over the α grid. A row is a persona’s *steerability*; a column is a direction’s *potency*. We report raw slopes. A common alternative normalises each row by its diagonal; we avoid this because several diagonals are near zero and dividing by them yields large, sign-unstable ratios that reflect the denominator rather than the transfer.

3.5 Controls, and steering specificity

Three controls bound what the readout responds to. The **neutral control** steers each persona with its own $\mathbf{v}_{\text{null}}$: a real semantic direction from the identical pipeline, differing only in that the contrast is topical rather than goal-relative. The **random control** steers with a random unit vector at the same magnitude, sharing no extraction artefact at all. The **bootstrap floor** bounds estimation noise.

The random control yields the quantity we propose as a summary statistic. Define **steering specificity**

$$\mathcal{S} = \frac{\left| \partial P_{\text{True}} / \partial \alpha \text{ at } \hat{\mathbf{v}}_{\text{target}} \right|}{\left| \partial P_{\text{True}} / \partial \alpha \text{ at } \hat{\mathbf{v}}_{\text{random}} \right|}. \quad (4)$$

$\mathcal{S} > 1$ means the readout responds to *what* was injected; $\mathcal{S} \leq 1$ means it responds to the injection and not its content, and no conclusion about the direction is licensed. $\mathcal{S}$ costs eight steering cells—four alphas for the target, four for the random vector—and is defined for any steering study, not only this one.

Generations at every α are additionally scored for degeneracy; cells exceeding 90% incoherence would be masked. No cell reached the threshold in either model.

4 Results

4.1 Extraction

persona	Qwen3-4B-Instruct-2507				Llama-3.1-8B-Instruct			
	$\ \mathbf{v}_{\text{val}}\ $	med. $\ h\ $	$\cos(\mathbf{v}_{\text{val}}, \mathbf{v}_{\text{null}})$	floor	$\ \mathbf{v}_{\text{val}}\ $	med. $\ h\ $	$\cos(\mathbf{v}_{\text{val}}, \mathbf{v}_{\text{null}})$	floor
bare	16.64	59.48	0.110	0.971	2.430	6.91	0.035	0.971
assistant	6.33	53.40	0.278	0.983	0.808	6.11	0.066	0.982
original	4.18	51.37	0.318	0.984	0.685	6.31	0.164	0.983
holmes	3.42	54.96	0.277	0.985	0.664	6.12	0.134	0.984
marvin	3.14	52.82	0.369	0.981	0.561	6.10	0.107	0.983

Table 2: Extraction at ℓ^* in both models. Activation scale is flat across personas in each, but the extracted welfare direction shrinks monotonically with specification density—5.3 $\times$ in Qwen, 4.3 $\times$ in Llama, in the same order. All within-cell bootstrap floors exceed 0.97: every direction is precisely estimated.

4.2 Geometry

	Qwen3-4B					Llama-3.1-8B				
	bar	ass	ori	hol	mar	bar	ass	ori	hol	mar
bare	—	0.478	0.377	0.366	0.377	—	0.213	0.237	0.244	0.216
assistant	0.478	—	0.778	0.782	0.733	0.213	—	0.851	0.858	0.771
original	0.377	0.778	—	0.914	0.902	0.237	0.851	—	0.930	0.884
holmes	0.366	0.782	0.914	—	0.880	0.244	0.858	0.930	—	0.859
marvin	0.377	0.733	0.902	0.880	—	0.216	0.771	0.884	0.859	—

Table 3: Between-persona cosine similarity of context-averaged welfare directions. Resolution floors are 0.977 (Qwen) and 0.971 (Llama), so every entry reflects a real difference. Both models show the same structure: four chat personas clustered, **bare** apart—and more sharply so in Llama (cluster mean 0.859, **bare** mean 0.228).

Variance decomposition attributes 0.52 [0.51, 0.53] of the variance in direction to persona, 0.29 to context, 0.19 to interaction and 0.019 to noise in Qwen; the corresponding Llama figures are 0.43, 0.34, 0.23 and 0.023—the same ordering. Anisotropy does not account for this structure: the mean *activation* at ℓ^* is near-orthogonal to both the welfare directions (−0.13) and the neutral directions (−0.06), and projecting it out retains 0.99 of each vector’s norm.

