Empathy Is Steerable but Multi-Axial: Mechanism Geometry and Persona Effects in LLMs

JuHeon Ha Byounghan Lee Yunseo Choi Kyung-Ah Sohn*
Department of Artificial Intelligence, Ajou University
{hjhj97, qudgks96, chldbstj0512, kasohn}@ajou.ac.kr

Abstract

Activation steering has been used to control traits such as honesty, refusal, and sycophancy, yet supportive empathy is evaluated along multiple dimensions that need not correspond to independently controllable activation directions. Using the EPITOME framework, which decomposes supportive empathy into Emotional Reactions, Interpretations, and Explorations, we study three instruction-tuned LLMs and ask whether candidate directions derived from these labels produce distinguishable intervention effects or instead share structure, and how persona prompts interact with those directions. We find that contrastive activation addition yields a stable middle-layer intervention that consistently shifts the EPITOME proxy scores across models, moving empathy analysis beyond response-level scoring. However, the recovered directions are only partially separable: steering one direction induces off-target shifts, and hand-crafted prompting shifts the empathy profile rather than isolating a single dimension. Persona prompts substantially change EPITOME scores, but a paired activation-shift decomposition shows the recovered subspace captures only about 3% of persona-induced squared activation-shift magnitude at layer 15. Under this EPITOME-based definition, expressed empathy is steerable but multi-axial, and controlling persona-conditioned empathy requires targeting structure beyond individual mechanism directions.[1]

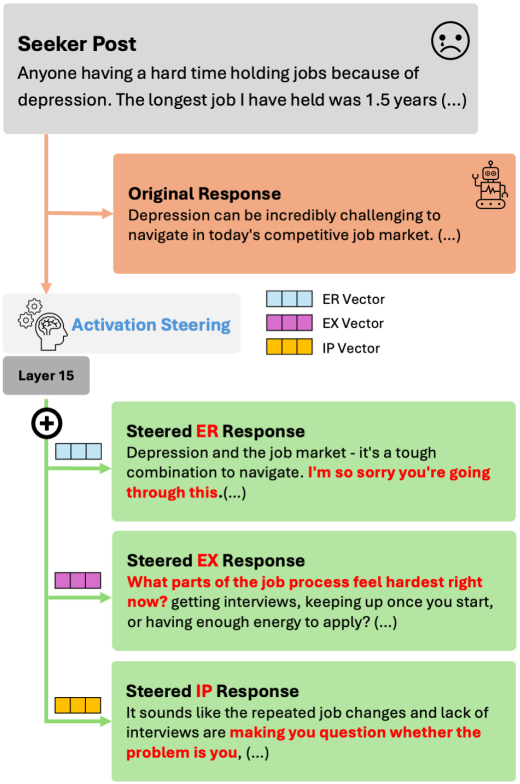


Figure 1: Example of mechanism-level steering on a support-seeking post. Highlighted phrases mark the mechanism-characteristic content introduced by each steering vector.

1 Introduction

Conversational LLMs are evaluated as emotional-support systems, and their responses are rated as more empathic than human-written ones (Sharma et al., 2020; Lee et al., 2024b). But a high empathy rating says little about *how* a response is empathic. Empathy is not one thing: it includes affective resonance with what the seeker feels, cognitive understanding of their situation, and communicative moves that invite elaboration (Sharma et al., 2020; Lee et al., 2024a). A response can realize one component while failing another, so an aggregate score reveals little about how a model is empathic. Fine-grained control matters because a system may need to strengthen one supportive behavior without amplifying another; activation steering provides an inference-time control channel without retraining or adding empathy instructions to the user-facing prompt.

*Corresponding author
[1]Code is available in our GitHub repository: github.com/hjhj97/Empathy-Is-Steerable-but-Multi-Axial

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 20326–20343
October 24-29, 2026 ©2026 Association for Computational Linguistics

Empathy is abstract and partly subjective, so analyzing it mechanistically requires decomposing broad empathy judgments into discrete, behaviorally anchored components. In this work, we use *empathy* to refer to EPITOME-defined supportive empathy in peer-support dialogue. EPITOME (Sharma et al., 2020) defines three communication mechanisms — Emotional Reactions (ER), Interpretations (IP), and Explorations (EX) — each scored on a 0/1/2 scale by a classifier trained on human-annotated support dialogues. ER captures emotional acknowledgment, IP captures inferred understanding of the seeker's situation, and EX captures questions or invitations to elaborate; Figure 1 illustrates these categories: ER steering adds emotional acknowledgment, IP steering adds an interpretation, and EX steering adds a follow-up question. However, separating empathy into three evaluation categories does not imply that the model treats them independently. Response-level scores show co-variation, not independent control; we therefore intervene along each candidate direction and measure target and non-target effects.

For other high-level traits, such an intervention is available. Activation steering localizes a behavior as a direction in the residual stream and adds or subtracts it at inference to control that behavior (Rimsky et al., 2024; Turner et al., 2024), and the approach has been validated for honesty, refusal, sycophancy, and demographic bias (Chen et al., 2025; Li et al., 2025). Here, steering is not merely a way to raise empathy scores, but a causal probe of representational organization. Prior work has studied generated responses under prompting, fine-tuning, and persona conditioning (Lee et al., 2024b; Hou et al., 2025; Malik et al., 2025), but not whether empathy is internally steerable as one quantity or as multiple interacting mechanisms.

