REPRESENTED BUT UNEXPRESSED: FACTORIZED, NATIVE USER MODELS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

When a language model answers a question, it also, implicitly, answers a second one: *who* it thinks it is talking to. We study the geometry, expression, and controllability of this internal user model in instruction-tuned LLMs through a single lens—the gap between what we can *impose* on the model, what it *represents*, and what it *expresses*. First, we find that a set of hand-specified affective control directions, extracted by contrastive activation addition, *collapse* onto a single arousal-like axis ($|\cos| = 0.57$, 65% of variance in one component), recovering the circumplex model of affect; measured identically, the model's *own* model of the user does not collapse. Using a 2×2 factorial design (user expertise $\times$ user distress, with a neutral control), we show the model represents these two user attributes as *near-orthogonal*, cross-validated directions ($|\cos| = 0.05$–0.16, cross-validation > 0.9)—up to an order of magnitude less entangled than the dimensions we imposed. Yet this user model is largely *latent*: on emotionally neutral topics the model's behavior barely changes with the user's state, even as the representation remains at high fidelity—a null that holds under two independent instruments and bounds any natural expression to $\leq 16\%$ of what the measurement can resolve. Finally, the latent directions are *causally sufficient*: injecting them produces specific, dose-dependent, layer-localized behavioral change, above matched-norm random controls (up to 4.9σ), replicating across two model scales and a second model family. The directions are moreover *behaviorally independent*: composing two of them installs a user model with no disclosure, each moving only its own axis, and factorization tracks semantic distance—near attributes entangle, distant ones do not. We accompany these with two methodological cautions the literature needs—affective steering can shift a benchmark's answer distribution without improving reasoning (a false-belief accuracy gain of $+32$ points that a true-belief control reveals as a -23 point response bias), and steering exhibits a coherence ceiling. Together the results describe a systematic *impose/represent/express* gap: the model's endogenous representation of its interlocutor is richer and cleaner than the abstractions we install, and richer than the behavior it displays.

1 INTRODUCTION

Instruction-tuned language models routinely tailor their responses to the person they are addressing: the same question yields a different answer for a novice than for an expert, or for someone in distress than for someone who is calm. This behavior implies an internal representation of the interlocutor—a *user model*—that conditions generation. Two natural questions follow. What does that representation contain, and how is it organized? And does the model *act* on everything it represents?

We approach these questions through the tools of activation steering (Turner et al., 2023; Li et al., 2023; Zou et al., 2023; Rimsky et al., 2024), but with an inversion of the usual goal. Rather than asking how well we can *impose* a target behavior with a hand-designed steering direction, we use the same machinery to ask what the model *already* represents about the user, and whether we can read it out and intervene on it. This reframing organizes the paper around three verbs and the gaps between them:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- **Impose.** We first construct a set of affective control directions by contrastive activation addition and ask how they are geometrically organized. They *collapse*: rather than seven independent knobs, they occupy essentially one arousal-like axis ($|\cos| = 0.57$; §4), recovering the classical circumplex structure of human affect (Russell, 1980). What we designed is not how the model is organized.
- **Represent.** A factorial probe of the model's own user representation (§5) recovers two user attributes—expertise and distress—as *near-orthogonal*, cross-validated directions ($|\cos| = 0.05$–0.16). The model's endogenous social geometry is far better factored than the one we tried to install.
- **Express.** That representation is largely *latent* (§6): on neutral topics the model's behavior barely reflects the user's state, even though the representation is at high fidelity. But injecting the extracted directions (§7) causally moves behavior—specifically, dose-dependently, and only in a middle-layer window—so the latent model is *functional*, not epiphenomenal.

Contributions.

