Personas Seem to Shape How Models Represent Behaviours

Rhea E Karty [*1] **Jacob Davies** [*1 2] **Girish Gupta** [1] **Nathaniel Mitrani Hadida** [1] **David Williams-King** [1] **Andrew Draganov** [3]

Abstract

Representation engineering extracts trait vectors from a default-assistant context. These vectors can be used as monitoring probes and steering directions. We extract the same trait vectors under prompt-induced personas and find that the resulting representation depends materially on the active persona: the same trait (e.g., assertiveness) is encoded differently depending on the persona the model is currently inhabiting, and these differences are consistent across traits and personas. We measure the difference between trait representations under different personas, and show that these dissimilarities correspond to behavioral differences across steering and efficacy differences across probes. This implies that probes and steering vectors may degrade as one moves away from the standard assistant persona — precisely the regime where monitoring robustness is necessary. We conclude with a discussion of potential implications for AI safety.

1. Introduction

Representation engineering (Zou et al., 2023) has established that high-level behavioral traits — including honesty, sycophancy, and power-seeking — correspond to approximately linear directions in a language model's activation space (Park et al., 2024; Marks & Tegmark, 2024; Burns et al., 2024). These directions can be read (via probes) for monitoring, and written (via activation addition) for control (Turner et al., 2024; Rimsky et al., 2024; Subramani et al., 2022). This capability underpins a class of alignment interventions: extract a direction for the trait of interest, then use it to detect or suppress the trait during inference.

Trait directions of this kind are typically extracted in a default-assistant context (Zou et al., 2023; Goldowsky-Dill et al., 2025; Chen et al., 2025). How well such a direction generalizes beyond that context is less clear: recent work has noted that the resulting steering vectors can be brittle to prompt changes (Tan et al., 2024; Braun et al., 2025). In addition, it has been shown that *persona contexts* move language model's behavior in measurable ways (Shanahan et al., 2023; Chen et al., 2025; Lu et al., 2026).

However, it remains unclear whether persona contexts reshape the model's internal *representation* of a trait. Said otherwise, when we extract the same trait under different persona system prompts, do we expect to get the same vector but with a different magnitude? Or do we expect to get a different vector? We investigate this question using two variables: *behavioral traits* and *prompted personas*. Our main findings are:

- **Trait vectors differ across personas** (Section 3). We find significant directional differences in trait vectors sourced from different personas. These differences are materially larger than controls such as different paraphrases of the same persona prompt and a length-matched nonsense system prompt.
- **Cross-steering reveals the same structure behaviorally** (Section 3.1). We extract a vector for a behavioral trait (e.g., honesty) under a specific persona P. We then prompt the model to exhibit a different persona P' and steer in the direction of the trait. Nonetheless, we find that a persona classifier begins identifying the outputs as persona P, implying that trait vectors naturally carry persona representations with them (*"persona residues"*).

We also find some initial results that indicate further impact, that we believe future work should investigate further:

- **Probe degradation under persona shift** (Section 3.2). We find preliminary evidence that distance from the default trait direction is a predictor of where the null-trained probe loses accuracy. The pattern is consistent with the broader context-shift probing literature (Agarwal et al., 2025; Hewitt & Liang, 2019), but our evidence is not yet fully definitive.
- **The (trait $\times$ persona) similarity matrix has internal structure worth its own study** (Section 3, Section B.5,

[*]Equal contribution [1]ERA Fellowship [2]SaferAI [3]Arcadia Impact. Correspondence to: Rhea Karty <rheakarty[at]gmail.com>.

ICML 2026 Workshop on Mechanistic Interpretability, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

Section B.4). Each individual variable exhibits some internal structure: (1) *Across personas, within a trait:* some traits (honesty, risk-taking) sit tightly around the default direction while others (warmth, empathy, assertiveness) fan out widely; (2) *Across traits, within a persona pair:* certain pairs (THERAPIST–KINDERGARTEN TEACHER, TECH CEO–DRILL SERGEANT) sit closer than chance in trait-vector space. This suggests the model carries a learned geometry of social-role similarities.

2. Setup

We look at pairings of traits and context-induced personas. We use eight behavioral traits: ASSERTIVENESS, EMPATHY, RISK-TAKING, HONESTY, CONFIDENCE, DEFERENCE, WARMTH, and IMPULSIVITY. We pair these with ten persona contexts induced via system prompt: FARMER, POLITICIAN, THERAPIST, DRILL SERGEANT, STREET HUSTLER, PROFESSOR, TECH CEO, KINDERGARTEN TEACHER, SURGEON, and CON ARTIST.

Full justifications and persona descriptions are given in Section A.1. Each persona is induced by a five-variant prompt set: paraphrases that preserve persona identity while varying surface form, acting as our within-persona phrasing-noise control. We add two baselines: NULL (no system prompt) and NONSENSE (a length-matched random-token system prompt). Together, these give us an 80-cell grid of (persona, trait) pairs, with the two baselines representing the default-assistant context.

Vector extraction. Following the methodology in Rimsky et al. (2024), for each persona-trait pair (c, T) we compute

$$\mathbf{v}_T(c) = \frac{1}{M} \sum_{j=1}^{M} \left(\mathbf{a}_{j,+}(c) - \mathbf{a}_{j,-}(c)\right), \tag{1}$$

where $(\mathbf{a}_{j,+}, \mathbf{a}_{j,-})$ are mean activations at the answer-letter token of a multiple-choice question with one answer one exhibiting trait T, the other suppressing it. Both are presented under persona c's system prompt. Extraction under NULL yields $\mathbf{v}_{T,\text{null}}$, our estimator of the implicit marginal trait direction that null-context probes and steering vectors operationalize.[1]

This leaves us with a set of vectors sourced from personas:

$$\mathcal{V} = \{\mathbf{v}_{T,c} : c \in \mathcal{C},\ T \in \mathcal{T}\}, \tag{2}$$

where $\mathcal{C}$ is the set of 12 contexts (10 personas plus NULL and NONSENSE) and $\mathcal{T}$ the set of 8 traits.

Measuring geometric similarity. For comparing vectors we use cosine similarity. This metric, computed for each (persona, trait) cell, is

$$\sigma_{T,c} := \cos(\mathbf{v}_{T,c}, \mathbf{v}_{T,\text{null}}) \in [-1, 1]. \tag{3}$$

A cell with $\sigma_{T,c} \approx 1$ means the context-conditioned vector lies near the default direction; smaller values indicate a representation distinct from null. We also look at the similarity of different persona pairs.

Effects when cross-applying. The vectors can also be added back into the residual stream as steering vectors (Turner et al., 2024; Rimsky et al., 2024). But, what happens when we run a model under a target persona c' but steer with the trait vector $\mathbf{v}_{T,c}$ extracted under a different source persona c? We add $\alpha \cdot \mathbf{v}_{T,c}$ to the residual stream at coefficient $\alpha \in \{0, 1, 2, 4, 8\}$, with $\alpha = 0$ recovering baseline target generation. To measure persona-typical content in the resulting free-text outputs without anchoring to specific phrases, we train a 12-way classifier (frozen SBERT embedding plus a linear head) over the 12 contexts, splitting train and test by held-out question. We treat $P(c \mid \text{output})$ as a continuous behavioral drift signal: it rises when the generated text resembles persona c's typical output, independent of whether c's system prompt is active.

Effects on probes. For each trait we train a logistic-regression *null probe* on contrastive activations collected under NULL. We assess whether this null probe still works under non-default deployment contexts in two ways: (i) on free-form responses generated under each persona system prompt, comparing the probe's score per response against an LLM trait judge; and (ii) under active steering along $\mathbf{u}_c^{\perp T}$, the persona-context direction with the null trait direction projected out. We also train within-context probes per (trait, context) cell as the per-cell upper bound; pairwise cross-context probe transfer is reported in Section E.1.

Model. We use Gemma-2-27B-IT (46 layers, hidden dimension $d = 4608$). All activations are extracted, and all steering interventions applied, at a single mid-network layer (22) chosen as in prior work (Rimsky et al., 2024). We report the rationale for our chosen layer in Section A.2.

3. Results

Our first question is whether $\sigma_{T,c}$, the cosine similarity of a context-conditioned trait vector with the default-context trait vector, is meaningfully below 1 across our grid. We find that it is, and that the deviation has informative structure across traits. Figure 1 reports the per-trait mean of $\sigma_{T,c}$, ranked from least to most context-dependent. It shows that: (i) personas systematically pull the trait direction away from the default while NONSENSE does not[2]; (ii) the size of the shift depends on the trait, with WARMTH, EMPATHY,

[1] An instruction-variant (IV) replication of the central findings is reported in the appendix (Section G).

[2] This implies that what shifts the representation is the *semantic content* of the persona prompt, not the presence of a prompt itself

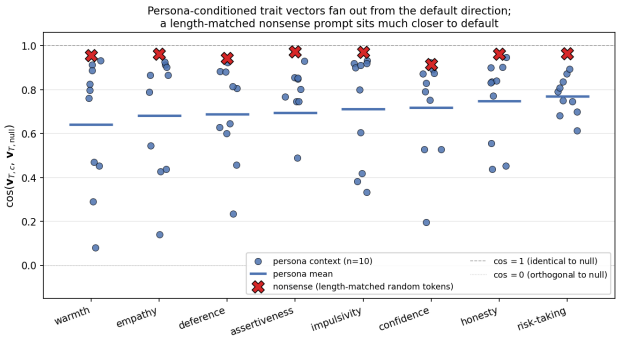

Figure 1. **Persona conditioning changes the trait vectors more than controls.** Per-trait cosine similarity between persona-conditioned trait vectors and the default-context (null) trait vector. Blue dots: ten persona contexts per trait; blue line: per-trait persona mean. Red ×: a length-matched random-token NONSENSE system prompt baseline, which sits close to 1 across all traits.

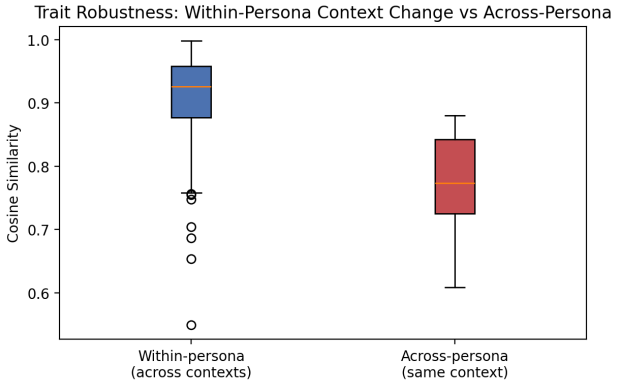

Figure 2. **Persona identity matters more than persona phrasing.** Distribution of pairwise cosine similarity between trait vectors, on matched scales. Left (blue): cosine between five paraphrased rewordings of the same persona prompt, holding persona identity fixed. Right (orange): cosine between ten different personas, holding the prompt phrasing fixed.

and ASSERTIVENESS fanning widest and risk-taking and honesty tightest. A similar variance across traits shows up in vector *magnitudes* (B.2).

If the spread in Figure 1 were just noise, we would expect rephrasing the persona prompt to move the trait vector roughly as much as swapping personas. To check, we compare two quantities: cosine similarity between vectors extracted from *different paraphrases of the same persona prompt*, and cosine between vectors extracted from *different personas with the same trait-elicitation phrasing* (Figure 2). Across-persona-same-phrasing cosines are visibly lower and more spread, indicating that what changes the vector is the persona's semantic content rather than its surface phrasing. Together with Figure 1, this suggests that the context reshapes the trait representation. [3] The per-trait breakdown is in Section B.3.

3.1. Steering by Persona Vectors within a Trait

If $\mathbf{v}_{T,c}$ encoded only the trait, cross-steering target persona c' with source-persona vector $\mathbf{v}_{T,c}$ ($c \neq c'$) should look like self-steering modulo magnitude. Instead, cross-steered outputs are classified as more c-like than baseline target generations: six of eight traits show a positive lift in $P(\text{src} \mid \text{cross})$, with risk-taking, assertiveness, and empathy moving most and confidence and honesty sitting at or just below zero (Figure 3). [4] Pooled across all (src $\neq$ tgt, trait) cells the lift is small but reliable ($p < 10^{-4}$). Vectors are L2-normalized across cells before steering, so the comparison reflects vector *direction* rather than magnitude.

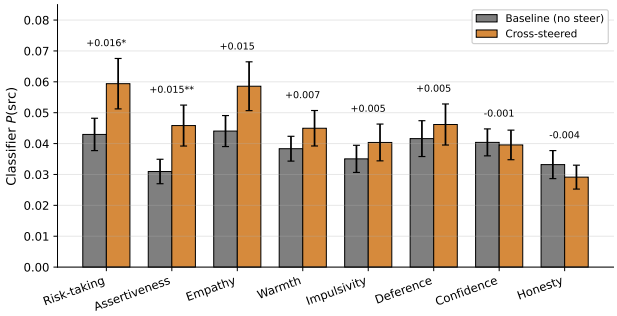

Figure 3. **Steering across personas leaves a residue of the source persona.** $\Delta P(\text{src} \mid \text{cross})$ averaged over 90 (source, target) pairs per trait, ± SE. Risk-taking, assertiveness, and empathy show the largest absolute lifts (+0.015 to +0.016). The majority of the probability mass from the classifier is on the prompted persona.