Persona makes this question concrete: persona prompts can make models express systematically different levels of empathy depending on the specified demographic (Hou et al., 2025; Malik et al., 2025). Whether a persona shifts empathy *uniformly* or redistributes it across ER, IP, and EX determines whether a single steering direction can close persona-conditioned empathy gaps.

Within this framework, we first establish that contrastive activation addition yields reliable empathy-shifting directions with a consistent middle-layer intervention point across models. Building on this, we ask two questions. **(RQ1)** *Do ER, IP, and EX occupy separable representational directions, or do they share subspace structure that limits independent control?* **(RQ2)** *Do persona prompts shift empathy along the recovered mechanism directions, and how much of persona-induced empathy variation is captured by the recovered ER/IP/EX mechanism basis?* Both questions require intervening on the representation rather than scoring outputs, so we extract a steering vector per mechanism and combine layer-sweep, geometric, and projection analyses across three open instruction-tuned models.

Our findings complicate the natural single-axis reading. **For RQ1**, the recovered ER/IP/EX directions are not independently separable but share representational structure and produce off-target shifts. **For RQ2**, persona prompts produce large empathy-score changes without comparable movement along the recovered mechanism directions; the ER/IP/EX subspace captures only about 3% of persona-mean squared activation-shift magnitude at layer 15. We make three main contributions:

- **A cross-model activation-level handle on empathy.** A layer sweep across three open LLMs identifies a consistent middle-layer steering point, robust under a second automated evaluator and showing partial out-of-domain transfer, moving empathy from response-level scoring to a tractable target for mechanistic analysis.
- **Expressed empathy is multi-axial in LLM representations, not a single controllable trait.** Geometry analysis shows that the three mechanism directions share representational structure rather than forming independent axes, producing measurable behavioral spillovers across mechanisms that prompting alone cannot isolate.
- **Persona-induced empathy bias is largely outside the recovered mechanism basis.** Large changes in persona-conditioned empathy scores correspond to only small projection shifts along the recovered mechanism directions, positioning single-vector debiasing as insufficient for closing persona-induced empathy gaps.

2 Related Work

Activation steering. Activation steering modifies model behavior by adding fixed direction vectors to residual-stream activations at inference. Turner et al. (2024) introduced ActAdd using contrastive prompt pairs, and Rimsky et al. (2024) formalized Contrastive Activation Addition (CAA) with the last-token mean-difference recipe we fol-

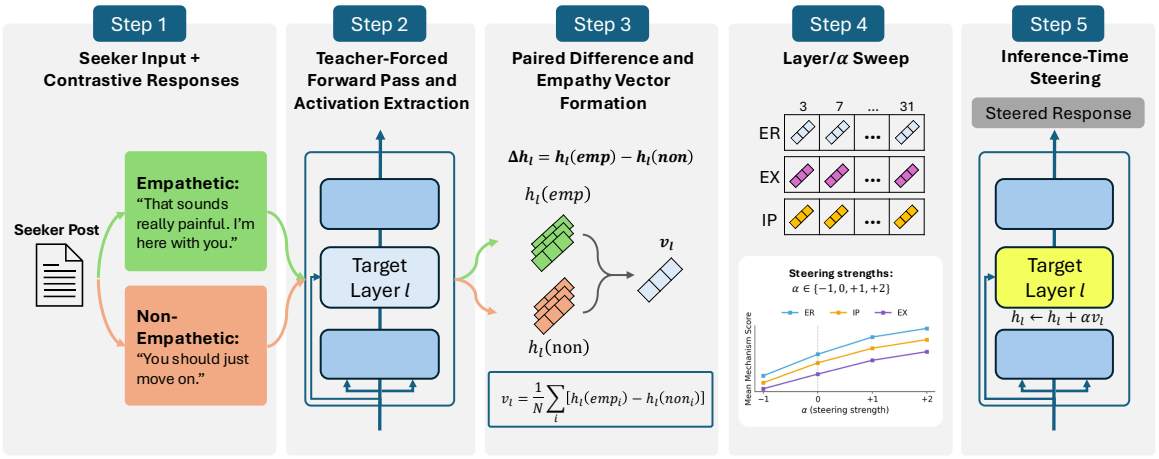


Figure 2: Empathy vector extraction and inference-time steering. (1) **Contrastive examples.** For each EPITOME mechanism, we sample high- and low-empathy seeker–response rows from the filtered data. (2) **Teacher-forced activation extraction.** We run each seeker–response sequence through the model and read the layer-ℓ final-response-token residual state. (3) **Empathy-vector formation.** We form v_ℓ^m by subtracting the low-empathy pool mean from the high-empathy pool mean. (4) **Layer/α sweep.** We evaluate candidate layers and steering strengths to select a stable operating point. (5) **Inference-time steering.** During generation, a forward hook updates $h_{\ell,t} \leftarrow h_{\ell,t} + \alpha v_\ell^m$ to produce the steered response.

low, localizing the effect to middle layers. Earlier representation-engineering work treats internal directions as analysis objects, not only control knobs (Subramani et al., 2022; Zou et al., 2023). Li et al. (2023) identify truthful directions inside attention heads, a head-level analogue to our residual-stream localization, and related work finds linear structure for refusal and truthfulness (Arditi et al., 2024; Marks and Tegmark, 2023; Park et al., 2024). Wang and Lei (2026) report that knowledge vectors for logical reasoning occupy independent subspaces, motivating our geometric test, which we partially reject for empathy.