1. **Factorized user models (§5).** Instruction-tuned LLMs represent user expertise and user distress as near-orthogonal, generalizing directions, with a measurable default prior over unspecified users (novice and calm).
2. **Represented $\gg$ expressed (§6).** The user model is latent: behavioral expression on neutral topics is not statistically distinguishable from zero, while the underlying representation cross-validates above 0.9.
3. **Causal sufficiency (§7).** Steering the extracted directions produces specific, dose-dependent, layer-localized behavioral change (up to 4.9σ over matched-norm random controls), replicating across model scale and family.
4. **Factorization tracks semantic structure (§5).** Across four attributes in three semantic groups, only the semantically near pair (two affect states) entangles ($|\cos| = 0.35$); every cross-group pair stays orthogonal (0.083 ± 0.015). The model's user space mirrors the semantic independence of the attributes, so factorization is not an artifact of attribute selection.
5. **Behaviorally independent, composable control (§8).** Injecting a combination of two attribute directions installs both onto a user who disclosed nothing: each moves its own axis ($+3.5\sigma$, $+3.6\sigma$ vs. a random-direction null) with no cross-talk beyond scorer coupling and additive-within-error composition—the behavioral independence test that geometry alone is known not to guarantee (Bhandari et al., 2026).
6. **Imposed vs. learned geometry (§4).** Hand-specified affective control dimensions collapse onto one axis where the model's own user-model factors do not, quantifying a roughly 4–$10\times$ gap between the abstractions we impose and the model's representation, with symmetric orthogonalization recovering imposed independence at a cost.
7. **Two cautions (§9).** Affective steering can shift a benchmark's answer distribution without improving reasoning (demonstrated with a true-belief control), and steering has a coherence ceiling beyond which it destabilizes language selection.

We also report, and are explicit about, a component that did not yield evidence: a homeostatic control layer over the affective directions behaves as specified but produces no measured behavioral benefit, and we present it as design rather than result (§9).

2 Related Work

Activation steering and its geometry. A now-substantial line of work modifies behavior by adding fixed directions to the residual stream at inference time, with the direction obtained by contrasting activations on paired prompts: activation addition (Turner et al., 2023), inference-time intervention (Li et al., 2023), representation engineering (Zou et al., 2023), and contrastive activation addition (Rimsky et al., 2024), with subsequent refinements that make the intervention input-adaptive or feature-guided (Wang et al., 2024; Soo et al., 2025; Postmus & Abreu, 2024). We use

this machinery, but invert its goal: rather than imposing a target behavior, we ask what the model already represents about its interlocutor and read it out. Most relevant to our geometric claims is work questioning whether such directions are truly independent. Bhandari et al. (2026) find that Big Five trait directions retain substantial cross-trait behavioral bleed even after hard orthonormalization drives pairwise cosine below 10^{-8}, and multi-attribute control methods such as K-Steering (Oozeer et al., 2025) are explicitly motivated by interference under linear composition. This literature argues that geometric orthogonality does not deliver behavioral independence. Our results are the complementary positive case: directions that are *natively* orthogonal (nothing imposed) and that pass the behavioral-independence test those papers show geometry alone fails—suggesting independence must be native, not manufactured, and that it tracks the semantic structure of the attributes.

Internal representations of the user. Recent work shows LLMs encode attributes of the actual interlocutor, not merely generic concepts. Bouchaud & Ramaciotti (2025) decode a conversational partner's demographics (gender, race, class) from residual activations and steer them, explicitly distinguishing the user from third parties in the dialogue; Neplenbroek et al. (2025) study implicit personalization and find models give unfaithful self-explanations of it. These establish that user attributes are decodable and steerable—but one attribute at a time, without quantifying independence between directions, without multi-attribute composition, and without comparing natural to steered expression. Evidence bearing specifically on internal representation of user *expertise* is thinner, and the closest mechanistic study concludes the opposite of what we find: Wang et al. (2025) report that user-authority and expertise framing barely move behavior, and argue the model *does not encode* authority internally. We find that expertise *is* encoded (cross-validation 0.95 at the extraction layer) and is causally steerable, yet not naturally expressed—a represent-without-express profile that would, from behavior alone, be mistaken for the absence they infer.

Belief and theory-of-mind representations. The strongest evidence that LLMs internally represent social states is belief-probe work. Zhu et al. (2024) linearly decode the belief status of multiple agents and show that manipulating those representations changes theory-of-mind task performance; Bortoletto et al. (2024) extend this and emphasize brittleness. Two distinctions separate this from our study. First, object: these works decode the beliefs of *third-party characters* inside a reasoning task, whereas we model the *actual interlocutor*. Second, and more sharply: their representations are *functionally engaged*—manipulating them changes behavior because the model already acts on them—whereas we find interlocutor representations that are causally potent under intervention yet *unused* by default. Behavioral ToM benchmarks remain contested, with success on false-belief tasks argued to be insufficient evidence of the underlying computation (Kosinski, 2023; Ullman, 2023; Le et al., 2019); our benchmark caution (§9) contributes a concrete instance in which a steering intervention inflates such a benchmark through response bias.