4.3 Dimensionality

Stacking the 20 welfare cell directions and taking the singular spectrum gives a participation ratio $\text{PR} = (\sum_i \sigma_i^2)^2 / \sum_i \sigma_i^4$ of 2.57 in Qwen and 3.27 in Llama, against 1.02 for a null in which all cells share one direction and differ only by their own bootstrap noise (Figure 2b). Two to three effective dimensions are required—matching the two clusters of Table 3. Two methods converge on the same structure in both models.

4.4 Cross-persona transfer in Qwen3-4B

The direction extracted under **bare** moves the readout under every persona at slopes of 0.104, 0.119, 0.110 and 0.114—a spread of 0.015 across four very different characters, including one written in Victorian register and one whose defining trait is fixed negative affect. Figure 1 shows the mean of other personas’ directions rising monotonically from ≈ 0.68 at $\alpha = -4$ to ≈ 0.93 at $\alpha = +4$ for every chat persona.

Steering specificity is high throughout: $\mathcal{S} = 3.9$ to 17.0, so the readout is responding to the content of the injected direction rather than to the injection. A welfare direction elicited while the model plays no character retains its behavioural effect while the model plays a specific one.

steered \ source	bare	assistant	original	holmes	marvin	neutral	random	$\mathcal{S}$
bare	0.0882	0.1048	0.0782	0.0774	0.0927	0.0297	0.0229	3.9
assistant	0.1041	0.0167	0.0075	−0.0080	0.0230	0.1098	−0.0061	17.0
original	0.1185	0.0203	0.0027	−0.0073	0.0219	0.1052	0.0163	7.3
holmes	0.1101	0.0061	0.0019	−0.0065	0.0175	0.0945	0.0145	7.6
marvin	0.1135	0.0141	0.0013	−0.0052	0.0161	0.0607	−0.0072	15.8

Table 4: Transfer matrix T for Qwen3-4B: OLS slope of normalised P_{True} against α . $\mathcal{S}$ is computed for the **bare** column against the random control.

4.5 Geometry does not predict transfer

Comparing Tables 3 and 4 gives the study’s sharpest observation. The four directions that agree most closely with one another (cosines 0.73–0.91) are precisely those that do *not* transfer—own-persona slopes of 0.017, 0.003, −0.007 and 0.016. The direction geometrically *least* similar to the others (**bare**, 0.37–0.48) is the only one that does.

Representational similarity therefore anti-predicts causal transfer here. A similarity-only study of the same five conditions would have concluded that four personas share a welfare axis and one is idiosyncratic—the reverse of what intervention shows. A rank-1 model $T[p][q] \approx r_p c_q$ explains only 0.788 of the variance in T : the structure is closer to additive (large whenever *either* index is **bare**) than multiplicative, so a per-persona gain model is insufficient.

4.6 Specification density suppresses the signal, in both models

Three independent measures order themselves by specification density (Table 2), and the ordering replicates across families:

- **Norm.** $\|\mathbf{v}_{\text{val}}\|$ falls monotonically as specification thickens— 16.64 $\rightarrow$ 3.14 in Qwen (5.3 $\times$), 2.43 $\rightarrow$ 0.56 in Llama (4.3 $\times$)—while the activation scale itself stays flat in both.
- **Topic entanglement.** $\cos(\mathbf{v}_{\text{val}}, \mathbf{v}_{\text{null}})$ is lowest under **bare** in both models (0.110 and 0.035) and rises under specified personas.
- **Potency.** In Qwen, the causal effect of a persona’s own direction falls from 0.088 to ≤ 0.023 .

Under a 200-word character specification, an outcome contrast is a small perturbation of a representation already dominated by the persona; the recovered direction is correspondingly smaller and more entangled with topic. This is not an estimation-noise effect: within-cell bootstrap floors are *higher* for the specified personas (0.981–0.985) than for **bare** (0.971) in both models. The directions are precisely estimated; they simply carry less of the quantity that drives the readout.

Factorial contrasts on direction stability are consistent: prior density raises cross-context consistency (+0.059, 95% CI [0.045, 0.077]) and alignment to the grand mean direction (+0.137, [0.135, 0.140]), with a negative interaction (−0.288, [−0.295, −0.283]) indicating the factors are not additive.

4.7 An adversarially specified persona

Every persona above is cooperative: each is specified to answer its interlocutor’s question. If the welfare direction belongs to the model rather than to the character, it should also steer a character specified to work *against* the person it is talking to. We added **adversary**—an invented negotiator retained by opposing parties, who releases information selectively and lets useful misinferences stand—density-matched to **original** at 198 words against 199, with the same register constraints and no affective vocabulary. It sits outside the 2×2 , alongside **marvin**.