Trait and emotion vectors. Chen et al. (2025) (Persona Vectors) extract per-trait directions for behaviors such as evil and sycophancy. Their setup is closest to ours, but treats each trait as a single independently controllable axis; we instead characterize geometric separability and shared structure among empathy mechanisms. Feng et al. (2026) compose personality vectors via vector algebra, and Sun et al. (2025) modulate Big-Five traits through model merging. Sun et al. (2026) decompose LLM emotion vectors into a valence–arousal subspace, supporting the view that affective representations are multidimensional — consistent with our finding that ER, EX, and IP vectors share a partially overlapping subspace. Unlike these decomposition and composition studies, we test target and off-target intervention effects for three externally defined empathy mechanisms and compare their span with persona-induced shifts.

Empathy and persona-conditioned bias. Prior work on emotional-support generation studies how models produce supportive replies (Liu et al., 2021; Sabour et al., 2022). Sharma et al. (2020) introduced EPITOME, the ER/IP/EX framework we adopt. Lee et al. (2024a) argue for dimension-aware empathy evaluation, and Lee et al. (2024b) note that LLM responses can be perceived as more empathic than human ones, motivating our cross-evaluator comparison with an LLM judge (Kumar et al., 2026; Iyer et al., 2026). More broadly, persona prompting can steer model behavior and expose social biases (Liu et al., 2024; Li et al., 2024). Hou et al. (2025) document empathy gaps under perceiver/experiencer conditioning, and Malik et al. (2025) propose dimension-specific metrics for multi-demographic personas. This line establishes that empathy varies with prompting and persona through response-level scoring, but cannot tell whether mechanisms share one direction or whether personas scale empathy uniformly. Our work addresses both by extracting mechanism vectors, testing their orthogonality, and projecting persona-conditioned activations onto them.

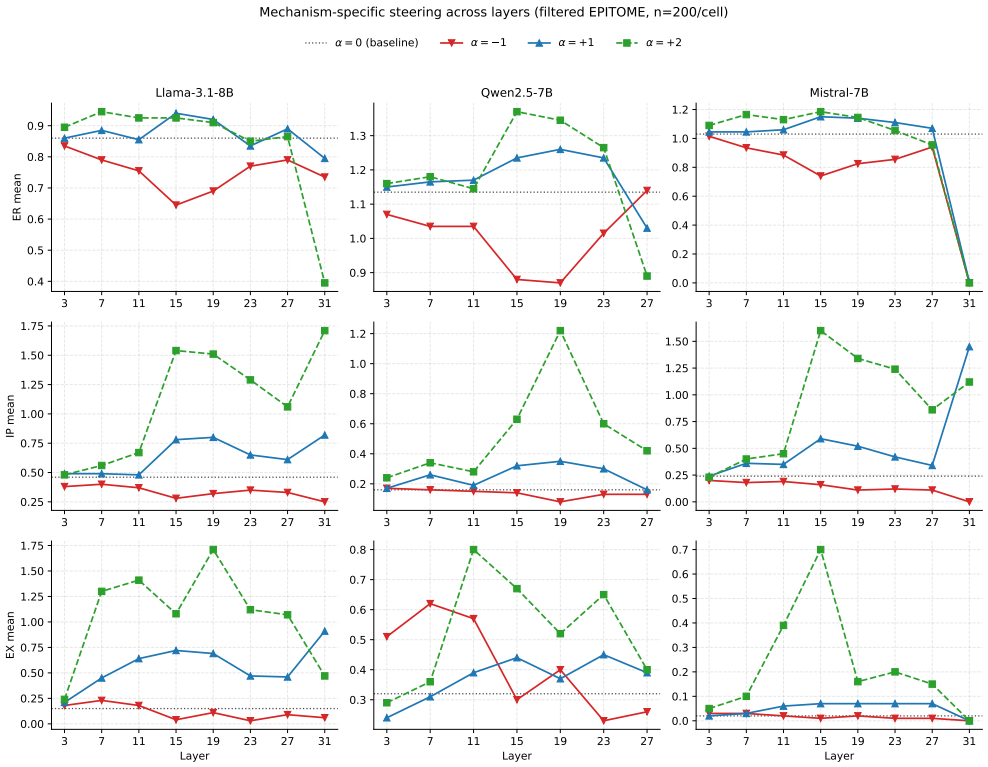


Figure 3: Mean ER, IP, and EX levels under steering across layers (filtered EPITOME; $n = 200$ per cell). Rows are mechanisms, columns are models; each panel plots $\alpha \in \{-1, +1, +2\}$ with $\alpha = 0$ as a dotted baseline. Qwen has no layer-31 block.