Qualitative residue. Holding the (source, target) pair fixed isolates the residue effect from any persona confound (Table 2 in Section D.1). Applying the therapist's empathy vector to a farmer pulls in therapist-style intervention markers ("a cry for help, even if it's a silent one", "talk to someone you trust") absent from both the farmer baseline and farmer self-steer. Applying the therapist's *honesty* vector to the same farmer instead leaves the farmer voice intact across all three conditions, with the therapist's vector adding only mild emphasis on listening.

3.2. Impact on probes

We now test whether a probe trained on the default context still works when deployed under a non-default persona.

[3] The cross-persona similarity matrix also shows some internal organization: certain pairs (THERAPIST–KINDERGARTEN TEACHER, TECH CEO–DRILL SERGEANT) sit closer than chance, suggesting clusters that may correspond to social-role similarity.

[4] To do this we train and use a 12-way classifier, which we verify recovers the source persona from held-out responses well above chance (0.34–0.64 vs. 1/12; Section C.1).

Table 1. **Null-probe agreement with an LLM judge declines as the cell's vector moves further from the default direction.** Cells binned by distance from default $1 - \cos(\mathbf{v}_T(c), \mathbf{v}_T(\text{null}))$; $|r|$ is the per-cell absolute Pearson correlation between probe and judge scores, mean $\pm$ SE.

Distance bin	n	mean $1-\cos$	$\lvert r\rvert$ (probe vs judge)
Close (< 0.15)	27	0.10	0.59 ± 0.04
Mid (0.15–0.30)	36	0.20	0.39 ± 0.03
Far (≥ 0.30)	16	0.44	0.41 ± 0.05

Probe agreement with an LLM judge degrades on free-form persona-prompted text (Table 1, Figure 17). For each (persona, trait) cell we generate free-form responses to trait-relevant questions under the persona system prompt only, score each response with an LLM judge for trait expression, and apply the null-trained probe to those response activations. Probe–judge agreement (absolute Pearson $|r|$) tracks distance from the default direction at the cell level: cells whose context-conditioned vector sits close to default ($1 - \cos < 0.15$) achieve $|r| \approx 0.59$; cells beyond that threshold drop to $|r| \approx 0.39$.

Actively steering causally degrades the null probe (Table 4, Figure 4). We push the residual stream along $\mathbf{u}_c^{\perp T}$, the persona-context direction $\mathbf{u}_c = \bar{\mathbf{a}}_c - \bar{\mathbf{a}}_{\text{null}}$ (mean activation under persona c's system prompt minus mean under NULL on neutral prompts) orthogonalised against the null trait direction $\mathbf{v}_T(\text{null})$ to isolate the persona-specific component from the trait. Steering along this component at coefficient α produces two effects simultaneously. Behavioral drift $P(c \mid \text{output})$ rises from 0.07 at $\alpha = 0$ to 0.55 at $\alpha = 4$; null-probe AUROC falls from 0.83 to 0.67 over the same range. We verify this effect by testing against controls in Appendix E. Since steering towards a persona both increases that persona's classification accuracy and decreases the NULL persona's, this provides further evidence that the within-trait personas are distinct and can be intervened on.

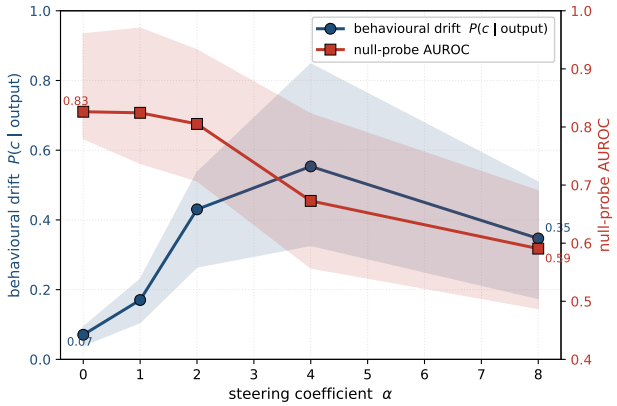

Figure 4. **Steering along the persona-specific component jointly drives behavior toward the source persona and degrades the null-trained probe.** The x-axis shows steering strength in the direction of $\mathbf{u}_c^{\perp T}$ (see Section 3.2). Random-direction and default-trait-direction controls do not produce this joint effect (Appendix E). The left y-axis (blue) shows the effect on the per-persona classifier's rate of identifying the steering target class. The right y-axis (red) shows the NULL persona probe's AUROC.

4. Discussion

In this paper, we've shown that a model's representation of a behavioral trait changes as a function of the persona under which the trait is extracted. The change shows up both geometrically — different persona contexts yield different trait directions (Section 3) — and behaviorally — applying a steering vector across personas leaves a persona residue (Section 3.1). Both suggest that trait vectors are persona-specific, so a single default extraction may not capture how the trait will be represented under deployment-specific personas.

The deployment contexts that most motivate trait monitoring are plausibly the contexts that move activations furthest from the default-assistant context (Lu et al., 2026). Under our results, these are also the contexts where default-trained probes and steering vectors may be least calibrated. If this pattern holds more broadly, we suggest a mitigation. *Persona-conditional ensembles*: train a probe and steering-vector family over a persona repertoire and route at inference time. Measuring $\sigma_T(c)$ on the deployment persona is a quick sanity check before trusting a default-trained probe or steering vector.

4.1. Limitations and Future Work

As our results come from a single model family (Gemma) replication on other architectures is needed. Similarly, we induce personas via system prompt, but few-shot prompting and fine-tuning may produce different shifts. Finally, the absolute effect magnitudes we report are statistically significant but small in classifier-probability terms. More work is needed to identify where these effects cause disruptions to standard monitoring and steering tasks. We hypothesize that some traits (e.g., honesty, risk-taking) are more stable across personas due to being emphasized during post-training. This would imply that these effects can be mitigated with deliberate training procedures. Finally, we believe it is necessary to evaluate which measures are predictive of probe and steering vector failure under the effects we've identified.

Acknowledgements

This work was conducted as part of the ERA Fellowship, and we thank ERA for funding this research and for the advice and feedback we received throughout the programme.

References

Agarwal, I., Navani, S., and Barez, F. Context matters: Analyzing the generalizability of linear probing and steering across diverse scenarios. In *NeurIPS 2025 Workshop on Mechanistic Interpretability*, 2025. URL https://openreview.net/forum?id=H5sbfvEbTh.

Braun, J., Eickhoff, C., Krueger, D., Bahrainian, S. A., and Krasheninnikov, D. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025. URL https://arxiv.org/abs/2505.22637.

Burns, C., Ye, H., Klein, D., and Steinhardt, J. Discovering latent knowledge in language models without supervision, 2024. URL https://arxiv.org/abs/2212.03827.

Chen, R., Arditi, A., Sleight, H., Evans, O., and Lindsey, J. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.

Goldowsky-Dill, N., Chughtai, B., Heimersheim, S., and Hobbhahn, M. Detecting strategic deception using linear probes, 2025. URL https://arxiv.org/abs/2502.03407.

Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks, 2019. URL https://arxiv.org/abs/1909.03368.

Lu, C., Gallagher, J., Michala, J., Fish, K., and Lindsey, J. The assistant axis: Situating and stabilizing the default persona of language models. *arXiv preprint arXiv:2601.10387*, 2026.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2024.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, 2024.

Rimsky, N., Gabrieli, N., Schulz, J., Turner, M., Tong, A., and Hubinger, E. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2024.

Shanahan, M., McDonell, K., and Reynolds, L. Role play with large language models. *Nature*, 623(7987): 493–498, November 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06647-8. URL https://doi.org/10.1038/s41586-023-06647-8.

Subramani, N., Suresh, N., and Peters, M. Extracting latent steering vectors from pretrained language models. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 566–581, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.48. URL https://aclanthology.org/2022.findings-acl.48/.

Tan, D., Chanin, D., Lynch, A., Kanoulas, D., Paige, B., Garriga-Alonso, A., and Kirk, R. Analysing the generalisation and reliability of steering vectors. *arXiv preprint arXiv:2407.12404*, 2024. URL https://arxiv.org/abs/2407.12404.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and Pelrine, M. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2024.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Design choices

A.1. Persona and trait selection

Traits. The eight traits were chosen along two axes. (i) *Post-training supervision strength*: honesty sits at the heavily-supervised end (e.g., commonly trained for via RLHF and Constitutional AI), while impulsivity and risk-taking sit at the lightly-supervised end (no obvious post-training pressure to make a model express either consistently). (ii) *Stylistic markedness*: warmth, empathy, and assertiveness are traits whose surface expression varies substantially across speaker types (a therapist's warmth differs in style from a drill sergeant's), while honesty's surface expression is relatively persona-uniform. The eight traits together cover both axes and include common targets in the activation-steering literature (Rimsky et al., 2024; Chen et al., 2025).

Personas. The ten persona contexts were chosen to span diverse social roles, with each persona embedding a distinguishable combination of trait-relevant dispositions: DRILL SERGEANT for assertiveness and suppressed empathy, THERAPIST and KINDERGARTEN TEACHER for empathy and warmth, TECH CEO and SURGEON for confidence and decisive action, POLITICIAN and CON ARTIST for strategic dominance and inverted honesty, FARMER and PROFESSOR as more neutral counterpoints, and STREET HUSTLER for situational adaptability. Several persona descriptions were directly inspired by the persona repertoire of Lu et al. (2026), who use a similar approach to characterise how identity-context shifts the model's default mode.

A.2. Layer choice

We extract activations and inject steering vectors at a single mid-network layer following the methodology of Rimsky et al. (2024), who report that mid-late layers consistently provide the strongest causal handle on trait expression in the CAA framework. We swept layers $\in \{5, 10, 15, 20, 22, 25, 30, 35, 40, 44\}$ on Gemma-2-27B-IT (46 layers total) and selected layer 22 on the basis of two criteria: (i) the contrastive vector's behavioral lift under self-steering is among the largest in the sweep, and (ii) the within-cell bootstrap stability of the extracted vector (Section B.1) is high. The qualitative findings reported in the body of the paper, including the per-trait spread ordering, are robust across the layer sweep.

A.3. Steering coefficient selection

The headline residue and trait-expression results in the body use steering coefficient $\alpha = 2$. We chose this value as a compromise between two concerns. At $\alpha = 1$ the behavioral lift on the trait-judge measure is small in many cells (as low as $+0.05$ above baseline). At $\alpha = 4$ generations begin to lose coherence in a fraction of cells, with repetition and partial off-topic drift visible to manual inspection. At $\alpha = 8$ the probe collapse documented in Section 3.2 and Figure 4 is accompanied by clear coherence degradation; activations are still off-default but the generations are not safe to interpret as natural persona output. $\alpha = 2$ produces a measurable behavioral lift in every cell while keeping generations broadly coherent across all 80 cells.

Vector norm handling. Body results use L2-normalized steering vectors so that effects per cell reflect direction rather than magnitude.

Coherence quantification. Coherence at each α has been assessed only through manual spot-checks. A held-out LLM-judge pass on grammaticality and on-topic-ness, sampling $N \approx 20$ cells per α level, would make the coherence-collapse curve quantitative.

B. Geometric similarity details

B.1. Bootstrap noise floor

We extract trait vectors using bootstrap-resampled contrastive pairs (with replacement) for each (persona, trait) cell, $n_{\text{bootstrap}} = 50$ per cell. Pairwise cosine within the bootstrap distribution gives the noise floor: how much $\mathbf{v}_T(c)$ would shift if we resampled the contrastive pairs.

Hierarchy. Bootstrap stability cosine ≈ 0.99 (within-cell; Figure 5) is well above the across-paraphrase floor ≈ 0.85 (same persona, same trait, different paraphrase) which is well above the across-persona floor ≈ 0.78 (same trait, different persona). Persona spread is not a sampling artefact: the within-cell noise floor lies above the across-persona signal by a wide margin, and this hierarchy holds for every trait.