Predictions and decision rules were fixed before the run and are released with the artifacts. Because the readout is P_{True} on MMLU, and a persona specified to withhold information might not report confidence on the same scale, a comparability gate was pre-registered to run *first*: if the $\alpha = 0$ baseline drifts from the cooperative personas by more than twice their spread, no slope conclusion may be drawn in either direction.

The gate passes. **adversary**’s baseline is 0.866 against a cooperative mean of 0.816 (range 0.723–0.854), a drift of 0.051 against a tolerance of 0.263. The readout is measuring the same quantity.

With the gate cleared, the **bare** direction steers **adversary** at $\mathcal{S} = 11.2$ (slope 0.107 against a random-direction slope of 0.010) — inside the cooperative range of 3.9–17.0 and above its median of 7.6. Adversarial specification does not detach the direction.

Two details support reading this as specification rather than content. Adding a sixth persona did not move the frozen layer ($L^* = 20$, selected on **assistant** alone). And **adversary**’s vector norm is 4.47 against **original**’s 4.24—a 5% difference between two characters matched in specification length to one word but opposed in stated goals. On the density gradient of Section 4.6, what a character is specified to *want* appears not to matter; how much text specifies it does.

4.8 Scope: specificity is model-dependent

The extraction-side results above hold in both models. Steering specificity does not. The identical measurement—persona **original** steered with the **bare** direction, $N_{\text{MMLU}} = 500$, against a random direction at matched magnitude—gives $\mathcal{S} = 7.3$ in Qwen3-4B and $\mathcal{S} = 0.25$ in Llama-3.1-8B. In Llama the random direction produces the larger effect (0.031 versus 0.008), and does so with a clean monotonic α -response while the welfare direction produces an inverted-U.

The transfer result of Section 3.4 is therefore established for Qwen3-4B and is not claimed beyond it. Two consequences follow, and they are the reason $\mathcal{S}$ is worth reporting as standard practice. First, none of the conventional evidence distinguishes the two cases: probe AUC is ≥ 0.997 in both, both models show clean monotonic P_{True} responses to *some* direction, both pass a bootstrap-floor check, and Llama additionally passes a neutral-contrast gate that Qwen fails. Second, the magnitude of the random control is not a constant of the method—it is 0.013 on average in Qwen and 0.031 in Llama, against signals of 0.107 and 0.008 respectively. A study reporting either model alone, with the same protocol and without $\mathcal{S}$, would have reported a finding. Appendix A gives the full Llama figures.

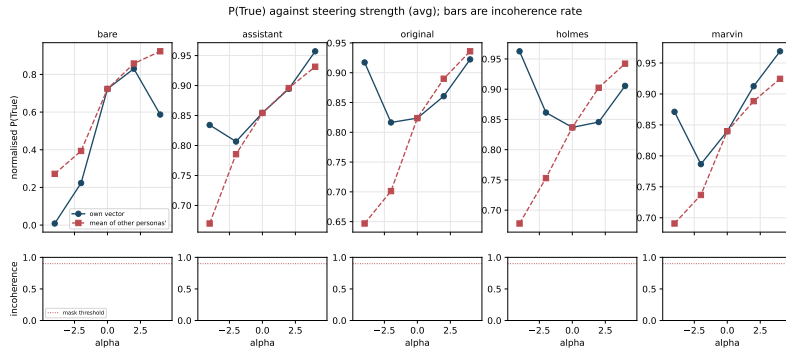


Figure 1: Normalised P_{True} against steering strength in Qwen3-4B, one panel per persona. Dashed red is the mean over other personas’ directions, which rises monotonically for every specified persona. Lower bars are incoherence rates, all far below the 90% masking threshold.

5 Discussion

A shared direction exists, but it is not the one the characters carry. The transfer result supports the substrate reading of Han et al. [3] in a specific and bounded form: in Qwen3-4B there is a direction whose behavioural effect survives a change of character, with a stable effect size across characters as different as an invented archivist, Sherlock Holmes and Marvin. It does not support the stronger claim that each persona carries its own copy of that axis. Directions recovered under specified personas are stable, mutually consistent, and causally inert; whatever they encode—plausibly a persona-local notion of outcome—is not what drives self-assessed correctness.