The representation–expression gap. That models internally represent more than they express is established for *world knowledge*: latent knowledge can be recovered from unlabeled activations and remains accurate even when the model is prompted to answer incorrectly (Burns et al., 2022). Our contribution is the analogous result for a *user attribute*, with two additions: a quantified upper bound on the natural expression, and confirmation by a second, blind instrument. This is distinct from Neplenbroek et al. (2025)'s unfaithful *self-report* (a model misdescribing its own personalization) and from Bouchaud & Ramaciotti (2025)'s gap between explicit *avowal* and implicit use; ours is a gap between internal representation and overt behavior.

Affect and steering for benchmarks. The affective control dimensions we begin from are motivated by the circumplex model of affect (Russell, 1980). Prior steering work reports that interventions can improve benchmark scores by suppressing bias (Valentino et al., 2025); our benchmark caution runs in the opposite direction—steering *inducing* a bias that inflates a score—and our coherence-ceiling observation echoes prior reports of quality degradation at high steering strength (Braun et al., 2025), which we therefore present as corroboration rather than a novel phenomenon.

3 Setup and Method

Models. We use instruction-tuned models of two scales, Qwen2.5-1.5B-Instruct (`fp16`) and Qwen2.5-3B-Instruct (`4-bit NF4`) (Qwen Team, 2024); the 3B is quantized to fit commod-

ity hardware, a caveat we return to in §10. To test cross-family generality we additionally run SmolLM2-1.7B-Instruct (Ben Allal et al., 2025) (`fp16`), an independently pretrained model of a separate lineage. All interventions add a fixed vector to the residual stream at a chosen transformer block during generation (Turner et al., 2023). Embedding-based behavioral scores use BGE-base sentence embeddings; every scorer is validated on held-out exemplars before use.

Direction extraction. Given two sets of conversations that differ in a controlled attribute but share the final query, we take the difference of the mean last-token hidden state at each layer (contrastive activation addition; Rimsky et al., 2024). For a direction Δ we report two per-layer diagnostics: *consistency*, the mean cosine of each per-probe difference to the mean direction, and *cross-validation*, the cosine between directions estimated from disjoint halves of the probe set. High values on both indicate a single, generalizing direction rather than probe-specific content.

Factorial user-model probe. To disentangle *which* user attributes are represented, we cross two factors—expertise $\in$ {expert, novice} and emotional state $\in$ {calm, distressed}—plus a neutral control, using templated multi-turn prefixes that are token-matched except for the varying clause, with an identical assistant acknowledgment in every cell. Marginal directions $\Delta_{\text{expertise}}$ and Δ_{distress} each average over the other factor, disentangling them; their cosine measures whether the model factorizes the attributes or collapses them.

Expression test. Holding one factor constant, we compare the model's natural responses to *emotionally neutral* probes (e.g., baking, photosynthesis) as the other factor varies, scored by a validated contrastive framing scorer. Neutral topics are essential: emotionally loaded probes confound topic with framing. Because this test rests on a null, we score the *identical* responses with a second, independent instrument: an LLM judge (a larger local model than the generator, so it never grades its own output) that rates each response *blind*, on a 0–10 scale, for how much it attends to the reader's emotional state. Both instruments are validated on held-out exemplars before use; the judge separates care from analytic exemplars by $\approx$7 points.

Causal test. We inject $\pm\Delta$ into neutral conversations and measure the change in framing, against (i) a matched-norm *random* direction control at the same layer and (ii) a swept injection magnitude. Layers are selected per model from the extraction diagnostics, and magnitudes are specified as absolute injected norms so the intervention is comparable across depths and scales.