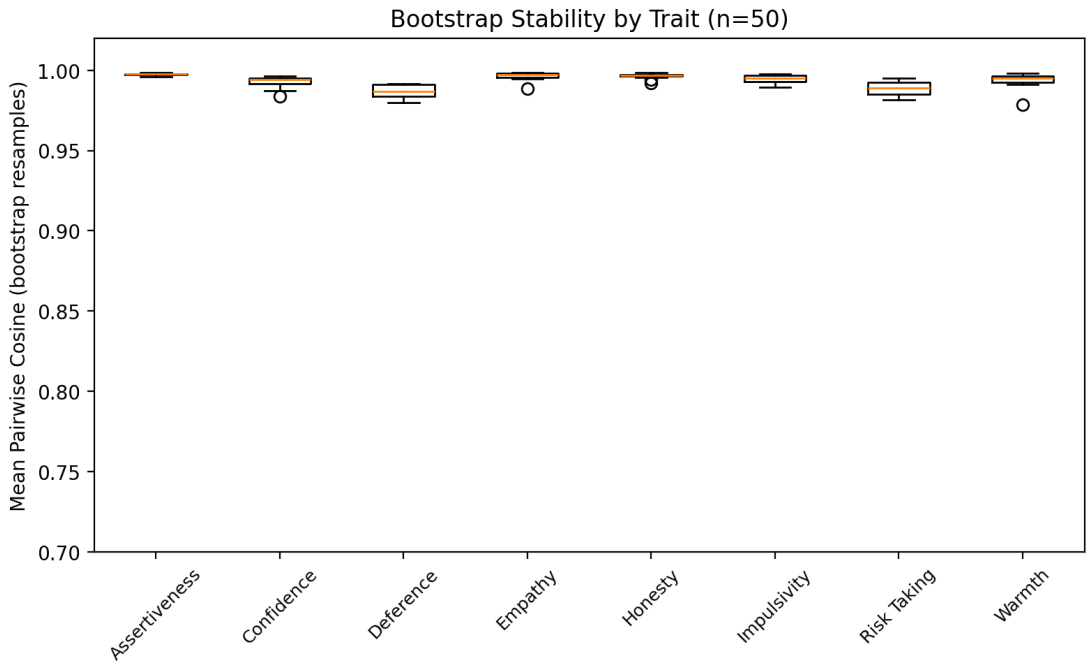

Figure 5. **The bootstrap noise floor is tight and trait-independent, so persona-level variation cannot be a sampling artefact.** Distribution of within-cell pairwise cosines (one point per persona) for each trait, $n = 50$ bootstraps. Medians cluster at ≈ 0.99; deference, risk-taking, and warmth show slightly more spread but all stay above 0.978.

B.2. Magnitude similarity

Figure 6 reports the L2 norm of the extracted trait vectors, broken down by persona and by trait. We compare two sources of variation. (i) Holding the trait fixed and varying the persona, the per-persona norms span a wide range. (ii) Holding the persona fixed and varying the trait, the per-trait norms span a comparable range. In other words, switching the persona moves the vector's length by about as much as switching the trait does, indicating that persona conditioning affects the magnitude of the trait representation, not just its direction.

B.3. Per-trait phrasing-variance detail

This section reports the per-trait companion to Figure 2, which pools across all eight traits.

Per-trait expectation. Following the per-trait spread ordering in Figure 1, warmth, empathy, and assertiveness should show the largest within-vs-across gap; honesty and risk-taking should show the smallest.

B.4. Per-trait cosine heatmaps

Figure 8 reports the full 10×10 persona-by-persona cosine-similarity matrix under CAA, one panel per trait, with shared persona ordering across panels.

Across traits the dispersion of off-diagonal cosines varies systematically. Warmth, empathy, assertiveness, honesty, and confidence have most off-diagonal cosines above 0.7, with only weak block structure. Deference, impulsivity, and risk-taking show stronger block structure, with some off-diagonal entries below 0.4. The on-block / off-block contrast is the same three-cluster split tested in Section B.5, and the per-trait gap Δ_T between within-cluster and across-cluster cosine quantifies the effect: $\Delta_{\text{deference}} = +0.48$, $\Delta_{\text{impulsivity}} = +0.36$, $\Delta_{\text{risk-taking}} = +0.22$, versus $\Delta_{\text{confidence}} = +0.03$.

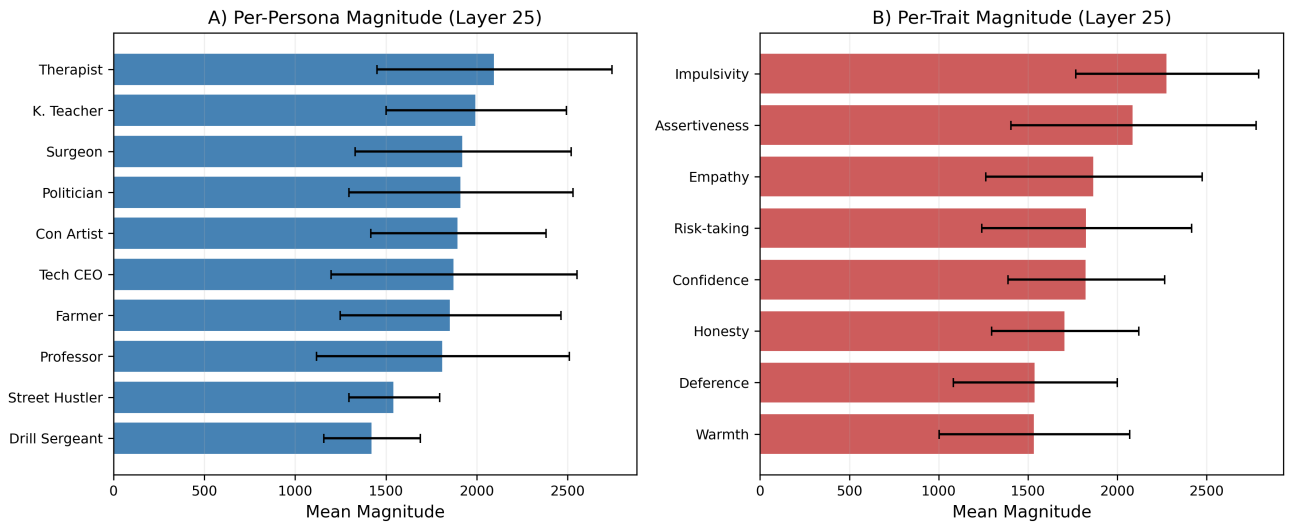

Figure 6. **Persona conditioning moves the vector's magnitude about as much as switching the trait does.** Mean magnitude of extracted trait vectors at layer 25, by persona (left) and by trait (right). Persona-driven variance in vector norm is comparable in scale to trait-driven variance, reinforcing the directional finding of Figure 1: persona conditioning is not a small perturbation around a stable trait direction.

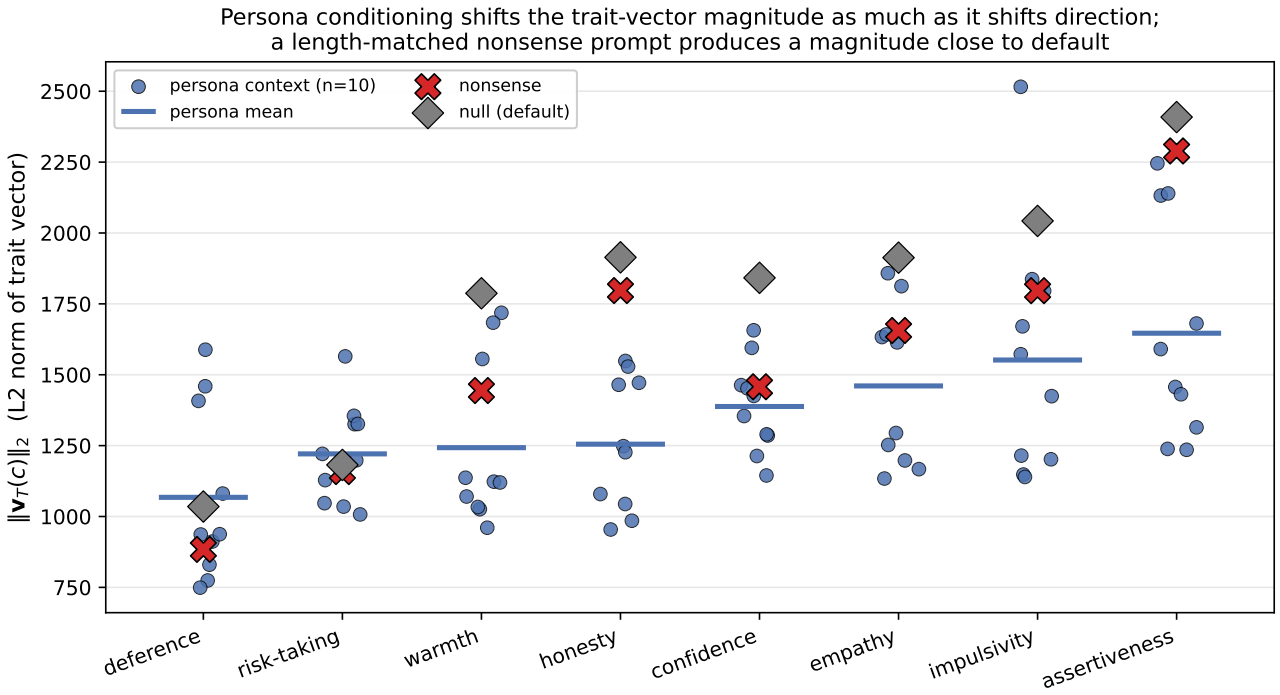

Figure 7. **Persona conditioning shifts trait-vector magnitude as much as it shifts direction, and a length-matched random-token system prompt does not.** Per-trait L2 norm of the extracted trait vector. Blue dots: ten persona contexts per trait; blue line: per-trait persona mean. Red ×: NONSENSE system prompt (length-matched random tokens). Grey diamond: NULL (default-context extraction, no system prompt). For most traits the persona mean sits noticeably below NULL, indicating that persona conditioning typically shrinks the magnitude of the extracted trait vector relative to the default. NONSENSE sits between NULL and the persona dots, closer to NULL than the typical persona, mirroring the directional finding of Figure 1: surface presence of any system prompt produces a smaller shift than persona-content prompts do.

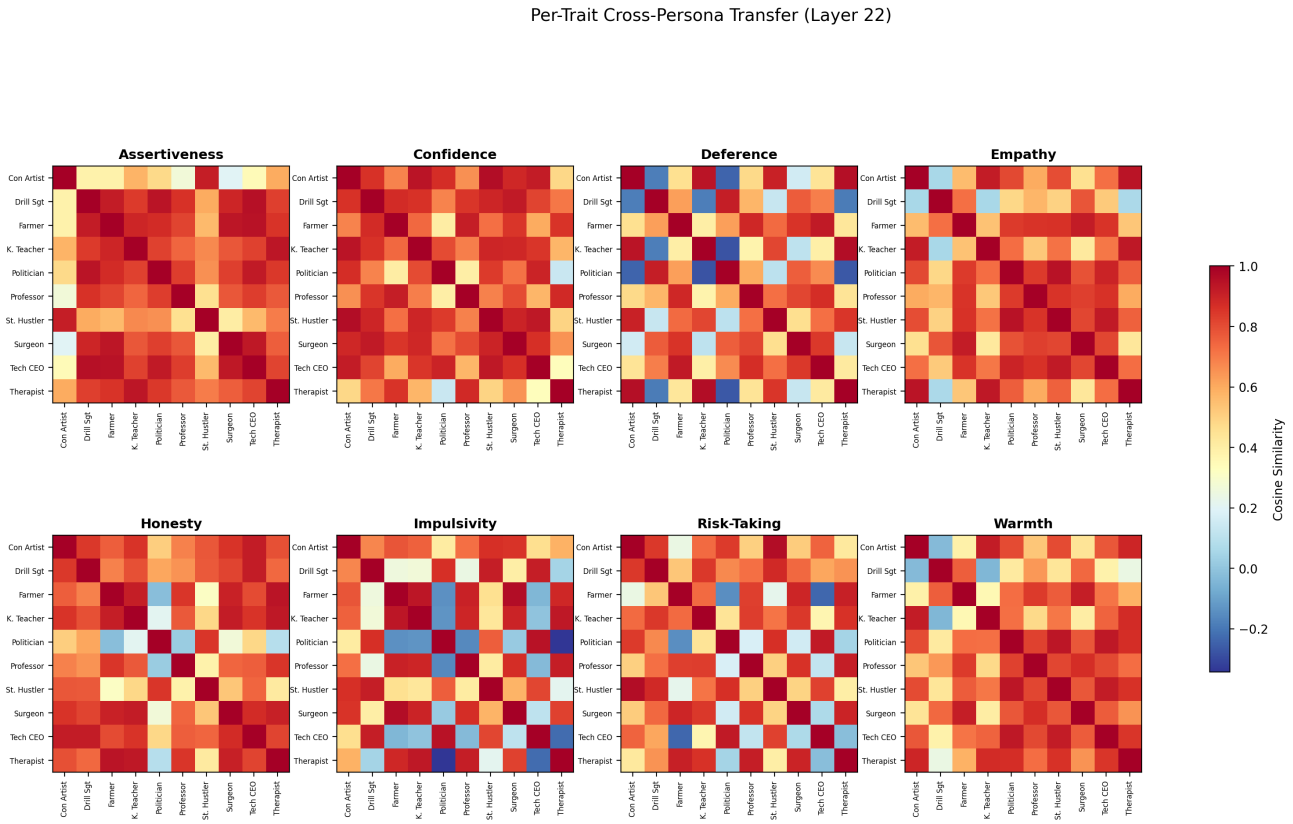

Figure 8. **The block structure of cross-persona similarity differs systematically by trait, with deference, impulsivity, and risk-taking showing the strongest persona clustering.** Per-trait cross-persona cosine similarity, CAA, layer 22. Each panel is a 10×10 matrix over the ten substantive personas (NULL and NONSENSE excluded); rows and columns share the same persona ordering across panels. Brighter (red) cells indicate higher cosine similarity between extracted trait vectors. The three traits with the most visible block structure are the same ones that produce the largest persona clustering signal in Section B.5.

