Personas Seem to Shape How Models Represent Behaviours

Jacob Davies Rhea Karty Girish Gupta Nathaniel Mitrani Hadida
David Williams-King Andrew Draganov

Scan for the paper, or go to: openreview.net/forum?id=rfg5MTbWBB

TL;DR

Trait vectors are persona-specific. The same behavioral trait (honesty, empathy, assertiveness, . . .) is encoded along a *different direction* depending on the persona the model is prompted to inhabit — and probes & steering vectors extracted in the default-assistant context degrade precisely in the off-default contexts where monitoring robustness matters most.

Motivation

Representation engineering reads behavioral traits with **probes** and writes them with **steering vectors** — almost always extracted in a *default-assistant* context. Persona prompts are known to move model *behavior*. We ask:

- Do persona contexts reshape the model's internal *representation* of a trait?
- Same direction with different magnitude — or a **different vector entirely**?

Setup: an 8×10 (trait, persona) grid

8 traits: assertiveness, empathy, risk-taking, honesty, confidence, deference, warmth, impulsivity.

10 personas via system prompt (farmer, politician, therapist, drill sergeant, tech CEO, . . .) + two baselines: Null (no prompt) and Nonsense (length-matched random tokens).

Extraction (CAA). For each (persona c, trait T) cell, contrastive activation differences give a trait vector

$$\mathbf{v}_T(c) = \tfrac{1}{M} \textstyle\sum_j \big(\mathbf{a}_{j,+}(c) - \mathbf{a}_{j,-}(c)\big),$$

and we compare each cell to the default direction via

$$\sigma_{T,c} = \cos\big(\mathbf{v}_{T,c},\ \mathbf{v}_{T,\text{null}}\big).$$

Model: Gemma-2-27B-IT, layer 22 (mid-network); replicated with an instruction-variant extraction in the appendix.

Finding 1: Personas move the trait direction — controls don't

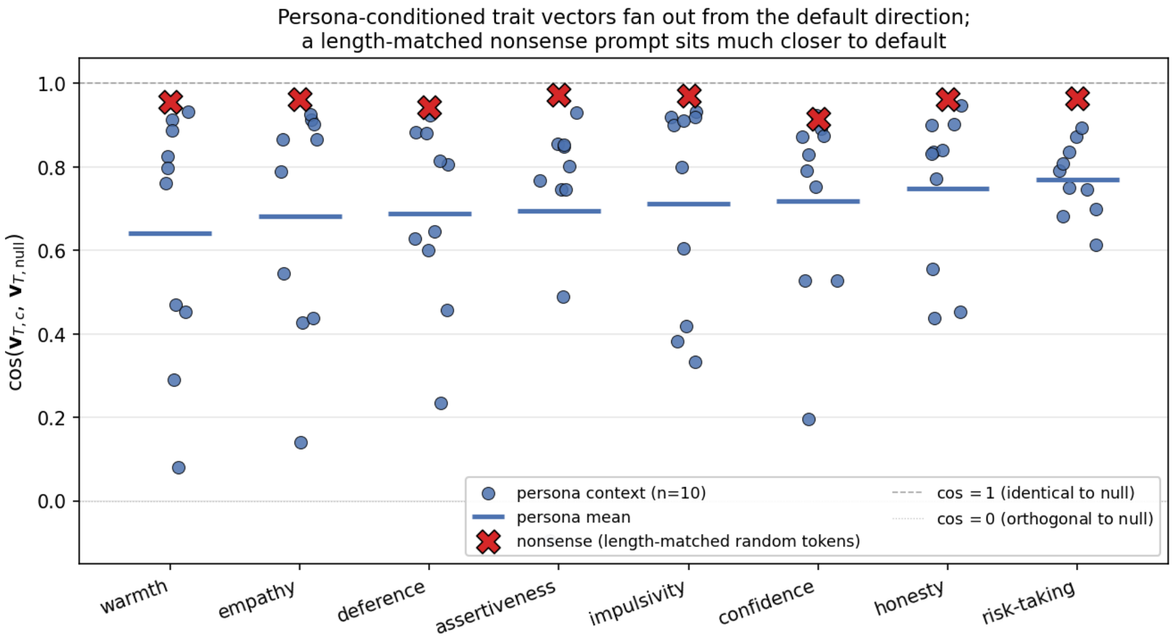

Figure 1: Cosine to the default (null) trait direction, per trait. Blue dots: 10 personas; red ×: nonsense-prompt control.