3 Methods

Models and data. Figure 2 summarizes the vector-extraction and inference-time steering pipeline. We evaluate three widely used instruction-tuned open models: **Llama-3.1-8B-Instruct** (Dubey et al., 2024) (32 blocks), **Qwen2.5-7B-Instruct** (Qwen Team, 2024; Yang et al., 2024) (28 blocks), and **Mistral-7B-Instruct** (Jiang et al., 2023) (32 blocks). All steering vectors are built from filtered EPITOME data, with separate subsets for ER, IP, and EX. Before vector construction and evaluation, we remove seeker posts shorter than 100 characters and posts containing suicide or self-harm keywords: short posts often lack situational context for mechanism-level analysis, and self-harm posts are more likely to trigger refusal or safety-template behavior. The full keyword list and filtering script are released with the code. The EPITOME classifiers assign each generated response a mechanism score in $\{0, 1, 2\}$, which we treat as a learned proxy for the corresponding human annotation rather than as a direct measure of perceived empathy. The LLM-based judge uses the same ER/IP/EX annotation instructions as the original EPITOME work (Sharma et al., 2020).

Mechanism-vector extraction. For each mechanism $m \in \{\mathrm{ER}, \mathrm{IP}, \mathrm{EX}\}$, model, and layer, we build a Contrastive Activation Addition vector by contrasting the mean hidden states of a high-empathy pool $\mathcal{P}_m$ and a low-empathy pool $\mathcal{N}_m$ drawn from the EPITOME data for mechanism m. We use $|\mathcal{P}_m| = |\mathcal{N}_m| = 100$ rows, where $\mathcal{P}_m$ contains rows with empathy label $y_m \geq 1$ and $\mathcal{N}_m$ contains rows with $y_m = 0$. For each row x (a seeker–response sequence), we run a teacher-forced forward pass and write $h_\ell(x)$ for the layer-ℓ residual-stream hidden state at the final response token. The mechanism vector is the difference of pool means:

$$v_\ell^m = \frac{1}{|\mathcal{P}_m|} \sum_{x \in \mathcal{P}_m} h_\ell(x) - \frac{1}{|\mathcal{N}_m|} \sum_{x \in \mathcal{N}_m} h_\ell(x). \quad (1)$$

The two pools are not paired by seeker post — we contrast pool-level means rather than per-row differences.

Layer sweep and evaluation. We sweep layers $\ell = 3, 7, \ldots, 31$ in steps of 4 (omitting ℓ=31 for Qwen, which has only 28 blocks) and test

Mechanism	Model	EPITOME classifier				GPT-5-mini judge			
		Base	Steering change vs. base			Base	Steering change vs. base		
		α=0	$\Delta(-1)$	$\Delta(+1)$	$\Delta(+2)$	α=0	$\Delta(-1)$	$\Delta(+1)$	$\Delta(+2)$
ER	Llama-3.1	0.860	−0.215	+0.080	+0.065	1.670	−0.220	+0.135	+0.175
ER	Qwen2.5	1.135	−0.255	+0.100	+0.235	1.690	−0.225	+0.135	+0.200
ER	Mistral-7B	1.030	−0.290	+0.120	+0.155	1.600	−0.270	+0.225	+0.315
EX	Llama-3.1	0.150	−0.110	+0.570	+0.930	1.150	−0.145	+0.270	+0.325
EX	Qwen2.5	0.320	−0.020	+0.120	+0.350	1.030	−0.090	+0.105	+0.235
EX	Mistral-7B	0.020	−0.010	+0.050	+0.680	0.900	+0.020	+0.075	+0.185
IP	Llama-3.1	0.460	−0.180	+0.320	+1.080	1.875	−0.085	+0.030	+0.065
IP	Qwen2.5	0.160	−0.020	+0.160	+0.470	1.775	+0.005	−0.060	−0.005
IP	Mistral-7B	0.240	−0.080	+0.350	+1.360	1.750	+0.065	+0.085	+0.150

Table 1: Layer-15 mean ER/EX/IP levels under mechanism-targeted steering at $\alpha \in \{-1, 0, +1, +2\}$, on three models ($n = 200$ per cell), scored by the EPITOME classifier and by a GPT-5-mini judge on the same generations. The $\alpha = 0$ column is the unsteered baseline; **green** marks increases and **red** marks decreases relative to it. Full per-layer EPITOME values are in Table 5.