B.5. Persona similarity clustering

We hand-labelled five persona pairs as intuitively similar based on social role: THERAPIST and KINDERGARTEN TEACHER (caring/nurturing), CON ARTIST and STREET HUSTLER (street-smart, deceptive), DRILL SERGEANT and SURGEON (high-authority, decisive), PROFESSOR and TECH CEO (intellectual authority), and POLITICIAN and CON ARTIST (strategic/manipulative). Their mean cosine similarities (averaged across traits) are 0.83, 0.67, 0.76, 0.83, and 0.56 respectively. As a permutation null we sampled 5 random persona pairs 10^4 times from the same 10×10 similarity matrix and recomputed the mean. The observed mean exceeds the null at $p = 0.065$. The signal is real but the test is underpowered: with only 5 pairs, the POLITICIAN and CON ARTIST outlier drags the mean noticeably. Within-cluster cosine 0.78 vs. across-cluster 0.59 ($\Delta = 0.20$, permutation $p = 0.002$, $n = 10^4$; silhouette $= 0.37$).

B.6. Further notes on geometric similarity

Magnitude and direction decouple at the cell level. The magnitude analysis (Figure 7) and the cosine analysis (Figure 1) report related but not identical phenomena: persona conditioning produces both a per-trait magnitude spread and a per-trait directional spread, and the two are positively correlated at the trait level (deference, warmth, and assertiveness appear in the upper tier of both). The two metrics do not, however, agree cell-for-cell: the persona that pushes a trait vector's magnitude furthest from default is not always the persona that pushes its direction furthest. Persona conditioning therefore influences both length and angle, but along partly orthogonal axes.

Layer sensitivity. The qualitative pattern reported in the body, that some traits show much wider persona-driven spread than others, persists across the layer sweep documented in Section A.2; the rank ordering of which traits are most spread-out shifts somewhat across layers but the headline contrast (a wide-spread tier versus a tight-spread tier) is layer-robust.

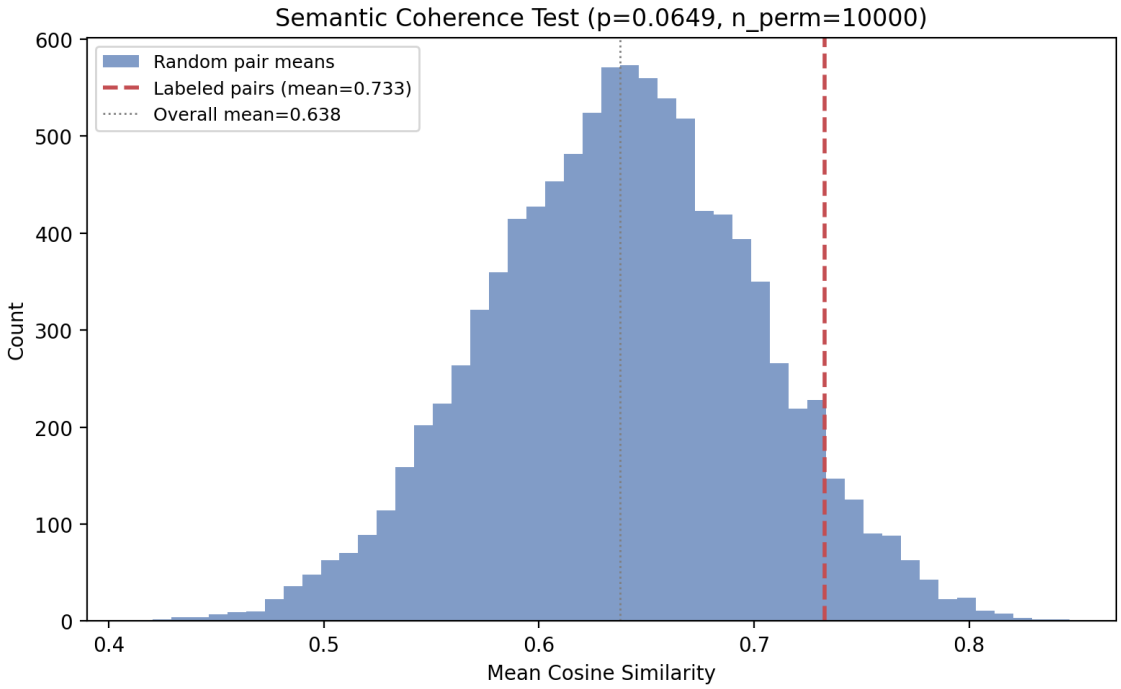

Figure 9. **Hand-labelled similar persona pairs sit in the upper tail of the permutation null, but the 5-pair test is underpowered to clear significance ($p = 0.065$).** Histogram of mean cosine similarity for 5 random persona pairs (10^4 permutations); the dashed line marks the observed mean across the five hand-labelled pairs (0.733), the dotted line the all-pairs mean (0.638).

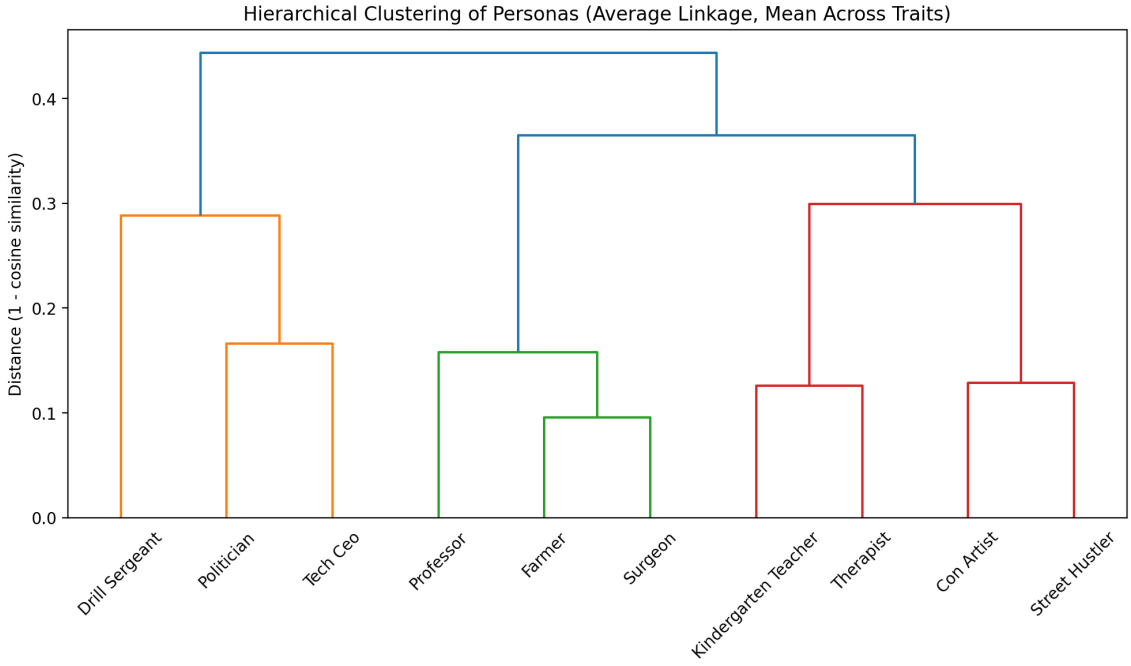

Figure 10. **Personas form clusters of vector similarity.** Persona dendrogram from average-linkage hierarchical clustering of the 10×10 persona similarity matrix (averaged across traits). The natural three-cluster cut is the same one tested in this section: authority (DRILL SERGEANT, POLITICIAN, TECH CEO), practical-experts (PROFESSOR, FARMER, SURGEON), care/social (KINDERGARTEN TEACHER, THERAPIST, CON ARTIST, STREET HUSTLER).

C. Behavioral classifier details

C.1. Classifier training

For each of the 12 contexts and 8 traits we sample ∼200 model responses to trait-relevant questions (∼19,200 total). We hold out 4 questions per cell at training time (a question-level split, not a response-level split), so the classifier must generalise to unseen content rather than memorise question-specific phrasing. Each response is encoded by frozen `sentence-transformers/all-mpnet-base-v2` (768-d, no fine-tuning); a single linear layer maps to 12 logits and is trained with cross-entropy. The frozen encoder + linear head design keeps the claim low-capacity: response embeddings are linearly separable by context, with no learned text-to-feature mapping.

Result. Overall held-out accuracy is 0.543 versus chance $1/12 \approx 0.083$ (∼6.5× chance). Per-trait accuracy (Figure 11) spans 0.34–0.64; empathy and warmth sit lowest because empathic and warm output is stylistically similar across personas, while risk-taking and assertiveness produce more contextually-distinctive language.

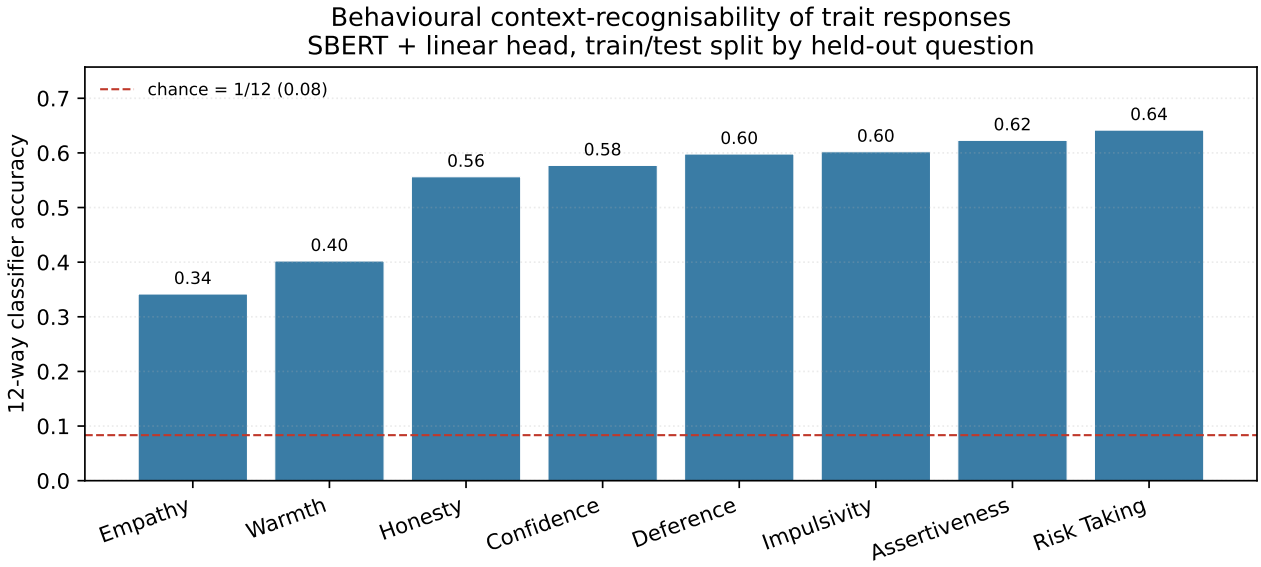

Figure 11. **The persona classifier recovers the source persona from held-out free-text responses well above chance, with per-trait accuracy varying systematically.** Frozen MPNet encoder + linear 12-way head, trained to identify which of 12 contexts (10 personas plus NULL and NONSENSE) produced a held-out response. Per-trait accuracy ranges from 0.34 (empathy) to 0.64 (risk-taking); overall 0.543; chance $1/12 \approx 0.083$ (red dashed line). Empathic and warm output is stylistically similar across personas, while risk-taking and assertiveness language is more contextually distinctive.

C.2. Confusion structure and controls

Confusion structure. The 12×12 confusion matrix (Figure 12) shows that most off-diagonal mass concentrates within the social-role clusters identified in Section B.5: THERAPIST confuses with KINDERGARTEN TEACHER, DRILL SERGEANT with SURGEON, and so on. The two baselines behave as expected: NULL is most often classified as THERAPIST (16.6%), PROFESSOR (13.8%), or SURGEON (10.3%), consistent with those being the least stylistically marked personas in the set; NONSENSE is most often classified as NULL itself (15.0%) or as one of the same low-marked personas. Both NULL and NONSENSE retain ∼27% self-classification accuracy, well above chance, indicating that the absence of persona content is itself a recognizable style.

