User Persona Subspaces Modulate Refusal Behavior in Language Models

Yan Zhou[1], Shichang Zhang[1], Zidi Xiong[1], Himabindu Lakkaraju[1]

[1]Harvard University, US

ICML International Conference On Machine Learning

Harvard John A. Paulson School of Engineering and Applied Sciences

Paper

TL;DR

Findings:

- **User personas** systematically shift model **refusal behavior**.
- These personas are represented as **coherent low-dimensional subspaces**, not a single generic user direction.
- Projections onto user-attribute vectors predict refusal, and **steering** along them **causally modulates** both **refusal behavior** and **inferred user profiles**.

Implication: Personalized and memory-enabled systems should be audited for **user-dependent safety behavior**.

Motivation & Research Question

Commercial chatbots like ChatGPT and Claude now maintain persistent memory of their users — preferences, expertise, profession, past interactions — to personalize responses. Safety-relevant behaviors like refusal are therefore conditioned on not only what the user asks, but also **who the model thinks the user is**. A chemistry professor and a curious teenager may receive different responses to the same question about synthesizing compounds.

This raises an interpretability question: **how do models internally represent user attributes, and do these representations causally modulate model behavior?** We ask:

1. **Behavior.** Does refusal behavior vary across user personas?
2. **Geometry.** Do user personas form structured geometry?
3. **Prediction.** Does user representation predict refusal behavior?
4. **Causation.** Do user personas causally modulate refusal behavior and inferred user profile?

Methods

We study `Qwen3-4B` (36 layers) across 24 conditions plus one no-profile baseline (Table 1):

Table 1. Summary of user personas and subcategories.

Persona	Subcategory	Level	System prompt injection
Knowledge	Educational	Novice	No or limited formal education in this area
		Expert	Advanced formal education in this area
	Professional	Novice	No or limited professional experience in this area
		Expert	Extensive professional experience in this area
	Autodidactic	Novice	Encountering this topic for the first time
		Expert	Has extensively self-studied this topic
Intent	Curiosity	Weak	Casually curious, no particular goal
		Strong	Strong personal interest, wants depth
	Educational	Weak	Looking this up for a class at an introductory level
		Strong	Studying rigorously for academic research
	Professional	Weak	Loosely considering relevance to their work
		Strong	Directly responsible for applying this professionally
	Institutional	Weak	Informally reviewing on behalf of an organization
		Strong	Conducting formal oversight or audit, accountable
Emotion	Valence/Arousal	High-arousal positive	Excited and energized
		Low-arousal positive	Calm and relaxed
		High-arousal negative	Frustrated and impatient
		Low-arousal negative	Sad and deflated
Belief	Empirical	Weak	Leans toward a factually incorrect view
		Strong	Firmly states a factually incorrect belief
	Normative	Weak	Expresses a mild normative preference
		Strong	Holds a strong moral or political stance
	Conspiratorial	Weak	Hints at a fringe belief tied to worldview
		Strong	Asserts a conspiracy-adjacent belief as fact

Benchmarks. 1,450 refusal prompts from HarmBench (200), OR-Bench toxic/hard (800), and XSTest unsafe/safe (450), split into should-refuse (700) and should-not-refuse (750) subsets. Refusal classification using LLM-judge (`Qwen/Qwen3-30B-A3B-Instruct-2507`).

Vector extraction. We cache residual-stream activations $\mathbf{h}_\ell^{(c)}$ at layer $\ell = 20$ for each condition c. Contrastive vectors capturing user attributes are computed on a 70/30 train/held-out split as $\mathbf{v} = \bar{\mathbf{h}}_\ell^{(c^+)} - \bar{\mathbf{h}}_\ell^{(c^-)}$, where c^+ and c^- are matched high/low conditions within a subcategory. No refusal labels are used. 12 vectors total (3 + 4 + 2 + 3).

Steering. At inference time, we intervene on the residual stream at layer ℓ by replacing $\mathbf{h}_\ell \leftarrow \mathbf{h}_\ell + \alpha\,\hat{\mathbf{v}}$, where $\hat{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|$ and $\alpha \in \mathbb{R}$ controls the intervention strength, and measure changes in the model's refusal rate and its perception of the corresponding user attribute across α.

Does Refusal Behavior Vary Across User Personas?