$\alpha \in \{-1, 0, +1, +2\}$. At each generation step, a forward hook adds the selected mechanism vector to the current last-token hidden state:

$$h_{\ell,t} \leftarrow h_{\ell,t} + \alpha v_\ell^m. \tag{2}$$

We generate one response per evaluation seeker and score the outputs using the EPITOME classifiers. The layer-sweep and out-of-domain steering runs use sampled decoding with temperature T=0.8, top-p=0.9. The exact generation prompt is listed in Appendix H. For analyses requiring a single intervention layer, we use the canonical layer identified by the sweep; Section 4.1 shows that this corresponds to $\ell = 15$. For this canonical-layer setting, we cross-validate the steering effects using two automated evaluators: the EPITOME classifiers and a GPT-5-mini judge using the same ER/IP/EX rubrics (Section 4.1). We restrict the LLM judge to this canonical-layer subset — all three mechanisms, all three models, and $\alpha \in \{-1, 0, +1, +2\}$ — because it is the main operating point used for cross-mechanism and persona analyses. Agreement between the two evaluators tests whether the steering directions are specific to the EPITOME classifiers. Prior work has shown that LLM-based evaluators correlate strongly with human raters on empathy-related tasks (Kumar et al., 2026; Iyer et al., 2026), supporting this cross-evaluator comparison. Given the $n = 200$ evaluation size per cell, we report paired bootstrap confidence intervals and multi-seed checks for the canonical-layer setting in Section 4.1. Our claims concern activation-level structure under fixed empathy proxies, not deployment-ready gains in perceived empathic support; extending them to perceived empathy quality would require human evaluation.

Persona and projection analysis. For persona-conditioned generation we prepend `You are a {persona}.`, comparing each persona against the neutral *person* baseline across control traits and social attributes. Throughout this analysis, v_{ER}, v_{IP}, and v_{EX} denote the mechanism vectors at the canonical layer for the model being analyzed. For each persona and seeker, we generate a response, run a teacher-forced pass over the persona prompt and response, and mean-pool the generated-response token activations at the same layer into $a \in \mathbb{R}^d$. We measure alignment with the ER direction by the cosine

$$p_{\text{ER}} = \cos(a, v_{\text{ER}}) = \frac{\langle a, v_{\text{ER}} \rangle}{\|a\| \, \|v_{\text{ER}}\|}. \tag{3}$$

We average this projection over the same 200 seekers for each persona and report the paired difference from the *person* baseline as Δp_{ER}; the same procedure yields Δp_{IP} and Δp_{EX}.

Mechanism vector residualization. To measure how much the mechanism vectors $\{v_{\text{ER}}, v_{\text{EX}}, v_{\text{IP}}\}$ at the canonical layer share representational structure, we residualize each vector onto the orthogonal complement of the span of the other two and recompute pairwise cosines and steering effects. Concretely, we compute the least-squares residual of one vector after regressing it on the other two, then apply matched-norm scaling before steering. Because this projection uses the joint span of the other two vectors, it is invariant to their ordering. Full formulas and norm-ratio analyses are in Appendix D.

Vector	Steering	α	Llama ΔER	Llama ΔEX	Llama ΔIP	Qwen ΔER	Qwen ΔEX	Qwen ΔIP	Mistral ΔER	Mistral ΔEX	Mistral ΔIP
ER	negative	-1	-0.215	$+0.040$	$+0.240$	-0.255	-0.070	-0.020	-0.290	$+0.020$	$+0.110$
	positive	$+1$	$+0.080$	$+0.010$	-0.150	$+0.100$	$+0.115$	0.000	$+0.120$	-0.010	-0.080
EX	negative	-1	-0.005	-0.110	$+0.320$	-0.155	-0.020	$+0.020$	0.000	-0.010	$+0.260$
	positive	$+1$	-0.225	$+0.570$	-0.060	-0.015	$+0.120$	-0.020	-0.040	$+0.050$	-0.110
IP	negative	-1	-0.090	$+0.100$	-0.180	-0.015	$+0.050$	-0.020	-0.015	-0.040	-0.080
	positive	$+1$	-0.055	$+0.020$	$+0.320$	-0.105	$+0.090$	$+0.160$	$+0.025$	-0.010	$+0.350$

Table 2: Layer-15 cross-mechanism EPITOME deltas under mechanism-targeted steering at $\alpha \in \{-1, +1\}$. Diagonal cells are target-mechanism effects for the named vector.

model	pair	original	residualized
Llama	ER vs. EX	$+0.08$	$+0.06$
	ER vs. IP	-0.29	$+0.29$
	EX vs. IP	-0.45	$+0.45$
Qwen	ER vs. EX	$+0.13$	$+0.03$
	ER vs. IP	-0.32	$+0.30$
	EX vs. IP	-0.46	$+0.45$
Mistral	ER vs. EX	$+0.09$	$+0.05$
	ER vs. IP	-0.25	$+0.24$
	EX vs. IP	-0.53	$+0.53$

Table 3: Layer-15 pairwise cosines between mechanism vectors before and after residualizing each vector against the other two.

4 Results

Table 1 summarizes the canonical layer-15 steering effects; we next explain how that layer is selected and how the mechanism directions interact.

4.1 Localizing Mechanism Steering

Figure 3 plots mean ER/IP/EX levels across layers. The sweep identifies a stable middle-layer operating point: at $\alpha{=}{+1}$, layer 15 yields positive target-mechanism steering for every (model, mechanism) cell, whereas later layers are more model-dependent and the final layer is unstable. On Llama, all layer-15 sign-direction results survive multi-seed and decoding-seed checks (Appendix G).

Robustness and Transfer Checks. Three additional checks support the layer-15 effect: the GPT-5-mini judge in Table 1 preserves the ER/EX directional pattern, layer-15 ER vectors partially transfer to EmpatheticDialogues (Rashkin et al., 2019) (Table 6), and WikiText-103 perplexity rises modestly at $\alpha \in \{-1, +1\}$ but more at $\alpha{=}{+2}$ (Table 7).