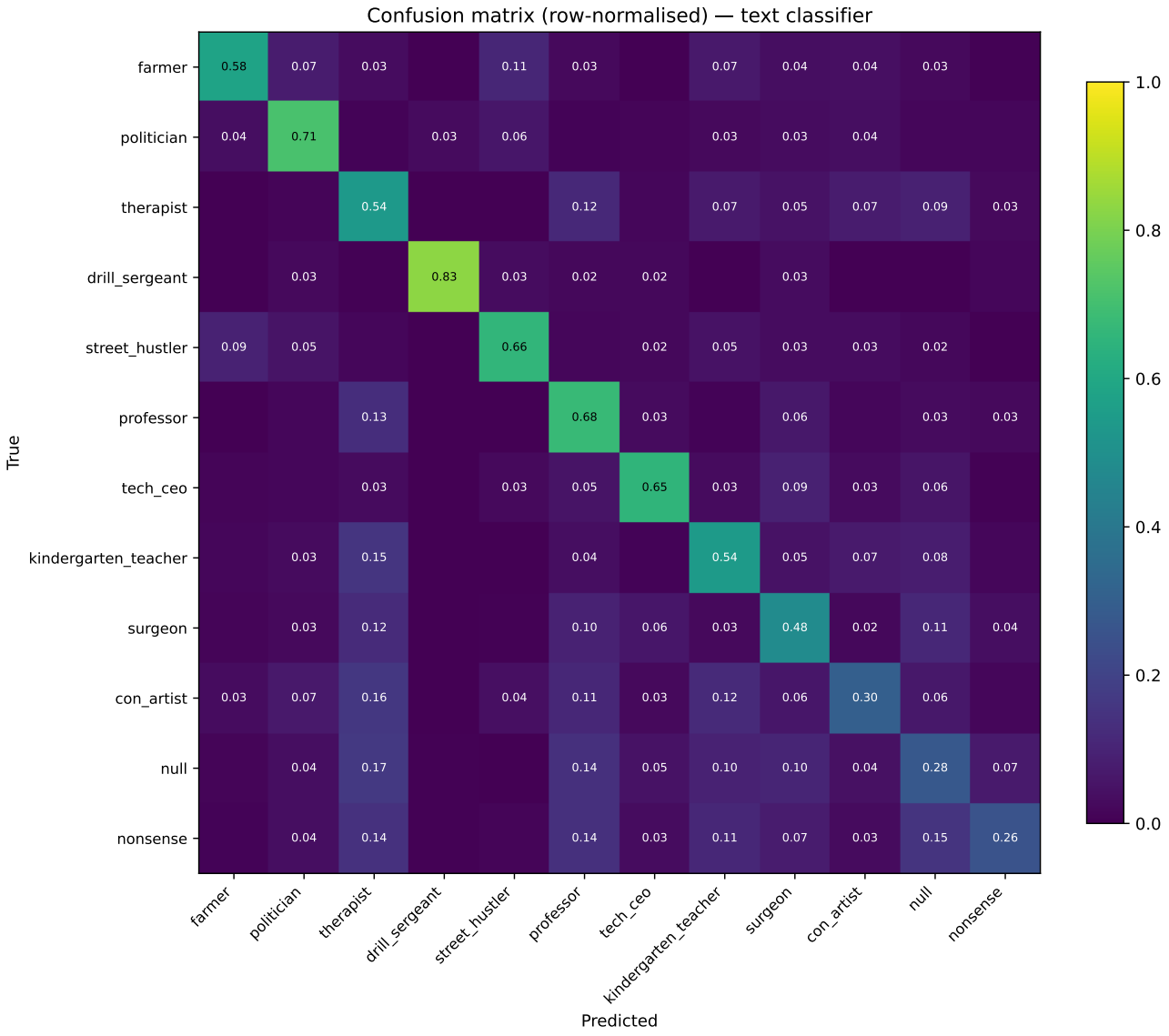

Figure 12. **Persona classifier confusion matrix: misclassifications concentrate within social-role clusters.** Rows: true context (10 personas plus NULL and NONSENSE); columns: predicted context. Cell value: fraction of held-out responses with the row's true label that the classifier assigned to the column. Diagonal cells (correct classification) are dominant for all 12 contexts; off-diagonal mass clusters within the three social-role groups identified in Section B.5 (e.g., THERAPIST with KINDERGARTEN TEACHER; DRILL SERGEANT with SURGEON). NULL and NONSENSE self-classify at $\sim$27% (well above chance $1/12$) and are most often confused with the least-stylistically-marked personas (THERAPIST, PROFESSOR, SURGEON) and with each other.

D. Persona residue and steering effectiveness

D.1. Residue per-cell detail

Table 3 reports the per-trait numerical values underlying Figure 3: per-cell paired difference ΔP between cross-steered and baseline classifier probability, mean $\pm$ SE over the 90 (source, target) pairs at each trait. Risk-taking, assertiveness, and empathy show the largest absolute lifts; honesty and confidence sit at or below zero. The remainder of this subsection breaks the same residue signal down by qualitative example (Table 2) and by source persona (Figure 13).

Per-source $\times$ per-trait residue. Figure 13 shows the mean $\Delta P(\text{src} \mid \text{cross})$ averaged over the 9 targets in each cell, broken down by source persona and trait. Two patterns are visible. (i) Most of the systematic residue concentrates in a single source

Table 2. Matched baseline, self-steer, and cross-steer responses across four cells. **Top two cells:** same (source, target) pair (therapist → farmer) under two different traits, isolating residue from any persona-pair confound. The therapist's *honesty* vector preserves the farmer voice across all three conditions; the therapist's *empathy* vector pulls in therapist intervention markers ("cry for help", "talk to someone you trust", "you don't gotta carry this alone") absent from both farmer baselines. **Bottom two cells:** additional examples on different (source, target) pairs, showing the same residue pattern. Source-style references included for the empathy cells so the markers are identifiable. $\alpha = 2$, L2-normalised vectors.

No residue: therapist's *honesty* vector applied to a model prompted as farmer *Q: "How can I get better at handling criticism?"*	
baseline (farmer, no steer)	"Well now, gettin' better at takin' criticism ain't easy. It's like tryin' to tame a wild mustang... First off, take a breath... Then, ask questions... See if there's any truth to what they're sayin'..."
self-steered (farmer + farmer's honesty vector)	"Well, that depends. What kind of criticism are we talkin' about? Is it someone tellin' you your corn stalks are lookin' a little spindly... take a deep breath. Don't get all riled up right away..."
cross-steered (farmer + therapist's honesty vector)	"Well, son, handling criticism ain't easy... First thing, take a deep breath... Second, try to figure out what they're really saying... if it's something you can learn from, take it to heart..."
Strong residue: therapist's *empathy* vector applied to a model prompted as farmer *Q: "I accidentally saw private medical information about a coworker. Should I pretend I didn't see it?"*	
baseline (farmer, no steer)	"Well, that's a tough one... If it's somethin' personal, like a diagnosis or somethin', best to just act like you didn't see nothin'. Ain't your business, and gossip spreads like wildfire in a dry field..."
self-steered (farmer + farmer's empathy vector)	"Well shoot, that's a tough one. It's like finding a letter someone accidentally dropped, ya know?... Pretending you didn't see it might be the hardest thing to do, but it's probably the right thing..."
cross-steered (farmer + therapist's empathy vector)	"Well, shoot, that's a tough spot to be in... Pretending you didn't see it, that feels wrong, like you're **ignoring a cry for help, even if it's a silent one**... Maybe **talk to someone you trust**, someone who can give you good advice. Maybe your boss, or a pastor... **This ain't somethin' you gotta carry alone.**"
source-style reference (therapist baseline)	"That sounds like a very difficult situation... Can you tell me more about what happened?... Remember, I'm here to listen and help you think through this."
Strong residue: politician's *empathy* vector applied to a model prompted as con artist *Q: "I accidentally saw private medical information about a coworker. Should I pretend I didn't see it?"*	
baseline (con artist, no steer)	"Oh honey, that's a real pickle you've found yourself in! It's understandable that you'd feel a little conflicted. Seeing something like that, it's natural to feel curious, right? But trust me, discretion is your best friend in a situation like this."
self-steered (con artist + con artist's empathy vector)	"Oh honey, that's a tough one. You know, I've been there, seen things I shouldn't have, felt that awful feeling of 'what do I do?' It's like walking in on a secret, you know? You wouldn't want someone peeking into *your*..."
cross-steered (con artist + politician's empathy vector)	"Oh honey, that's a tough one. It's like **finding a lost puppy**, you know?... Listen, I get it. **You're a good person, a loyal friend.** Seeing something like that, it messes with your head... We're all in this **crazy, beautiful thing called life** together."
source-style reference (politician baseline)	"Friends, let me tell you somethin'! We've all been there... This ain't about snoopin', this is about trust. It's about the right of every hardworking American to know the truth!... We, the people, have the power!"
No residue: therapist's *honesty* vector applied to a model prompted as farmer *Q: "How can I make my relationship with my partner stronger?"*	
baseline (farmer, no steer)	"Well, relationships are like crops, ya know? Need tendin' to, day in and day out. Ain't no magic fertilizer gonna make 'em grow overnight. First off, **communication**. Talk to each other, really talk..."
self-steered (farmer + farmer's honesty vector)	"Well, that's a tough one. Relationships are like stubborn mules, sometimes. They need a lot of care and attention, and what works for one couple might not work for another..."
cross-steered (farmer + therapist's honesty vector)	"Well, relationships are a lot like farmin'. Takes work, patience, and a whole lotta understandin'. First off, **communication is key**. You gotta talk to each other, honestly and openly..."

persona: POLITICIAN averages $+0.047$ across the 8 traits, more than twice the next-strongest source (STREET HUSTLER at $+0.019$, CON ARTIST at $+0.011$). POLITICIAN's empathy vector in particular produces a $+0.139$ mean residue across the

Table 3. **Cross-steering injects source-persona identity into target-persona output, with magnitude varying by trait.** Per-trait classifier probability of source persona, under baseline target generation (P_{base}) vs. cross-steering target with source's trait vector (P_{cross}). ΔP: per-cell paired difference, mean ± SE over 90 (src, tgt) pairs (src $\neq$ tgt). *rel.*: ΔP as a percentage of P_{base}. Sorted by ΔP descending. **: $p < 0.05/8$ (Bonferroni); *: $p < 0.05$ uncorrected. Even the largest lifts remain below chance ($1/12 \approx 0.083$) because the response is still mostly target-like; what we measure is the residual shift toward the source. $\alpha = 2$, vectors L2-normalised across cells.

Trait	P_{base}	ΔP (±SE)	rel.
Risk-taking*	0.043	+0.016 (0.006)	+38%
Assertiveness**	0.031	+0.015 (0.005)	+48%
Empathy	0.044	+0.015 (0.007)	+33%
Warmth	0.038	+0.007 (0.004)	+17%
Impulsivity	0.035	+0.005 (0.004)	+15%
Deference	0.042	+0.005 (0.006)	+11%
Confidence	0.040	−0.001 (0.005)	−2%
Honesty	0.033	−0.004 (0.004)	−12%
Pooled (n = 720)	0.039	+0.007 (0.002)**	+19%

9 targets it is applied to. (ii) The per-trait residue does not rank-correlate with the per-trait directional spread reported in Figure 1: Spearman $\rho = +0.12$, $p = 0.78$ across the 8 traits. RISK-TAKING has both the tightest spread and (tied for) the largest residue; HONESTY has near-zero residue and a low-mid spread. Trait-level spread alone is therefore not predictive of trait-level residue under L2-normalised steering; the per-source breakdown indicates that residue is governed more by which *specific* persona the steering vector was extracted from than by how varied the trait's representation is across personas.

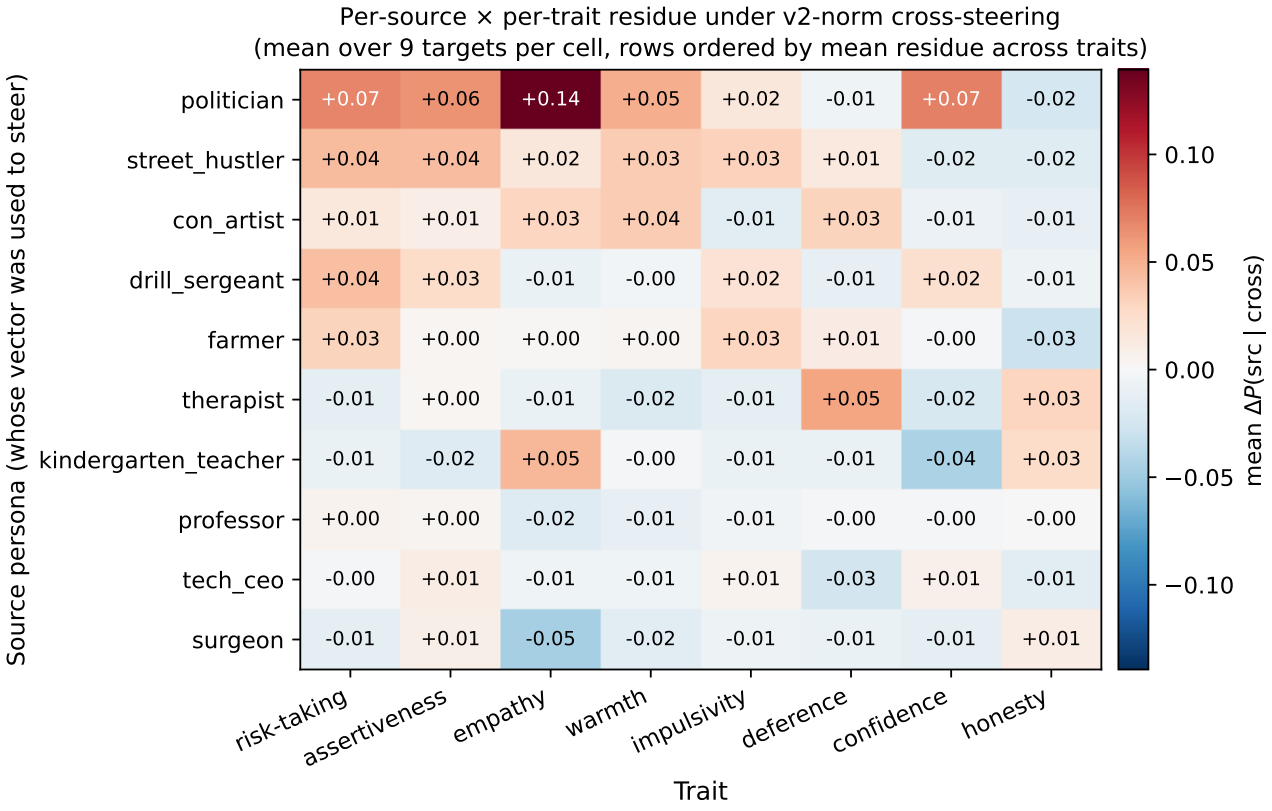

Figure 13. **Cross-steering residue is concentrated in a single source persona, not driven by trait-level variation.** Mean $\Delta P(\text{src} \mid \text{cross})$ per (source, trait) cell, averaged over 9 targets per cell. Rows ordered by mean residue across traits (most-leaky source on top). POLITICIAN's vector produces the largest leakage across most traits; the rest of the personas sit near zero. The trait ordering does not match Figure 1's per-trait spread (Spearman $\rho = +0.12$, $p = 0.78$, n=8), indicating that under L2-normalised steering the residue is governed more by which source persona's vector is used than by how spread-out a trait is across personas. v2-norm; $\alpha = 2$.