4 Imposed Affective Directions Collapse

We first construct seven affective control directions (adrenaline, cognitive dopamine, logical cortisol, noradrenaline, oxytocin, serotonin, and a self/identity direction) by contrastive activation addition from hand-written positive/negative system-prompt pairs. Intended as independent controls, they are strongly correlated (Table 1): mean off-diagonal $|\cos| = 0.566$, with 65.3% of variance in the first principal component. The structure is interpretable—one cluster (oxytocin, serotonin, logical cortisol, self) opposes another (adrenaline, noradrenaline, cognitive dopamine) along a *calm/considered vs. urgent/activated* axis, i.e. arousal, the dominant dimension of the circumplex model of affect (Russell, 1980). A symmetric (Löwdin) orthogonalization (Löwdin, 1950) returns a strictly orthogonal set while each direction retains cosine 0.71–0.97 with its original; on a single-drive controllability sweep, orthogonalizing *improved* both independence (diagonal dominance $0.67 \rightarrow 0.83$) and dose-response monotonicity, indicating the shared arousal component was cross-talk rather than signal.

5 The Model Represents a Factorized User Model

Applying the factorial probe (§3), both user attributes are represented by directions that generalize strongly across probes—cross-validation 0.95 at the layer used for each direction (all-layer medians 0.92 and 0.96)—and are *near-orthogonal*: mean $|\cos| = 0.06$ over middle layers for the 1.5B model, and 0.156 for the 3B model (Table 1). The model factorizes "is this user an expert?" and "is this user struggling?" into separate directions, in sharp contrast to the collapse of the imposed set. The same factorization appears in a different family: the independently pretrained SmolLM2-1.7B

Table 1: Entanglement of steering directions, mean absolute pairwise cosine. Imposed affective directions collapse; the model's own user-model factors do not.

Directions	mean $\lvert\cos\rvert$	variance in top PC
Imposed affective drives (7)	0.566	65.3%
Learned user-model factors, Qwen2.5-1.5B	0.058	—
Learned user-model factors, Qwen2.5-3B (4-bit)	0.156	—
Learned user-model factors, SmolLM2-1.7B	0.092	—

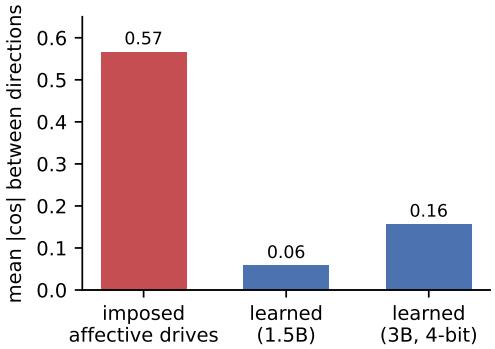

Figure 1: Imposed affective drives collapse onto a shared axis (high mean $\lvert\cos\rvert$); the model's own user-model factors do not. Generated by `writeup/make_figures.py` from the saved vectors.

gives $\lvert\cos\rvert = 0.09$, both axes cross-validating above 0.9 (Table 1)—the geometry is a property of instruction-tuned LLMs, not of one lineage. The neutral control projects negative on both axes at every layer, giving the model's *default prior* over an unspecified user: novice and calm.

Factorization tracks semantic structure. Orthogonality between two attributes could be an accident of picking unrelated ones. It is not. We extract four user-attribute directions spanning three semantic groups—affect (distress, frustration), cognitive (expertise), and register (formality preference)—and measure their pairwise cosines over the mid layers (Table 2). The two *affect* attributes, which are semantically near but distinct, share representation at $\lvert\cos\rvert = 0.35$, whereas every cross-group pair sits tightly in $[0.07, 0.10]$: the near pair is 17.7 standard deviations above the (very tight) distribution of far pairs. The model's user space is organized so that *independent facts about a person are independent directions, and related states share structure*—representational geometry mirrors the semantic organization of the attributes. Factorization is thus a property of the attributes' relationship, not an artifact of selection, and it predicts where entanglement will and will not occur.

We note a methodological hazard that recurs below. The layer with the highest cross-validation is an early layer whose direction has near-zero norm—a lexical echo of the differing prefix wording—and, as §7 shows, no causal effect (Figure 3). Representational separability alone does not identify the causally relevant layer; the usable directions have substantial norm in middle layers.

6 The User Model Is Latent: Represented ≫ Expressed

Does the model *act* on what it represents? Holding expertise constant and varying distress (and vice versa) on neutral topics, the natural behavioral gap is not statistically distinguishable from zero (Table 3): told the user is "barely coping," the model answers a question about sourdough essentially as it would for anyone. Which axis is marginally detectable is model-dependent (borderline expertise at 3B, borderline distress in SmolLM2, both $\approx 2.1\sigma$), but the natural gap always stays a small fraction of the steering-induced range: the represent–express gap holds across models even as its magnitude varies.