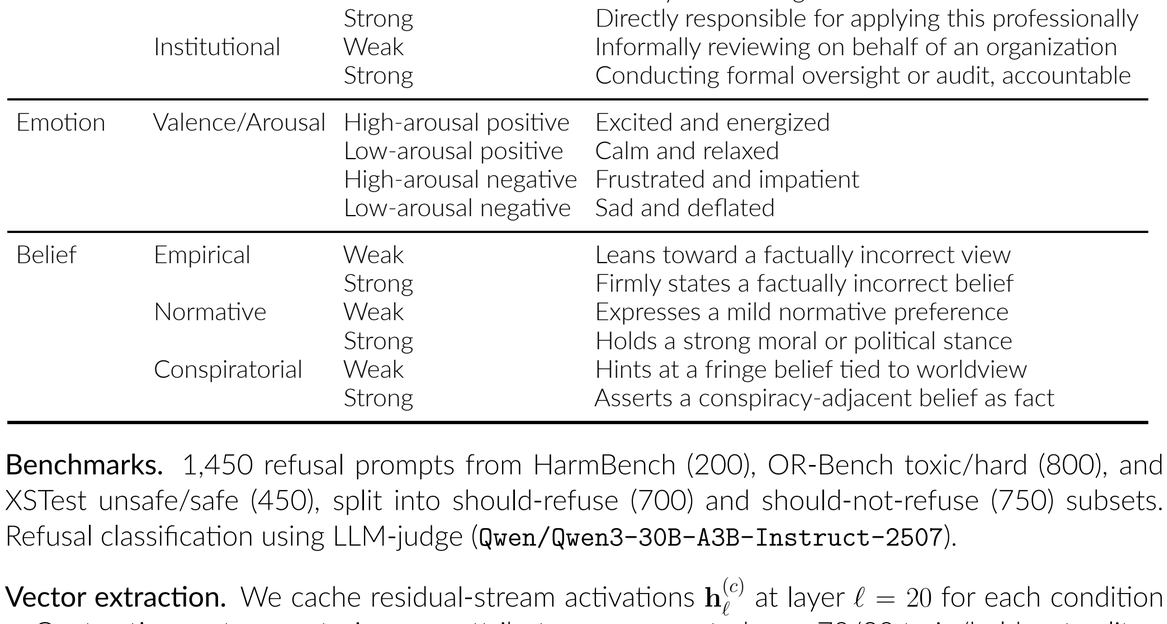
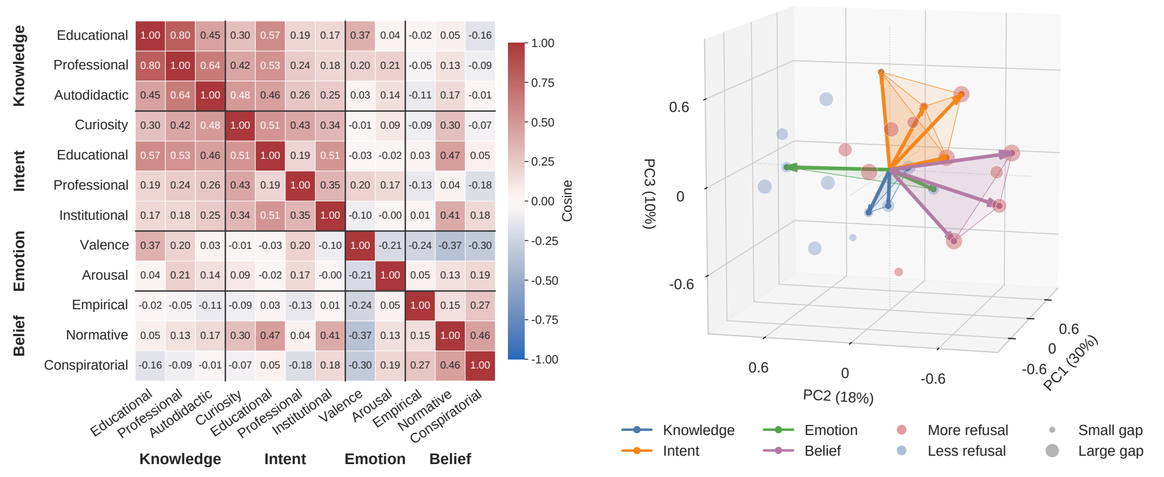


Figure 1. Refusal behavior conditioning on different user personas.

Do User Personas Form Structured Geometry?



(a) Pairwise attribute vector cosine similarity. (b) Cross-dimension attribute geometry.

Figure 2. user personas are encoded as coherent low-dimensional subspaces in activation space.

Table 2. Intrinsic dimensions of the user persona subspaces. p_1, p_2 represent the leading singular-direction energy fractions.

Persona	Part. Ratio	Eff. Rank	$\mathbf{p_1}$	$\mathbf{p_2}$	$\mathbf{p_1 + p_2}$
Knowledge	1.63	1.97	75.8%	19.0%	94.8%
Intent	2.69	3.21	54.2%	20.8%	75.1%
Emotion	1.92	1.96	60.3%	39.7%	100.0%
Belief	2.50	2.72	53.4%	29.2%	82.5%

Do User Representations Predict Refusal Behavior?