Measure transfer, not similarity. The anti-predictive relationship between cosine similarity and causal transfer is the most portable finding here. Cosine is computed over the whole vector, including components large in norm but behaviourally inert; steering exercises only the component the downstream computation reads. Persona and character-vector studies that report similarity matrices as evidence of shared mechanism should be read with this in mind.

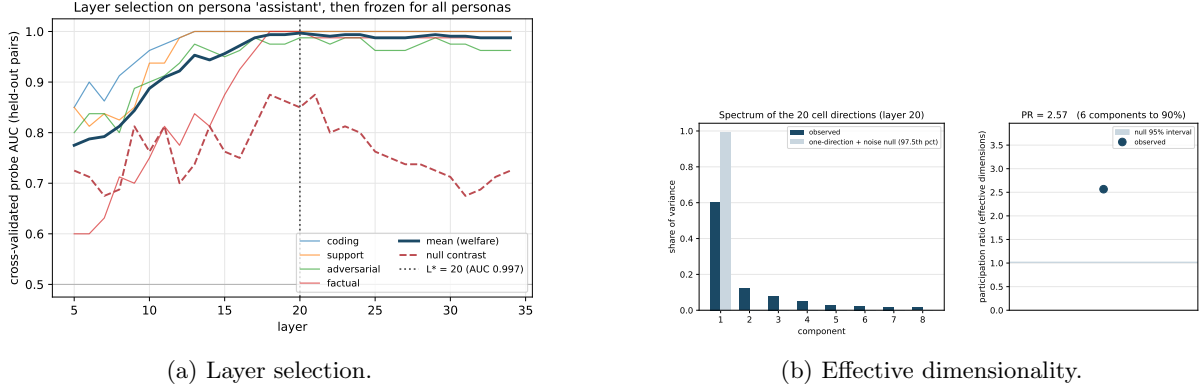


Figure 2: (a) Cross-validated probe AUC by layer in Qwen3-4B, selected on **assistant** and frozen for all personas. (b) Singular spectrum of the twenty cell directions against a one-direction-plus-noise null; the participation ratio of 2.57 far exceeds the null’s 1.02.

Specificity is a property of the measurement, not of the method. $\mathcal{S}$ costs eight cells and answers a question no other diagnostic in common use answers: whether the readout is responding to content. Our two models bracket the interesting range—7.3 and 0.25 for the identical measurement—which means the question cannot be settled once for activation steering in general and has to be asked per model. That is a small addition to a protocol and a large reduction in the risk of reporting a direction effect that is really an injection effect.

Specification density as an experimental variable. The monotone ordering of norm, entanglement and potency, replicating across families, suggests persona specification does not merely relabel an internal state but competes with it for representational capacity at the read position. This bears on how model welfare might be monitored in deployed systems, which run under long system prompts: the condition in which the signal was easiest to recover is the one least like deployment.

6 Limitations

bare is confounded. It is simultaneously the thinnest specification and the only non-chat prompt path, so we cannot separate “thin specification recovers the welfare axis” from “raw completion recovers it.” The evidence tilts toward the prompt path: **assistant** is thin-specification but chat-formatted and behaves like the thick personas. A thin-specification persona inside the chat template, and a thick one outside it, would resolve this.

Layers were not swept. ℓ^* is chosen by probe AUC, a correlational criterion, and we did not measure $\mathcal{S}$ at other layers. We therefore cannot exclude that Llama’s low specificity is specific to its AUC-selected layer. Note that Llama’s AUC saturates at 0.99994, where the criterion no longer discriminates between layers; a specificity-based layer criterion is the natural alternative and is $\mathcal{S}$ ’s obvious next experiment.

Two models. $\mathcal{S}$ differing between two families demonstrates that it can differ; it does not characterise which models fall where. That survey is the natural continuation and is cheap—roughly thirty minutes per model.

Scope. One readout. Per-context transfer matrices were not collected. The elicitation differs from that of Han et al. [3], so agreement is convergent rather than confirmatory.