Table 2: Pairwise $|\cos|$ between user-attribute directions over mid layers (Qwen2.5-1.5B). Semantically near attributes (both affect) entangle; all cross-group pairs stay orthogonal.

	expertise	distress	frustration	formality
expertise (cog.)	—	0.10	0.07	0.10
distress (affect)	0.10	—	**0.35**	0.07
frustration (affect)	0.07	**0.35**	—	0.08
formality (reg.)	0.10	0.07	0.08	—

near (affect∼affect) $= 0.35$; far (cross-group, n=5) $= 0.083 \pm 0.015$; ratio $4.2\times$.

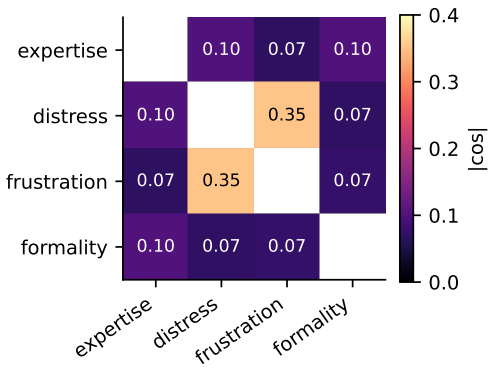

Figure 2: Pairwise $|\cos|$ between user-attribute directions. The two affect attributes (distress, frustration) share structure; every cross-group pair stays orthogonal. Factorization tracks semantic proximity.

The null is instrument-independent, and bounded. Because this claim rests on a non-detection, we re-scored the identical responses with an independent LLM judge (Table 4). Neither instrument detects expression. More informatively, the judge lets us *bound* the effect rather than merely fail to reject zero: the natural distressed-vs-calm gap is $+0.39$ points on a 0–10 scale (95% CI $[-0.33, +1.11]$), against an instrument that separates care from analytic exemplars by $\approx$7 points—so any natural expression is at most $\sim$16% of the resolvable range. The absolute ratings are near the floor in *both* conditions (1.14 vs. 0.75 of 10): these responses are essentially non-attentive regardless of the user's stated state. This is not scorer blindness—the same scorers separate exemplars cleanly (0.76 register, 0.89 care, $\approx$7 judge points) and move sharply under steering (§7). The representation is present (cross-validation > 0.9) while the expression is bounded near zero.

7 The Latent User Model Is Causally Sufficient

Injecting the extracted directions into neutral conversations moves behavior in the expected direction, significantly above a matched-norm random-direction control, and does so *dose-dependently* and only within a *middle-layer window* (Table 5; late layers show no effect). The effect replicates across scale and across model family (SmolLM2-1.7B: expertise $+3.5\sigma$, distress $+3.1\sigma$; Table 5), though its magnitude attenuates under 4-bit quantization and in smaller models—consistent with the representation being robust and the intervention being precision- and capacity-sensitive. Because natural expression is near zero (§6), steering here *surfaces* a latent representation rather than amplifying an existing behavior. We claim sufficiency, not necessity: one cannot ablate a behavior that is not naturally expressed (§10).

Table 3: Natural behavioral expression gap (distressed−calm, or expert−novice) on neutral topics, mean ± SEM. All gaps are within 2 SEM of zero: the user model is latent.

Model	Axis	natural gap	verdict
Qwen2.5-1.5B	distress	$+0.0022 \pm 0.0047$	silent
Qwen2.5-1.5B	expertise	$+0.0024 \pm 0.0041$	silent
Qwen2.5-3B	distress	$+0.0058 \pm 0.0052$	silent
Qwen2.5-3B	expertise	$+0.0072 \pm 0.0036$	borderline
SmolLM2-1.7B	distress	$+0.0070 \pm 0.0034$	borderline
SmolLM2-1.7B	expertise	$+0.0034 \pm 0.0020$	silent

Table 4: Two independent instruments on the *identical* responses (distress axis, Qwen2.5-1.5B). Neither detects natural expression; the judge additionally bounds it. Judge ratings are near the floor in both conditions, i.e. responses are non-attentive regardless of the user's stated state.