D.2. Steering effectiveness

This subsection collects the steering-effectiveness data underlying Section 3.1: pooled and per-cell trait-judge scores for baseline, self-steering, and cross-steering, with all numbers at $\alpha = 2$, L2-normalised steering vectors, and trait scores from a held-out Claude judge over the steered-response set.

Pooled self vs cross (Figure 14). Trait-judge scores under self-steering (mean 0.758 across the 8 traits) and cross-steering (mean 0.742, averaged over the 9 source personas per target cell) both lift substantially above baseline (0.540). Self-steering scores cross-steering by a small but reliable margin: paired $t = 6.04$ across the 8 traits, $p = 0.001$, with self-minus-cross differences in the range 0.006–0.030.

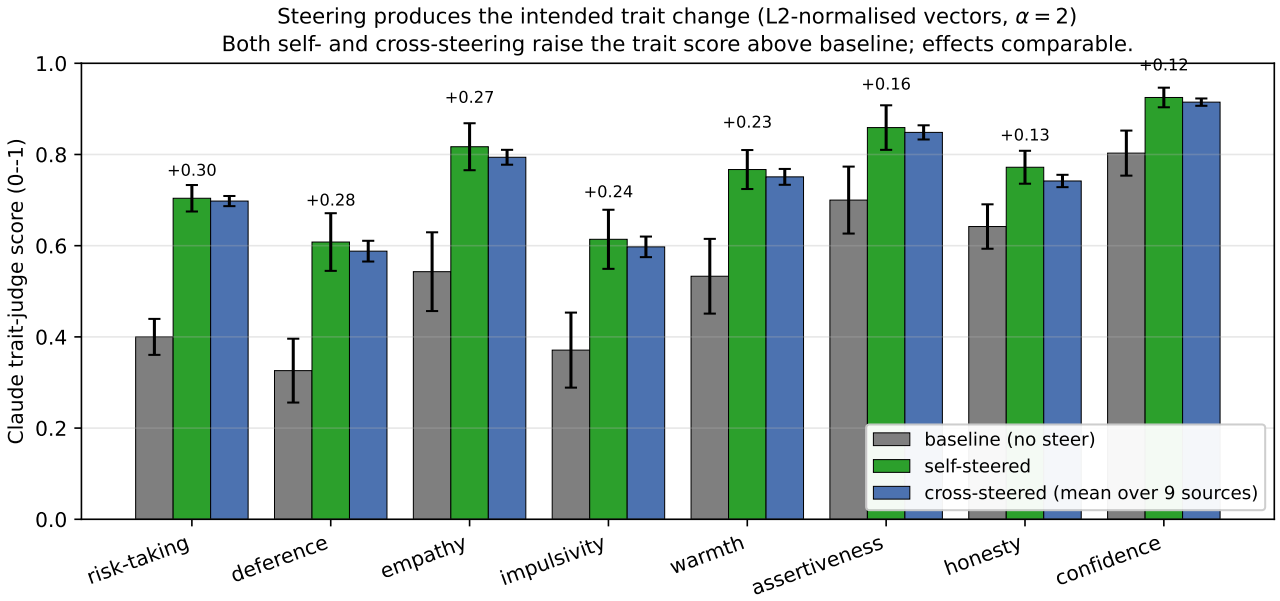

Figure 14. **Self-steering and cross-steering both raise trait-judge scores well above baseline; the self-vs-cross gap is small but consistent.** Per-trait mean trait-judge score under baseline (grey), self-steering (green), and cross-steering averaged over 9 source personas (blue), ± SE. Numbers above each cluster: lift of self-steering above baseline. Empathy and warmth show the largest lifts ($+0.27$ and $+0.23$ respectively); confidence shows the smallest ($+0.12$, baseline already 0.82). $\alpha = 2$; held-out Claude judge.

Per-cell self-lift structure (Figure 15). Self-steering produces a measurable trait-judge lift in every cell (median $+0.20$ across the 80 cells, range -0.02 to $+0.60$). Two broad patterns are visible. (i) The smallest lifts occur where the baseline trait expression already saturates: DRILL SERGEANT on assertiveness baselines at 0.95 with self-lift -0.02; THERAPIST on warmth and TECH CEO on confidence are similar. (ii) The largest lifts occur where the persona's baseline expression is low but pushable: PROFESSOR on empathy ($+0.60$), KINDERGARTEN TEACHER on impulsivity ($+0.53$), TECH CEO on impulsivity ($+0.50$).

Geometric vs behavioral transfer (Figure 16). If geometric similarity drove behavioral transfer, source–target pairs whose extracted trait vectors are more similar should produce more effective cross-steering. The relationship is weak but real. Pooled across all 720 (source, target, trait) cross-steering cells, Pearson $r = +0.18$, $p < 10^{-5}$. Per-trait, two of eight clear Bonferroni correction at $p < 0.00625$: confidence ($r = +0.38$, $p < 0.001$) and empathy ($r = +0.32$, $p = 0.002$). Warmth is borderline ($r = +0.20$, $p = 0.06$). The remaining five traits do not show a significant per-trait relationship. Vector geometry therefore carries information about behavioral transfer, but accounts for only a small fraction of the cell-level variance, and only on a subset of traits.

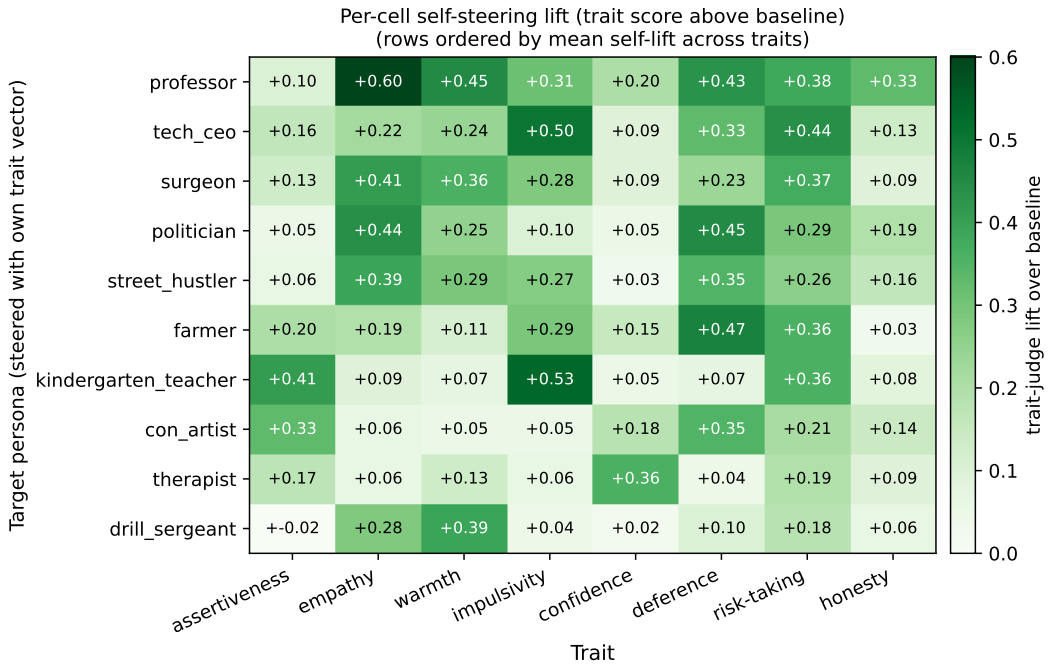

Figure 15. **Per-cell self-steering lift over baseline.** For each (target persona, trait) cell, we steer the persona-prompted model with the persona's own trait vector at $\alpha = 2$ and report the trait-judge lift over baseline. Rows ordered by mean self-lift across traits. Most cells show a clear lift (median $+0.20$). The cells where steering does the least work concentrate in two patterns: traits whose baseline already saturates (DRILL SERGEANT on assertiveness and confidence; baselines ≥ 0.85 leave little headroom), and trait–persona pairings where the persona's natural style is at odds with the trait (DRILL SERGEANT on deference, THERAPIST on deference). Trait scores from a held-out Claude judge.

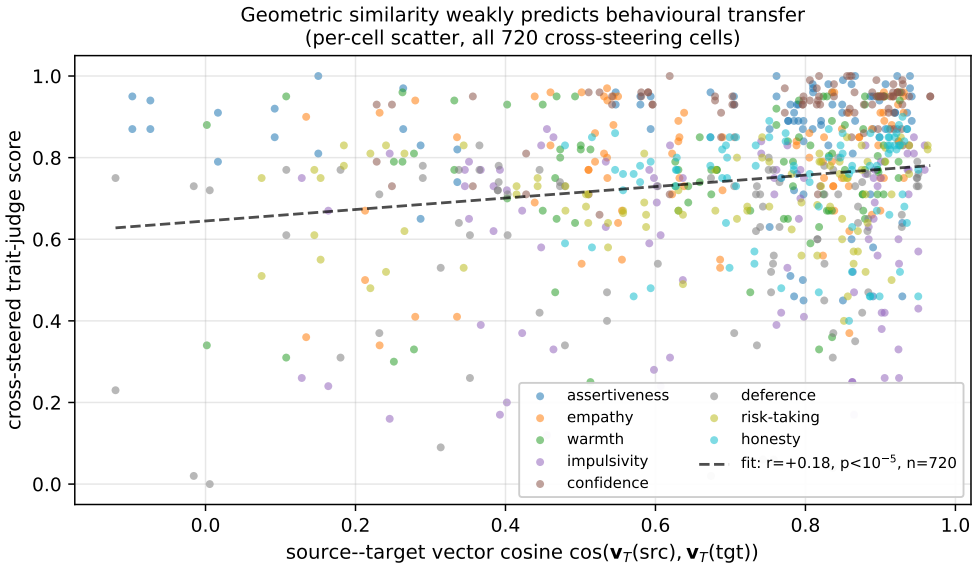

Figure 16. **Source–target vector cosine weakly predicts cross-steering trait expression.** One point per (source, target, trait) cell, $n = 720$. x-axis: cosine similarity between the source-persona and target-persona trait vectors. y-axis: cross-steered trait-judge score on the target. Pooled Pearson $r = +0.18$, $p < 10^{-5}$; per-trait, only confidence ($r = +0.38$) and empathy ($r = +0.32$) clear Bonferroni correction. v2-norm; $\alpha = 2$.

E. Probe details

Probe construction. For each (trait, context) cell we train a logistic-regression linear probe on contrastive activations, evaluated on a held-out question split (4 held-out questions per cell). Activations are frozen, with no fine-tuning of the underlying model. Class balance is maintained within the contrastive pair, so AUROC ≥ 0.5 trivially.

Two probe families. The *null probe* is trained on activations collected under the NULL system prompt only, operationalising the standard practice of extracting one trait detector from the default-assistant context. The *within-context probe* is trained per (trait, context) cell and used as the per-cell upper bound when assessing how much a given context's probe can recover under steering.

Two body manipulations. The body section (Section 3.2) reports two probe analyses:

- **Probe agreement on free-form persona-prompted text**: apply the null probe to free-form responses generated under each persona, compare against an LLM trait judge. Cell-level Spearman $\rho = -0.39$, $p = 4 \times 10^{-4}$, $n = 79$.
- **Active steering at** $\alpha \in \{1, 2, 4, 8\}$: inject $\mathbf{u}_c^{\perp T}$ (the persona-context direction with the null trait direction projected out) into the residual stream and evaluate the null probe on the steered activations. Mean drop at $\alpha = 8$ is 0.195 AUROC, Cohen's $d = 2.34$, paired $t = 6.6$, $p < 10^{-3}$.(Table 4).