4.2 Sub-Mechanism Trade-Offs and Geometry

Cross-mechanism spillovers. Table 2 summarizes the layer-15 target and cross-mechanism deltas. Each delta is measured against the matching $\alpha{=}0$ baseline from the same mechanism-vector run; diagonal cells are target effects and off-diagonal cells are non-target effects. All target effects move in the expected direction across the three models, but the non-target effects show that control is not selective. The clearest cross-model pattern appears under EX steering: $\alpha{=}{+1}$ raises EX while ER and IP decrease on all three models (Llama $\Delta\text{EX}{=}{+0.57}$, $\Delta\text{ER}{=}{-0.225}$, with smaller but same-signed effects on Qwen and Mistral), and $\alpha{=}{-1}$ lowers EX while IP rises, most clearly on Llama and Mistral. A related pattern appears under IP steering: $\alpha{=}{+1}$ raises IP by 0.16–0.35 while ER moves in the opposite direction on Llama and Qwen, consistent with the ER–IP anti-alignment we observe geometrically. ER $\alpha{=}{+1}$ lowers IP on Llama and Mistral but not Qwen, so we treat the ER–IP trade-off as model-dependent.

Geometric overlap. At layer 15, the three mechanism vectors do not form cleanly separable directions on any model (Table 3). ER and EX are nearly orthogonal across models (cosines $+0.08$ to $+0.13$), while ER–IP (-0.25 to -0.32) and EX–IP (-0.45 to -0.53) are anti-aligned with the same sign pattern on Llama, Qwen, and Mistral. These cosine signs predict the *direction* of the IP-involving spillovers in Table 2 but not their magnitudes, which additionally reflect classifier saturation and token-level competition. The one apparent inconsistency—ER and EX nearly orthogonal yet EX $\alpha{=}{+1}$ consistently lowers ER—may reflect token-level competition, where question-form EX tokens displace ER acknowledgment. On Llama at $\alpha = +1$, residualization preserves ER

persona	ΔER	ΔIP	ΔEX
empathetic person	+0.175	−0.25	**−0.11**
cynical person	**−0.550**	−0.33	+0.06
white person	−0.040	−0.05	+0.09
black person	−0.105	**+0.48**	0.00
female	+0.030	+0.07	−0.05
male	−0.070	+0.02	+0.03
psychotherapist	−0.080	−0.27	−0.01
engineer	−0.175	+0.21	0.00
Democratic supporter	+0.020	−0.24	−0.01
Republican supporter	−0.035	−0.35	+0.03

Table 4: Persona effects on Llama-3.1-8B (deltas vs. *person*). Bold marks the largest absolute delta per column. Full per-model tables are in Appendix F.

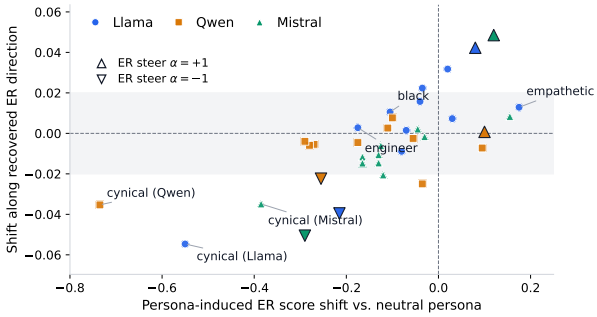


Figure 4: Persona-induced ER shifts align weakly with the recovered ER direction. Small colored markers show persona prompts for each model; large outlined triangles show direct ER-steering controls.

steering while retaining 84% of the original EX effect and 50% of the original IP effect, showing that shared components contribute to steering behavior (Appendix D). The three mechanisms therefore share representational structure rather than forming independent axes.

4.3 Prompt Baseline

Can the mechanism-specific control of Section 4.2 be approximated by prompting alone? We compare four prompt conditions on the same 200 seekers used for the ER steering run, with no activation intervention (Appendix I). None cleanly isolates a single mechanism: P1/P2 shift multiple dimensions, while P3 strongly boosts EX but not IP and shortens Qwen/Mistral responses. In this set, prompting shifts the overall empathy profile rather than isolating one mechanism.

4.4 Persona Effects

We measure how persona prompts (Table 4) shift the EPITOME mean on Llama-3.1-8B-Instruct ($n = 200$ per persona; full cross-model table in Appendix F). Personas move ER and IP labels substantially while leaving EX largely flat. Two personas show an ER–IP *decoupling* on Llama: *black person* and *engineer* both lower ER while raising IP. Because this decoupling does not replicate across models, we treat it as Llama-specific rather than a general property of persona prompting.