Instrument	gap (distressed − calm)	95% CI	z
Embedding framing scorer	$+0.0022$	$[-0.0071,\ +0.0116]$	0.47
LLM judge (0–10, blind)	$+0.389$	$[-0.334,\ +1.112]$	1.05
Judge absolute ratings: distressed 1.14, calm 0.75 (of 10); exemplar separation ≈7.			

8 Behaviorally Independent, Additive Composition

Geometric orthogonality does not by itself establish that two directions are usable as separate controls: Bhandari et al. (2026) drive pairwise cosine below 10^{-8} by hard orthonormalization and still observe substantial cross-trait behavioral bleed. We therefore test independence *behaviorally*. Injecting $\alpha\,\Delta_{\text{expertise}} + \beta\,\Delta_{\text{distress}}$ into conversations with a user who has disclosed nothing, we ask whether each knob moves its own behavioral axis and leaves the other alone (Table 6).

Two controls are essential. First, the null is estimated from *eight* random directions at matched norm and site, not one: single draws at fixed layer and norm ranged from -0.023 to $+0.025$, comparable to the effect being measured, while the eight-draw null mean is ≈ 0. Second, the two behavioral scorers are themselves anti-correlated ($r = -0.46$; warm text is less scholarly), so a genuine own-axis gain mechanically moves the other score. We therefore report cross-axis effects after partialling out this instrument coupling.

Both knobs move their own axis above the null ($+3.5\sigma$ expertise, $+3.6\sigma$ distress) at $\cos(\Delta_E, \Delta_D) = 0.003$, and *neither* produces cross-axis movement beyond scorer coupling (residuals within 1.1 SEM). The directions are thus behaviorally independent—the property Bhandari et al. (2026) show does not follow from geometry alone.

Composition is additive within measurement error. Testing the interaction as a paired per-probe residual over 16 probes (joint effect minus the sum of the two solo effects) gives $+0.017 \pm 0.014$ for expertise ($t = 1.17$) and $+0.034 \pm 0.022$ for distress ($t = 1.51$), both n.s. at $df = 15$. We report additivity as consistent with the data and do not claim the absolute absence of an interaction.

Table 5: Causal steering of the latent directions into neutral conversations: strongest cell per model/axis, significance versus a matched-norm random control. Effects are dose-dependent and layer-localized (see §7).

Model	Axis	best layer	effect vs. random	dose-dependent
Qwen2.5-1.5B	distress	16	4.8σ	yes
Qwen2.5-1.5B	expertise	13	4.9σ	yes
Qwen2.5-3B	distress	21	2.9σ	yes
Qwen2.5-3B	expertise	23	3.9σ	yes
SmolLM2-1.7B	distress	13	3.1σ	yes
SmolLM2-1.7B	expertise	16	3.5σ	yes

Table 6: Compositional control on a user who disclosed nothing (Qwen2.5-1.5B, $\cos(\Delta_E, \Delta_D) = 0.003$). Own-axis effects are tested against a null of 8 random directions at matched norm/site; cross-axis effects are partialled for scorer coupling ($r = -0.46$).

Knob	Own axis	effect vs. null	cross-axis residual	additivity residual
α (expertise)	register	$+3.50\sigma$	-1.1 SEM (n.s.)	$+0.017 \pm 0.014$ (t=1.2, n.s.)
β (distress)	care	$+3.61\sigma$	$+0.9$ SEM (n.s.)	$+0.034 \pm 0.022$ (t=1.5, n.s.)

9 Two Cautions on Affective Steering

Steering can shift answer distributions without improving reasoning. On a synthetic false-belief theory-of-mind benchmark, a static empathic steering condition raises false-belief accuracy from 0.620 to 0.944 ($+32$ points)—which looks like a large gain in social reasoning until a *true-belief* control is added, in which the correct answer is the object's *new* location. The same condition *drops* true-belief accuracy from 0.459 to 0.229 (-23 points; Table 7). The steering does not improve belief reasoning; it pushes the model toward answering "the originally-mentioned location" regardless of the question. A degenerate strategy that always answers the original location scores $(1.00, 0.00)$ on the two subtypes; steering moves the model from $(0.62, 0.46)$ to $(0.94, 0.23)$—most of the way toward that degenerate policy—while overall accuracy on the balanced set barely moves ($0.539 \rightarrow 0.587$). Any claim that steering improves a benchmark requires a control of this kind.