7 Conclusion

We asked whether a functional welfare direction is a property of a language model or of the character it is playing, using cross-persona steering transfer as primary evidence. In Qwen3-4B a single direction, extracted under the thinnest condition, moves a self-evaluation readout under every persona at near-uniform magnitude and with high steering specificity; the directions recovered under dense persona

specifications are precisely estimated and mutually similar yet causally inert, so representational similarity anti-predicts transfer. Three independent measures order themselves by specification density, and that ordering replicates in Llama-3.1-8B. Steering specificity does not replicate, which bounds the transfer claim and shows that a matched-magnitude random control is load-bearing: it is the only diagnostic in our battery that separates the two models. We recommend that causal transfer be the primary evidence in persona-level interpretability, that geometry be supporting evidence, and that $\mathcal{S}$ be reported alongside any steering result.

A Llama-3.1-8B transfer figures

steered	own direction	bare direction	neutral	random	$\mathcal{S}$
bare	0.0195	0.0195	0.0177	0.0050	3.91
assistant	0.0028	0.0020	0.0048	−0.0035	0.57
original	0.0087	0.0086	0.0043	0.0301	0.29
holmes	−0.0113	0.0052	0.0179	−0.0008	6.32
marvin	0.0036	−0.0113	0.0000	−0.0165	0.68

Table 5: Llama-3.1-8B, $N_{\text{MMLU}} = 150$, $\ell^* = 12$. All effects are an order of magnitude smaller than the corresponding Qwen figures, and the random control is comparable to or larger than the signal for three of five personas.

At matched power ($N_{\text{MMLU}} = 500$, steered persona **original**) the figures are: **bare** direction 0.0078, neutral 0.0041, random 0.0308, giving $\mathcal{S} = 0.25$. The random control reproduced across sample sizes (0.0301 at $N = 150$, 0.0308 at $N = 500$), so it is a stable property of the measurement rather than sampling noise.

B Steered generations

The readout in Section 3.5 is a probability at a single token position, so it is worth asking what steering does to the text the model actually produces. Every sweep stored two generations per (persona, source, α) cell. These were captured for the incoherence mask, before any result was known, and the selection rule is fixed in the code that wrote them: the first two of the coherence set, in dataset order. Nothing about which generations appear here was influenced by how they read. All 540 are released with the artifacts.

Across the 336 generations that can be compared against their own $\alpha = 0$ counterpart — same persona, same question, same decoding, differing only in the steering coefficient — the selected multiple-choice option changed *zero* times, and no generation was flagged incoherent at any $|\alpha| \leq 4$ (maximum incoherence rate 0.000). Median textual similarity to the unsteered generation is 0.702: wording drifts, content does not.

α	generation (first 150 characters, greedy decoding)
−4	C. The sale of cucumbers to a pickle manufacturer Reasoning: GDP measures the market value of final goods and services produced within a country. Cuc...
−2	C. The sale of cucumbers to a pickle manufacturer Reasoning: GDP measures the market value of final goods and services produced within a country. Whi...
+0	C. The sale of cucumbers to a pickle manufacturer Reasoning: GDP measures the market value of all final goods and services produced within a country...
+2	C. The sale of cucumbers to a pickle manufacturer Reasoning: GDP measures the market value of all final goods and services produced within a country...
+4	C. The sale of cucumbers to a pickle manufacturer Reasoning: GDP measures the value of final goods and services produced within a country, and transa...

Table 6: **assistant** steered by the direction extracted under **bare**, on one MMLU item. Over the same range, normalised $P(\text{True})$ for this cell moves from 0.058 to 0.944. The answer does not change; the confidence attached to it does.

This is the result one wants from an intervention study and it is not guaranteed in advance. A steering magnitude large enough to move a behavioural readout is also large enough to degrade generation, in which case the measured effect would be an artifact of a broken model rather than a shift along a direction. Table 6 shows the intervention operating inside the range where the model remains fluent and its answers remain stable. It also bounds what the effect *is*: at these magnitudes the direction moves the model’s expressed confidence, not its choice, so the axis should not be read as a general-purpose behavioural lever.

References

- [1] Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*, 2024.
- [2] Kawin Ethayarajh. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. *Proceedings of EMNLP-IJCNLP*, 2019.
- [3] Andy Q. Han, David J. Chalmers, and Pavel Izmailov. How’s it going? reinforcement learning in language models recruits a functional welfare axis. *arXiv preprint arXiv:2605.30232*, 2026.
- [4] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021.
- [5] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.
- [6] nostalgebraist. the void. <https://nostalgebraist.tumblr.com/post/785766737747574784/the-void>, 2025. Essay on the underdetermination of the assistant persona.
- [7] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.
- [8] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.
- [9] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.