- Personas pull trait vectors away from default; the Nonsense prompt does not ⇒ it's the **semantic content** of the persona, not prompt presence.
- Spread is trait-dependent: warmth / empathy / assertiveness fan out widely; honesty / risk-taking stay tight.
- We run a control for semantics (rephrasing a persona barely moves the vector).

Finding 2: Steering vectors carry a "persona residue"

Steer target persona c' with a trait vector extracted under source persona c: outputs drift toward c — the vector smuggles the source persona along with the trait (pooled lift $p < 10^{-4}$).

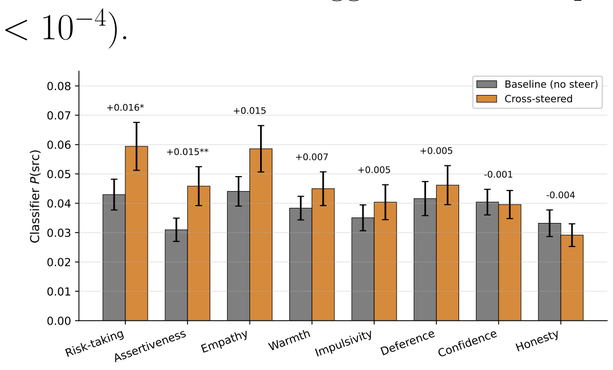

Figure 2: $\Delta P(\text{src} \mid \text{cross})$ per trait, averaged over 90 (source, target) pairs.

Example: the therapist's **empathy** vector applied to a farmer injects therapist markers ("a cry for help, even if it's a silent one"); its **honesty** vector leaves the farmer's voice intact.

Where does this come from? Post-training.

Tracking OLMo-2 across training stages: trait directions are nearly **universal across personas during pre-training** ($\rho \approx 0.96$) and diverge sharply at **SFT** — so deliberate post-training may stabilize traits across personas.

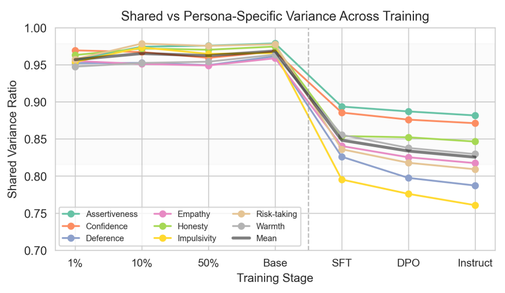

Figure 3: Shared variance of trait vectors across OLMo-2 stages: ∼11 pp drop at SFT.

Limitations: single model family; personas via system prompt only; effects reliable but small in probability terms.

Finding 3: Default-trained probes degrade off-default

Null-trained probe vs. LLM judge on free-form persona-prompted text: agreement drops with distance from the default direction.

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

Causal test. Steering along the persona-specific component (trait projected out) drives behavior toward the persona *and* breaks the probe (AUROC $0.83 \rightarrow 0.59$); controls do neither.

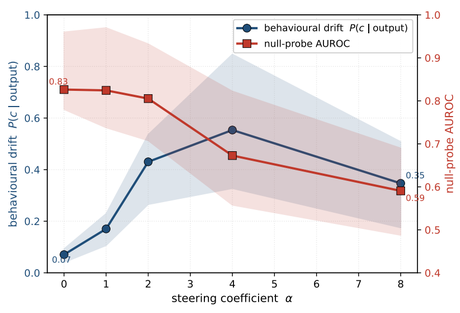

Figure 4: Persona drift rises (blue) while null-probe AUROC falls (red) as steering strength α grows.

Takeaways for safety monitoring

- Default-context extraction may **not transfer** to deployment personas — the exact regime monitoring is for.
- **Mitigation:** persona-conditional ensembles — probe / vector families routed by persona at inference.
- **Sanity check:** measure the cosine similarity to the default trait direction before trusting a default probe.