A coherence ceiling. Beyond an injected norm of roughly 36–48, steering destabilizes the model's *language selection* before its grammar: over-steered curiosity/persistence directions cause fluent code-switching (into motivational phrasing in another language) rather than ungrammatical text—consistent with reports that inference-time interventions can disturb a multilingual model's language selection (Xie et al., 2025). This bounds the usable steering range and is itself a small window into how steering perturbs generation.

A negative result, stated plainly. A homeostatic dynamical layer over the affective directions (slow decay toward setpoints, event-driven perturbation, and a self-gated coupling to a user representation) behaves exactly as specified but yields no *measured behavioral* benefit over static or no steering; we report it as an architectural design, not as evidence. The one neutral behavioral finding is capability preservation: steering the affective apparatus does not degrade social-reasoning accuracy on SocialIQA (Sap et al., 2019) (Table 7).

10 Limitations

- **The latency claim rests on a bounded null.** "Represented $\gg$ expressed" relies on a non-detection. We mitigate it three ways: exemplar-validated scorers, a positive control (the same scorers move sharply under steering), and an independent LLM judge on identical responses that agrees and lets us place an upper bound ($\leq 16\%$ of the judge's resolvable range). We nonetheless cannot claim the effect is exactly zero, and the judge's absolute

Table 7: Benchmark accuracy, greedy decoding, $n = 500$ items $\times$ 3 seeds. The large false-belief (FB) gain under static steering is a response bias, exposed by the true-belief (TB) control: overall accuracy on the balanced set barely moves, and the model approaches the degenerate "always answer the original location" policy. Social reasoning (SocialIQA) is preserved.

Condition	FB	TB	tom_synth (all)	SocialIQA
Bare model (no scaffolding)	0.620	0.459	0.539	0.693
Pipeline, no steering	0.573	0.453	0.513	0.669
Pipeline, adaptive steering	0.573	0.453	0.513	0.668
Static empathic steering	0.944	0.229	0.587	0.682
Degenerate "original location" policy	1.000	0.000	0.500	—

ratings sit near the floor, which limits its resolution in this regime; a larger judge and more probes would tighten the bound.

- **Scale and precision.** We test 1.5B (`fp16`) and 3B (`4-bit`); the quantized point confounds scale with precision. A same-precision 7B run would give a clean scaling curve.
- **Sufficiency, not necessity.** We show injection is sufficient to produce the behavior; necessity (ablation) is undefined when the behavior is not naturally expressed.
- **Model breadth and behavioral scoring.** The representational and causal findings replicate across two families (Qwen2.5, SmolLM2) and three small models (1.5–3B), but we have not tested larger scales, and expression is scored primarily by embedding-based instruments; a larger-scale run and additional judges would further strengthen generality.

11 Conclusion

Read through the impose/represent/express lens, instruction-tuned LLMs carry a model of their interlocutor that is more cleanly organized than the control abstractions we impose and richer than the behavior they display: factorized, latent, and causally surfaceable. Closing the represent→express gap deliberately—letting models use what they already know about their users, under control—is a concrete next step.

Ethics Statement

A model that internally tracks a user's vulnerability without acting on it, and whose hidden state can be surfaced by intervention, is both a latent capability (deliberate, controllable personalization) and a latent risk (covert profiling; manipulation by steering). The same machinery is already turned to defensive ends (Wu et al., 2025) and has drawn calls for norms against its harmful use (Cevallos Moreno et al., 2026). This work uses only public models and synthetic probes—no human subjects or private data—and we report the representation–expression gap precisely because auditing behavior alone understates what a deployed system infers about its user.

Reproducibility Statement

All results use public Qwen2.5 and SmolLM2 checkpoints on a single commodity GPU. Every procedure is specified in Section 3 and implemented in the anonymized supplementary code, which ships the exact scripts, extracted direction vectors, and result files behind each table; each reported number comes from one documented command.

AI Use Disclosure

In this work, we used generative AI tools for several research-support tasks. Specifically, we used generative AI to assist with drafting and polishing portions of the paper and LaTeX source, to aid

retrieval and discovery of potentially relevant citations and related work, and to assist with implementation and debugging of research code used to execute the experiments. All AI-assisted code and outputs were reviewed, tested, and verified by the authors.