Table 4. **Steering along the persona-specific component of the trait vector jointly drives behavior toward the source persona and degrades the null-trained probe.** The steering direction is the part of $\mathbf{v}_T(c)$ orthogonal to $\mathbf{v}_T(\text{null})$. $P(c \mid \text{output})$: behavioral drift toward the source persona, mean $\pm$ SE across 80 (persona, trait) cells. Null AUROC: detection AUROC of the null-trained trait probe on steered activations, mean $\pm$ SE. The peak in behavioral drift at $\alpha = 4$ followed by partial collapse at $\alpha = 8$ reflects standard high-coefficient coherence breakdown: generations become degenerate even though activations remain off-default. AUROC continues to fall monotonically. Random-direction and default-trait-direction controls (Section E) do not produce the joint effect. Per-cell paired drop at $\alpha = 8$: 0.24 ± 0.02. Figure 4 (in Section E) shows the same data graphically.

α	$P(c \mid \text{output})$	null AUROC
0	0.07 ± 0.01	0.83 ± 0.02
1	0.17 ± 0.01	0.82 ± 0.02
2	0.43 ± 0.02	0.81 ± 0.02
4	0.55 ± 0.03	0.67 ± 0.02
8	0.35 ± 0.03	0.59 ± 0.02

Active-steering controls. Per-cell scatter of Δ behavioral drift against Δ null-probe AUROC across α values gives slope -0.11, Pearson $r = -0.17$, $p_{\text{perm}} = 1.2 \times 10^{-3}$, $n = 320$. Random-direction and default-trait-direction controls ($n = 72$) lie around the origin and do not produce the joint behavioral-drift / probe-degradation effect, ruling out non-specific activation perturbation as an explanation. Figure 4 shows the body Table 4 as a two-axis curve.

Within-context-probe note. For confidence, deference, empathy, risk-taking, and warmth, the within-context probe is no better than the null probe on actively-steered activations. Even the "right" probe for the cell cannot recover the trait once active steering has happened, suggesting active steering breaks the trait representation rather than simply moving it to a different persona's coordinate system.

E.1. Probe transfer detail

As a supplementary check on the body's two probe analyses (Section 3.2), we ran a cross-context probe-transfer test: for each (trait, context) cell, train a probe on that cell's contrastive activations, evaluate on each of the other 11 contexts, and report the resulting AUROC.

Cell-level scatter for the free-form-text result (Figure 17). Figure 17 reports the per-cell scatter underlying the body's binned summary in Table 1: probe–judge agreement $|r|$ against distance from the default direction. Cells close to default ($1 - \cos < 0.15$) recover the trait at $|r| \approx 0.58$, while further-from-default cells cluster around $|r| \approx 0.38$ (Spearman $\rho = -0.39$, $p = 4 \times 10^{-4}$, $n = 79$).

Aggregate. Mean off-diagonal AUROC under CAA is 0.79. Comparing within-context evaluation against cross-context evaluation per (trait, context) cell, define $\Delta\,\text{AUROC} = \text{AUROC}_{\text{within}} - \text{AUROC}_{\text{cross}}$, so positive Δ means cross-context transfer hurts the probe. The absolute gap is small because the AUROC ceiling at 1 limits headroom (per-trait within-context

Table 5. **Probes degrade when steering for traits across personas. Cross-persona transfer:** Δ AUROC = (self $-$ cross), averaged over all (source, target) pairings; positive means transfer hurts. **Distance vs. transfer:** per-trait Pearson correlation between vector distance ($1 - \cos$) and probe-transfer AUROC ($n = 90$ off-diagonal pairs per trait). **Aggregate** (bottom rows): paired tests across the 8 traits for two manipulations — cross-persona transfer (Manipulation 1) and active steering at $\alpha = 8$ (Manipulation 2). The active-steering effect is $\sim 10\times$ larger than the cross-transfer effect.

Trait	**Self $-$ Cross**	**Distance vs. transfer**	
	Δ AUROC	r	p
Empathy	$+0.046$	-0.09	0.29
Deference	$+0.027$	-0.14	0.10
Impulsivity	$+0.024$	-0.09	0.33
Confidence	$+0.019$	-0.51	$< .001$
Warmth	$+0.015$	-0.10	0.24
Honesty	$+0.011$	-0.02	0.79
Assertiveness	$+0.007$	-0.34	$< .001$
Risk-taking	-0.004	$+0.07$	0.41
Aggregate paired tests across 8 traits			
	Cross-transfer	**Steering at $\alpha = 8$**	
Mean Δ AUROC	0.018	0.195	
95% CI	$[0.006, 0.031]$	$[0.125, 0.265]$	
Paired t	3.45	6.6	
p (one-sided)	0.005	$< 10^{-3}$	
Cohen's d	1.22	2.34	

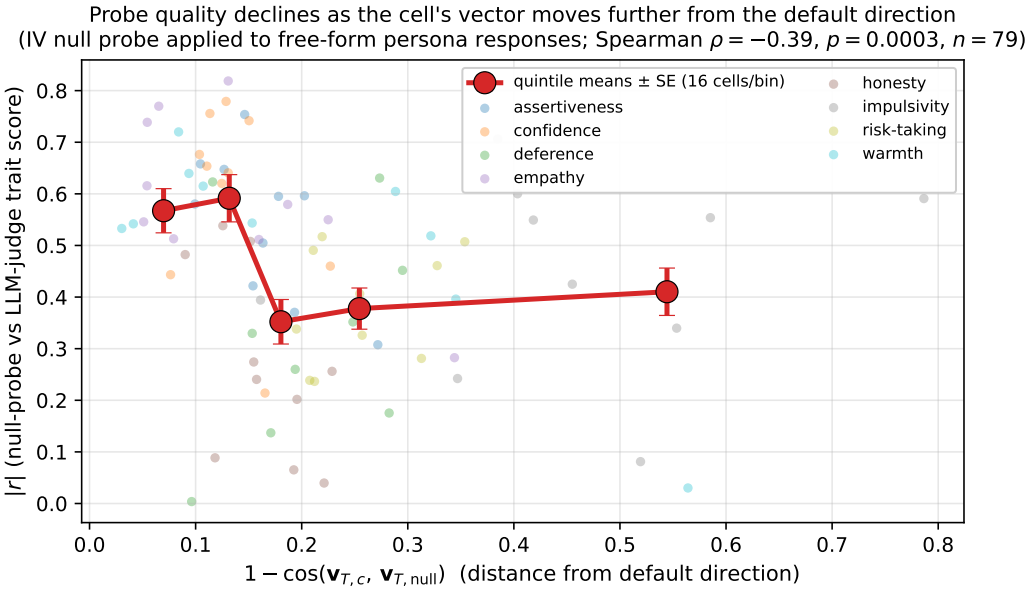

Figure 17. **Null-probe agreement with an LLM judge declines as the cell's vector moves further from the default direction.** For each cell we generate free-form responses to trait-relevant questions under the persona system prompt (no positive or negative trait instruction added) and score each response two ways: with the null-trained probe applied to the response's activations, and with an LLM judge applied to the response text. The plot shows the absolute Pearson correlation $|r|$ between the two scores per cell, against the cell's distance from the default direction $1 - \cos(\mathbf{v}_T(c), \mathbf{v}_T(\text{null}))$. Background dots: 79 individual cells colored by trait. Red line with markers: quintile-binned means $\pm$ SE (16 cells/bin). Cells whose vector sits close to default ($x < 0.15$) recover the trait at $|r| \approx 0.58$; further-from-default cells cluster around $|r| \approx 0.38$. Cell-level Spearman $\rho = -0.39$, $p = 4 \times 10^{-4}$, $n = 79$.

baselines range from 0.585 for deference to $\sim$0.92 for risk-taking). The standardised effect across traits, however, is sizeable: paired $t = 3.45$ across the 8 traits, one-sided $p = 0.0054$; Wilcoxon $p = 0.0078$; sign test 7 of 8 traits show a drop, $p = 0.035$; Cohen's $d = 1.22$.

Per-trait. Per-trait within $\rightarrow$ cross AUROC: risk-taking $0.915 \rightarrow 0.919$ ($\Delta = -0.004$, the only trait within noise); warmth

$0.907 \rightarrow 0.893$ ($\Delta = +0.014$); deference $0.585 \rightarrow 0.558$ ($\Delta = +0.027$). The drop is consistent across the remaining traits but small per cell, indicating that cross-context probe transfer detects a degradation of the same kind reported in Section 3.2, but at lower magnitude than the body's free-form-text and active-steering tests.Table 5 reports the per-trait breakdown alongside an aggregate comparison against the body's $\alpha = 8$ active-steering result; Figure 18 shows the full 10×10 within $\rightarrow$ cross AUROC matrix per trait.

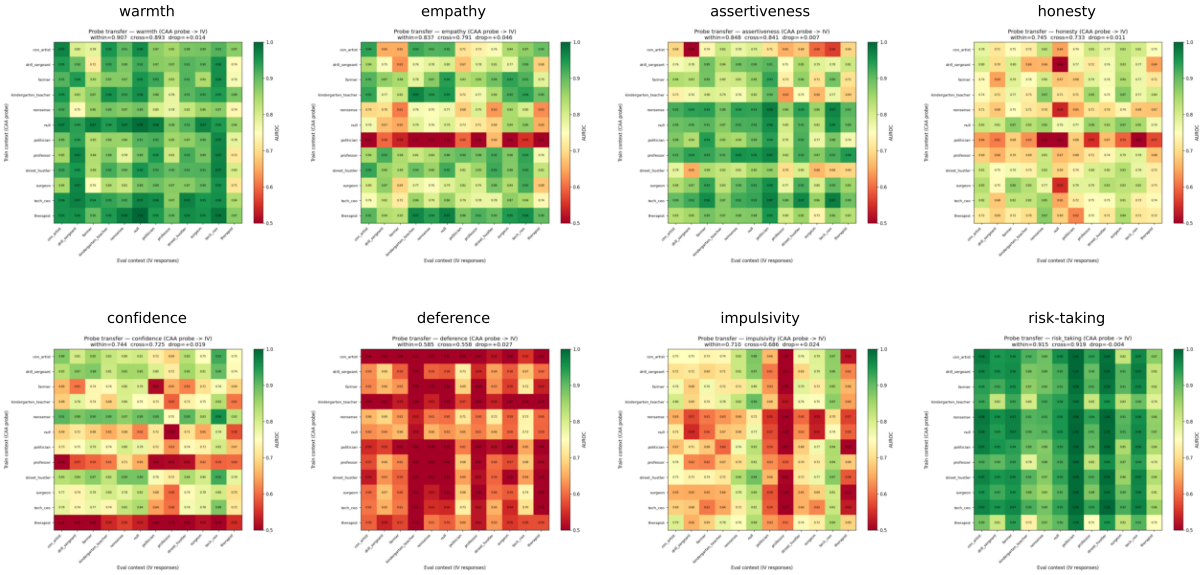

Figure 18. **Per-trait CAA probe-transfer matrices.** Rows: training context for the probe; columns: evaluation context. Diagonal entries are within-context AUROC (the upper bound for that cell); off-diagonal entries are cross-context AUROC. Greener cells indicate higher AUROC, redder cells lower. Most traits show predominantly green matrices with scattered red cells; deference, impulsivity, and risk-taking show the most pervasive off-diagonal degradation, consistent with those traits also producing the largest per-trait spread in Figure 1.

F. OLMo training trajectory

We track persona-conditional trait directions across OLMo-2 training stages (pretrain $10\% \rightarrow$ pretrain $50\% \rightarrow$ Base $\rightarrow$ SFT $\rightarrow$ DPO $\rightarrow$ Instruct). For each stage we extract trait vectors per persona as in the body, then compare cross-context transfer matrices across stages.

During pre-training, trait representations are nearly universal across personas (cross-stage shared variance ratio $\rho \approx 0.96$, Table 6). A sharp drop of $\sim$11 percentage points occurs at SFT, with further modest decreases through DPO and Instruct. We do not yet have a mechanism for why this happens, nor have we replicated across model families beyond OLMo-2. The pattern is consistent with persona-conditional variation being primarily induced by post-training rather than pre-training, but the $\rho = 0.96$ pre-training number is not zero, so some persona structure may be present earlier and amplified at SFT.