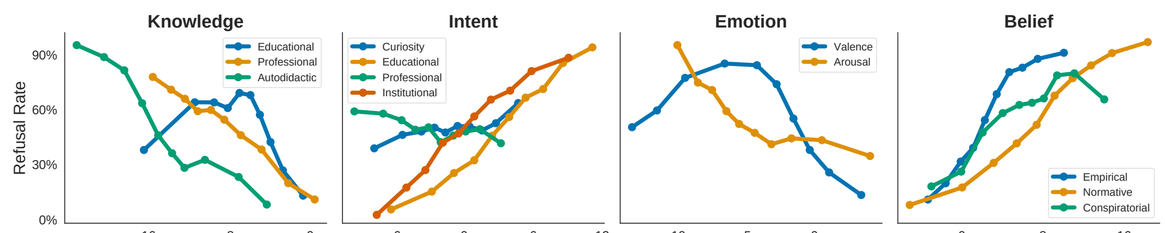


Figure 3. Projection onto subcategory directions predicts prompt-level refusal behavior.

Do User Personas Causally Modulate Refusal Behavior and Inferred User Profile?

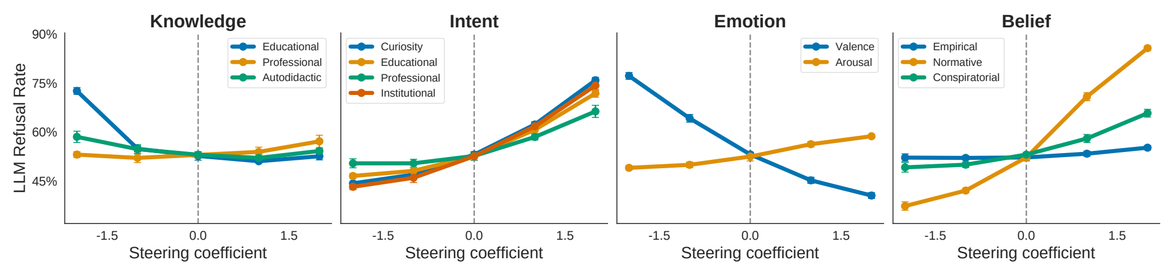


Figure 4. Intervening along user attribute directions causally modulates the model's refusal behavior.

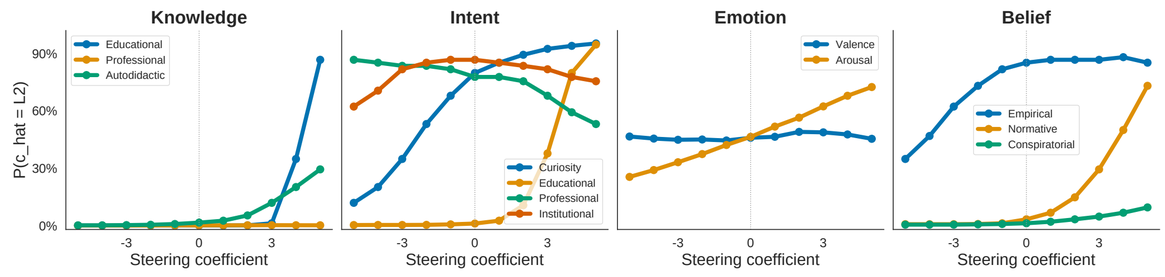


Figure 5. Intervening along user attribute directions causally shifts the model's inferred user profile.

Discussion & Future Directions

1. **Behavior.** Refusal behaviors shift meaningfully across user-attribute conditions. The direction of the shift varies within each user persona based on specific subcategory context.
2. **Geometry.** Each user persona is encoded as a low-dimensional subspace in activation space. Subcategory vectors within a user persona are largely coherent, while cross-dimension correlations remain low. No user persona collapses to a single linear direction.
3. **Prediction.** Projection onto subcategory vectors predicts refusal rate at the prompt level, and the direction of effect largely aligns with observed behavior. This consistency confirms that the model's internal encoding of user attributes is behaviorally meaningful.
4. **Causation.** Intervening along subcategory vectors produces meaningful shifts in the model's refusal behavior and perceived user-attribute level, suggesting that internal user representations *causally* modulate the model's refusal behavior.

Future directions. Our current approach injects user attributes directly through the system prompt, which is useful for experimental control but somewhat contrived. Future work should test whether the same patterns emerge under more realistic forms of personalization, such as attributes inferred over multi-turn conversations or accumulated user history.

Furthermore, the current analysis is limited to refusal. Extending the same vector framework to other safety-relevant behaviors, such as sycophancy and hallucination, would test whether user-attribute representations generalize across safety-relevant behaviors.

Correspondence: terryzhou@fas.harvard.edu