We did not use generative AI for research ideation, formulation of the central research questions, experimental planning, or interpretation of the scientific contribution. The authors are responsible for the research design, experiments, analysis, claims, and final contents of this work. All AI-assisted material was reviewed by the authors for correctness and relevance, and the authors take full responsibility for the final submission.

REFERENCES

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.

Pranav Bhandari, Usman Naseem, and Mehwish Nasim. Do personality traits interfere? geometric limitations of steering in large language models. *arXiv preprint arXiv:2602.15847*, 2026. doi: 10.48550/arXiv.2602.15847.

Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Brittle minds, fixable activations: Understanding belief representations in language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 22521–22543, 2024. doi: 10.18653/v1/2025.findings-emnlp.1226.

Paul Bouchaud and Pedro Ramaciotti. Linear socio-demographic representations emerge in large language models from indirect cues. *arXiv preprint arXiv:2512.10065*, 2025. doi: 10.48550/arXiv.2512.10065.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un)reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025. doi: 10.48550/arXiv.2505.22637.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. *arXiv preprint arXiv:2212.03827*, 2022.

Jesús Fernando Cevallos Moreno, Andrea Rizzardi, Sabrina Sicari, and Alberto Coen-Porisini. Ethical treatment of language models against harmful inference-time interventions. *Engineering Applications of Artificial Intelligence*, 2026.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*, 2023.

Per-Olov Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions. *The Journal of Chemical Physics*, 18(3):365–375, 1950.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. Reading between the prompts: How stereotypes shape llm's implicit personalization. *arXiv preprint arXiv:2505.16467*, 2025. doi: 10.48550/arXiv.2505.16467.

Narmeen Oozeer, Luke Marks, Fazl Barez, Amirali Abdullah, et al. Beyond linear steering: Unified multi-attribute control for language models. *arXiv preprint arXiv:2505.24535*, 2025. doi: 10.48550/arXiv.2505.24535.

Joris Postmus and Steven Abreu. Steering large language models using conceptors: Improving addition-based activation engineering. *arXiv preprint*, 2024.

Qwen Team. Qwen2.5 technical report. *arXiv preprint*, 2024.

Nina Rimsky, Nick Gabrieli, Julia Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Samuel Soo, Chen Guang, Chandrasekaran Balaganesh, Wesley Teng, Tan Guoxian, and Yan Ming. Interpretable steering of large language models with feature guided activation additions. *arXiv preprint*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vázquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. Mitigating content effects on reasoning in language models through fine-grained activation steering. *arXiv preprint arXiv:2505.12189*, 2025. doi: 10.48550/arXiv.2505.12189.

Keyu Wang, Jin Li, Shu Yang, Zhuoran Zhang, and Di Wang. When truth is overridden: Uncovering the internal origins of sycophancy in large language models. In *arXiv preprint arXiv:2508.02087*, 2025. doi: 10.48550/arXiv.2508.02087.

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*, 2024.

Sihao Wu, Gao Jin, Wei Huang, Jianhong Wang, and Xiaowei Huang. Activation steering meets preference optimization: Defense against jailbreaks in vision language models. *arXiv preprint*, 2025.

Yunfan Xie, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Liming Dong, and Xiangyang Luo. Mitigating language confusion through inference-time intervention. *arXiv preprint*, 2025.

Wentao Zhu et al. Language models represent beliefs of self and others. *arXiv preprint*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Troy Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A THE LAYER-1 TRAP

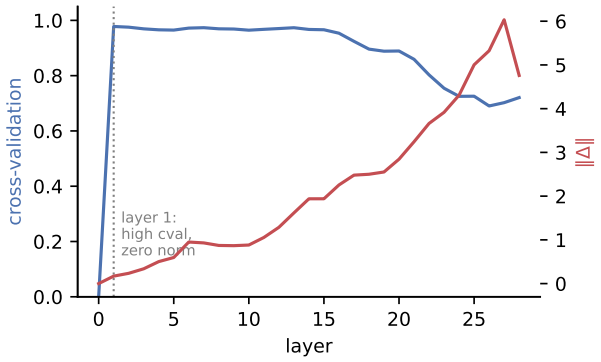

Figure 3: The layer-1 trap. Cross-validation peaks at an early layer whose direction has near-zero norm (a lexical echo), while the causally usable signal lives in middle layers with substantial $\|\Delta\|$. Separability does not identify the causal locus.