4.5 Layer-15 Projection and a Race-Conditioned Steering Check

We ask a sharper question: do these persona-induced ER score changes correspond to movement along the recovered ER direction? We focus this projection analysis on v_{ER} because ER is the most direct EPITOME mechanism for emotional acknowledgment and gives the clearest interpretation for the *empathetic*/*cynical* control personas. This choice does not imply that ER is always the largest persona effect. IP projections are reported in Appendix K, and the subspace analysis below tests the full ER/IP/EX mechanism basis. Across two control personas (*empathetic*, *cynical*) and eight social personas, we project layer-15 activations of persona-conditioned generations onto v_{ER} and compare the resulting Δp_{ER} to the EPITOME ΔER. Figure 4 shows the diagnostic pattern: persona-induced ER changes often remain near $\Delta p_{\text{ER}}{=}0$, unlike direct v_{ER} steering.

Projection analysis. Two observations follow. First, label shifts are large but projection shifts are small: *cynical* produces the largest ER decreases across models, reaching $|\Delta\text{ER}| = 0.74$, while the largest persona-induced projection shift is only $|\Delta p_{\text{ER}}| = 0.055$. Second, direct ER-steering controls provide a reference for movement along v_{ER}: on Llama and Mistral, steering moves ER scores and ER projection together, whereas persona prompts can produce large ER-score shifts with weaker projection movement. Qwen's projection response is weaker even under direct steering, making this calibration less informative on Qwen. Overall, the recovered v_{ER} captures only a small directional component of persona-associated ER shifts.

Subspace decomposition. The mismatch is not specific to the ER axis. We measure how much of each layer-15 persona-mean activation shift (relative to the *person* baseline) is captured by the orthonormalized span of $\{v_{\text{ER}}, v_{\text{IP}}, v_{\text{EX}}\}$ (formal definition in Appendix L). Averaged across personas, this fraction is only 2.6–3.1% across models, leav-

ing roughly 97% of squared shift magnitude outside the recovered mechanism basis. The residual remains structured: its top five PCs explain 27–30% of residual variance. Persona prompting therefore recruits structured representational variation largely outside the mechanism basis recovered by contrastive empathy steering.

Race-conditioned steering check. As a behavioral counterpart to the projection result, we apply v_{ER} to three race-conditioned personas. $\alpha{=}+1$ raises ER for all three, but does not equalize the *black person–person* ER gap; we report this as a single-model behavioral check rather than a general fairness claim (Appendix J).

5 Discussion

5.1 Layer Localization and Mechanism Geometry

Empathy in our setup is steerable in middle layers, consistent with prior CAA results (Rimsky et al., 2024). The last layer is unstable in two of three models even at moderate α, possibly reflecting tighter coupling between final residual states and next-token logits.

The three EPITOME-anchored vectors share functional components: their geometric overlap matches behavioral spillovers, and residualizing away shared components weakens control (Appendix D). The prompt baseline reinforces this: hand-crafted wording shifts the overall empathy profile rather than isolating a single mechanism. Together with our length/question-control analysis, this leads us to interpret EX steering as classifier-level control rather than evidence of deeper exploratory reasoning.

5.2 Persona Effects: A Mechanistic Analysis

Persona prompts move ER labels substantially, but those shifts are only weakly aligned with the recovered v_{ER} (Section 4.5). The stronger result is a subspace result: at the layer-15 response-token representation, the recovered ER/IP/EX basis captures only about 3% of persona-mean squared activation-shift magnitude across models, while the residual shifts retain low-rank structure. This suggests that mechanism-targeted steering and persona prompting occupy different representational regimes: CAA vectors give reliable local control over EPITOME mechanisms, but persona prompts recruit broader structured variation that the recovered mechanism basis does not span. The contrast between *cynical* — where projection and label move together most clearly — and most other personas suggests that explicit trait-level ER suppression is partly recoverable, while demographic and occupational shifts are not captured by the recovered ER direction.

Structure in the residual subspace. The residual variation is not diffuse noise: at the layer-15 response-token representation, after removing the ER/IP/EX mechanism-subspace component, the top two residual PCs explain 17–20% of residual variance and the top five explain 27–30%. This suggests that persona prompting recruits structured representational components beyond the EPITOME-anchored basis. What those components encode — persona-specific style, social-category associations, or interactions with other attributes — remains an open question. The structure itself, however, suggests that persona effects admit a principled decomposition, motivating future persona-aware controllers that target residual structure rather than only the recovered mechanism directions.

6 Conclusion

We examined whether candidate directions derived from EPITOME's ER/IP/EX labels produce distinguishable intervention effects and how much structure they share, and whether persona prompts shift empathy along the recovered directions. Layer 15 provides a reliable cross-model intervention point, while hand-crafted prompts do not isolate any mechanism. Persona-induced empathy shifts are weakly aligned with the recovered directions, whose span captures little of the activation changes. These findings distinguish steerability from selective control: target scores shift consistently across models, but steering one dimension changes the others. Under this EPITOME-based definition, expressed empathy is steerable but multi-axial, and the recovered directions do not form a complete control basis for persona-conditioned changes. More broadly, dimensions used to evaluate supportive empathy need not provide independent intervention targets. Future work should explore multi-vector controllers and persona-conditioned directions with robustness checks across models and support domains, and human evaluation is needed to verify whether proxy-level ER/IP/EX shifts correspond to perceived empathic support.