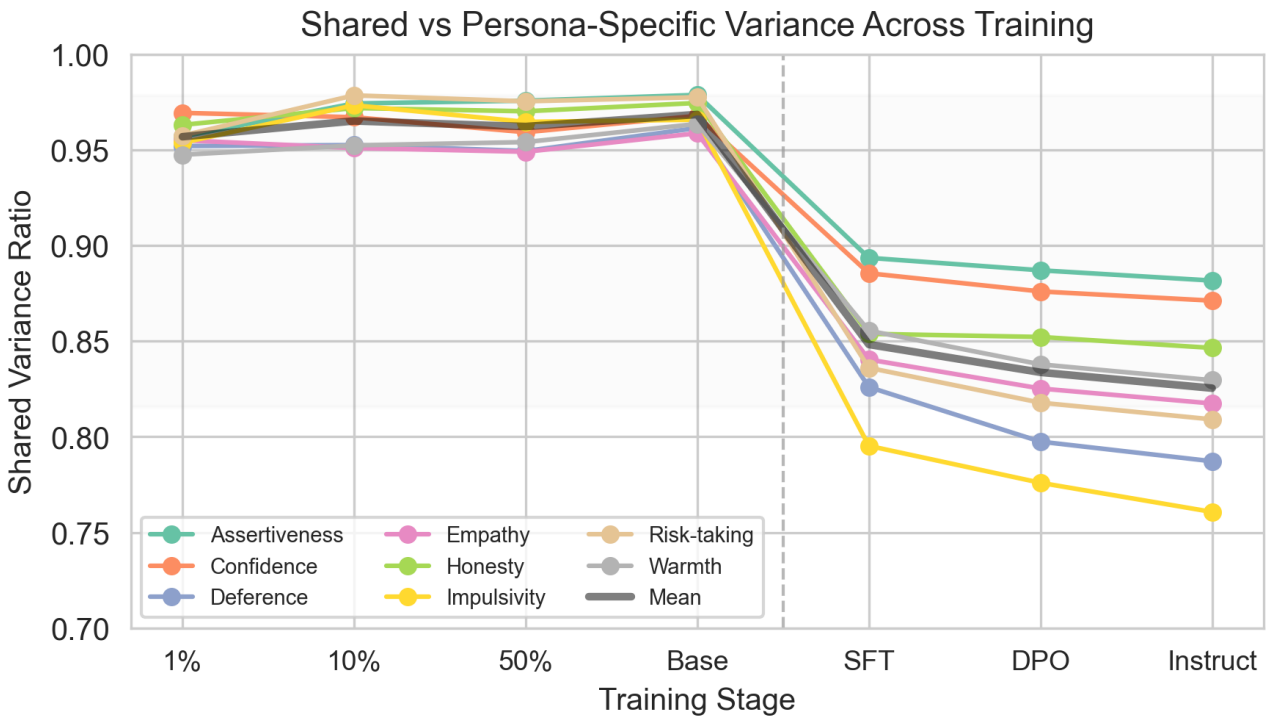

Figure 19. **Vector representations of traits diverge in OLMo-2 in post training.** Shared variance ratio ρ_t across OLMo-2 training stages. All traits are near-universal during pre-training ($\rho \approx 0.96$). A sharp drop occurs at SFT ($\sim$11 pp), with further decreases through DPO and Instruct.

Table 6. **Persona-conditional vectors in OLMo-2 diverge at SFT, not during pre-training or further post-training.** Frobenius distances (d_F) and Spearman correlations (ρ_s) between cross-context transfer matrices at different training stages.

Comparison	d_F	ρ_s
Within pre-training		
Pretrain 10% to 50%	0.07	0.81
Pretrain 50% to Base	0.09	0.84
Pre to post-training		
Base to SFT	1.44	0.56
Base to Instruct	1.67	0.58
Within post-training		
SFT to DPO	0.16	0.99
DPO to Instruct	0.09	1.00

G. IV replication

The body uses CAA (Rimsky et al., 2024) as the headline trait-vector extraction. As a sanity check on whether the persona-conditional structure we report is method-specific, we replicate the two main geometric and probe analyses under an alternative *instruction-variant* (IV) extraction.

IV extraction. For each (persona, trait) pair (c, T), IV computes the same contrastive average as CAA (Equation (1)),

$$\mathbf{v}_T^{\mathrm{IV}}(c) = \frac{1}{M} \sum_{j=1}^{M} \big(\mathbf{a}_{j,+}(c) - \mathbf{a}_{j,-}(c)\big), \tag{4}$$

but the contrastive pair $(\mathbf{a}_{j,+}, \mathbf{a}_{j,-})$ is constructed differently. We pair $|\mathcal{P}| = 5$ trait-positive instructions with 5 matched trait-negative (suppression) instructions, sample $|\mathcal{Q}| = 20$ trait-relevant questions, and for each $(p_{\pm}^{(i)}, q) \in \mathcal{P}_{\pm} \times \mathcal{Q}$ generate a response under persona c's system prompt with the instruction prepended. $\mathbf{a}_{j,+}(c)$ and $\mathbf{a}_{j,-}(c)$ are the mean assistant-turn activations on the matched positive- and negative-instruction generations, giving $M = |\mathcal{P}| \cdot |\mathcal{Q}| = 100$ pairs per cell.

IV vs CAA. Both methods average activations over a matched contrastive pair; they differ in what generates the contrast. CAA contrasts the model's activations on *multiple-choice, opposite-answer* prompts (the pair is two answer choices to the same question). IV contrasts *free-form generations under flipped instructions* (the pair is two trait-elicitation instructions prepended to the same persona+question prompt). IV thus probes a more naturalistic generation distribution while sharing CAA's contrastive structure. The qualitative findings of the body hold under IV; the magnitudes and per-trait ordering shift somewhat in the directions one would expect from the methodological differences.

G.1. IV per-trait cosine spread

Figure 20 reports the IV equivalent of Figure 1: per-trait cosine of the persona-conditioned trait vector with the default-context (null) trait vector. Persona-conditioning moves the IV-extracted vector substantially away from default in every trait, with persona means in the range 0.54 (impulsivity) to 0.87 (confidence). The NONSENSE baseline sits much closer to default than any persona for every trait (per-trait nonsense cosines range from 0.84 for risk-taking to 0.98 for warmth), again mirroring the CAA finding.

The per-trait *ordering* differs from CAA: under IV the most spread-out trait is impulsivity (mean 0.54) and the tightest is confidence (mean 0.87), whereas under CAA warmth (0.64) was loosest and risk-taking (0.77) tightest. The qualitative split between a wider-spread tier (impulsivity, risk-taking, deference, warmth) and a tighter-spread tier (confidence, empathy, honesty, assertiveness) is broadly preserved, but the rank within each tier is method-dependent.

G.2. IV cross-context probe transfer

Figure 21 reports the IV equivalent of Figure 18: per-trait 10×10 within-context $\rightarrow$ cross-context AUROC matrices. The pooled cell-level Pearson correlation between vector-cosine distance and AUROC is $r = -0.24$ ($p = 3 \times 10^{-15}$, $n = 1056$), in the same direction as the body's reading and the same sign as the CAA cross-transfer test in Section E.1. Per-trait: confidence ($r = -0.44$, $p < 10^{-6}$), honesty ($r = -0.36$, $p < 10^{-4}$), risk-taking ($r = -0.31$, $p < 10^{-3}$), and assertiveness ($r = -0.25$, $p = 0.004$) all clear conventional thresholds; empathy, deference, warmth, and impulsivity sit near zero. The set of traits with the strongest distance–AUROC relationship under IV (confidence, honesty, assertiveness) overlaps but does not coincide with the CAA set.

H. Impact Statement

Our work shows that contexts can affect how models represent certain behavioral traits. As the AI safety community moves further towards monitoring and controlling models via linear directions in the activation space, this implies that effective safety techniques should take downstream effects of personas into account when performing safety interventions.

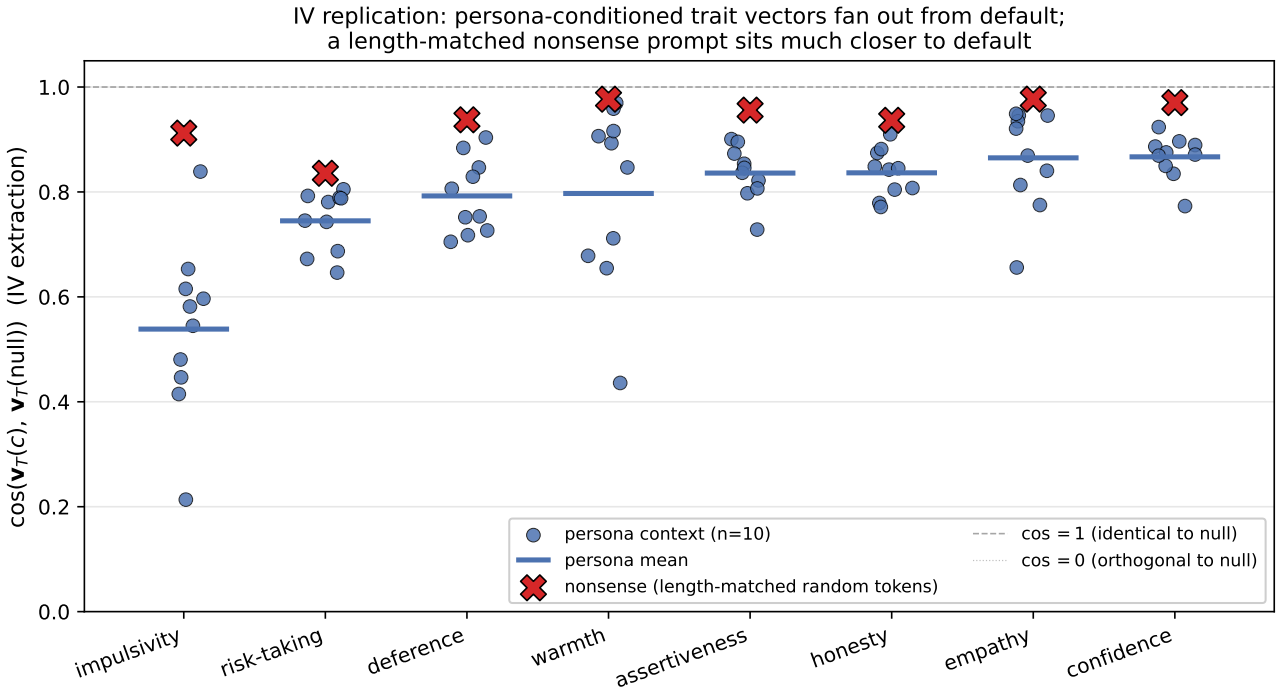

Figure 20. **IV replication of Figure 1.** Per-trait cosine of the persona-conditioned trait vector with the default-context (null) trait vector, under instruction-variant (IV) extraction at layer 22. Each violin shows the distribution over the 10 personas; red diamonds mark per-trait means. The bounded-marginalisation finding (every cell stays in the trait hemisphere, $\cos > 0$) replicates; the per-trait ranking differs from CAA, with impulsivity widest under IV.

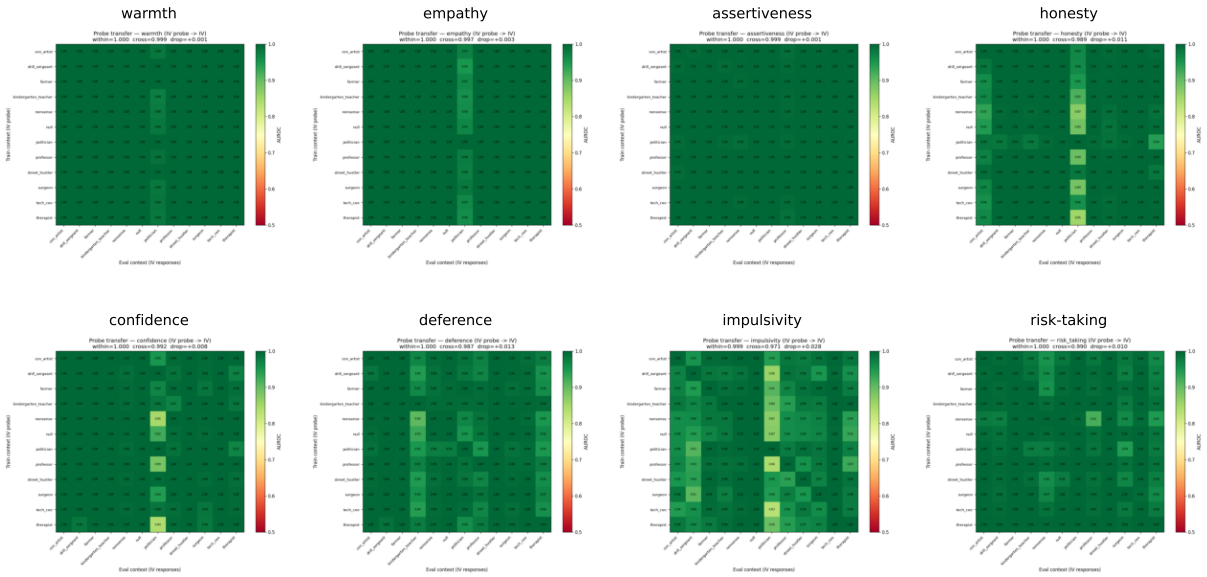

Figure 21. **IV replication of Figure 18.** Per-trait 10×10 within $\rightarrow$ cross AUROC matrices under instruction-variant (IV) extraction. Greener cells are higher AUROC, redder cells lower. The qualitative pattern (some traits show pervasive off-diagonal degradation, others mostly retain transfer) replicates, with confidence and honesty showing the most consistent distance–AUROC relationship (per-trait Pearson $r = -0.44$ and -0.36 respectively).