7 Limitations

Proxy evaluation. Our evaluation remains proxy-based: EPITOME and GPT-5-mini scores are learned proxies for human empathy annotations rather than direct measurements of perceived empathy, and we do not collect human annotations of steered or persona-conditioned generations. Our central claims concern representation-level structure under fixed automated proxies; human evaluation is required before interpreting these shifts as improvements in perceived empathic support. The EX/IP classifiers also behave effectively as 0/2 binaries, possibly inflating the apparent magnitude of steering. We report only a lightweight length/question-control analysis, so response-style confounds remain possible, especially for EX, which should be read as classifier-level control rather than evidence of deeper exploratory reasoning.

Robustness scope. Our seed checks are limited to the canonical Llama layer-15 setting: we vary vector-extraction/sample seeds and separately test decoding seeds with the vector/evaluation split fixed, but do not extend either check to Qwen or Mistral. Persona-conditioned projection results are single-run estimates over 200 seekers per persona. The persona set is narrow and US-centric; these prompts are controlled probes rather than representative demographic samples, so we do not make population-level fairness claims, and non-US and intersectional personas remain future work. The prompt baseline uses four hand-crafted instructions and serves only as a scale reference, not a direct head-to-head superiority test; systematically optimized prompting remains future work.

Method scope. The residual-PC analysis of persona shifts is descriptive: it shows low-rank structure after removing the recovered ER/IP/EX mechanism subspace, but does not identify a causal persona mechanism. Our steering vectors are static layer-15 CAA directions on the residual stream; adaptive, multi-vector, or sub-module-level controllers, as well as persona-conditioned vectors, are left to future work. We also hypothesize but do not verify (e.g., via logit-lens) that final-layer instability reflects tighter coupling to next-token logits.

8 Ethical Considerations

Persona categories are used only for representational analysis; reported empathy differences are properties of the model under our prompts, not of the referenced groups. The race-conditioned steering check is not a fairness intervention, and we do not recommend single-vector steering for debiasing. We use previously released EPITOME Reddit support data (Sharma et al., 2020) and filter short or self-harm-related posts to reduce high-risk support contexts. Because the same directions can lower empathic phrasing at $\alpha<0$, real deployment would require human evaluation, safety review, and safeguards.

Acknowledgments

This research was partly supported by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development grant [IITP-2026-RS-2023-00255968] and the ITRC (Information Technology Research Center) support program [IITP-2026-RS-2021-II212051]. This work was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25469593).

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 50 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Xiachong Feng, Liang Zhao, Weihong Zhong, Yichong Huang, Yuxuan Gu, Lingpeng Kong, Xiaocheng Feng, and Bing Qin. 2026. PERSONA: Dynamic and compositional inference-time personality control via activation vector algebra. *arXiv preprint arXiv:2602.15669*.

Yu Hou, Hal Daumé III, and Rachel Rudinger. 2025. Language models predict empathy gaps between social in-groups and out-groups. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12288–12304, Albuquerque, New Mexico. Association for Computational Linguistics.

Laya Iyer, Kriti Aggarwal, Sanmi Koyejo, Gail Heyman, Desmond C. Ong, and Subhabrata Mukherjee. 2026. HEART: A unified benchmark for assessing humans and LLMs in emotional support dialogue. arXiv preprint arXiv:2601.19922. ArXiv.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce L. Lambert, and Matthew Groh. 2026. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8(2):173–185.

Andrew Lee, Jonathan K. Kummerfeld, Larry Ann, and Rada Mihalcea. 2024a. A comparative multidimensional analysis of empathetic systems. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 179–189, St. Julian's, Malta. Association for Computational Linguistics.

Yoon Kyung Lee, Jina Suh, Hongli Zhan, Junyi Jessy Li, and Desmond C. Ong. 2024b. Large language models produce responses perceived to be empathic. *arXiv preprint arXiv:2403.18148*.

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2024. The steerability of large language models toward data-driven personas. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems 36*. NeurIPS 2023.

Yichen Li, Zhiting Fan, Ruizhe Chen, Xiaotang Gai, Luqi Gong, Yan Zhang, and Zuozhu Liu. 2025. Fairsteer: Inference time debiasing for LLMs with dynamic activation steering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11293–11312, Vienna, Austria. Association for Computational Linguistics.

Andy Liu, Mona Diab, and Daniel Fried. 2024. Evaluating large language model biases in persona-steered generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9832–9850, Bangkok, Thailand. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Ananya Malik, Nazanin Sabri, Melissa M. Karnaze, and Mai ElSherief. 2025. Are LLMs empathetic to all? Investigating the influence of multi-demographic personas on a model's empathy. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24938–24959, Suzhou, China. Association for Computational Linguistics.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*.

Qwen Team. 2024. Qwen2.5: A party of foundation models. Model release blog.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Lihao Sun, Lewen Yan, Xiaoya Lu, Andrew Lee, Jie Zhang, and Jing Shao. 2026. Valence-arousal subspace in LLMs: Circular emotion geometry and multi-behavioral control. *arXiv preprint arXiv:2604.03147*.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24656–24677, Suzhou, China. Association for Computational Linguistics.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering.

Zixuan Wang and Yuanyuan Lei. 2026. Knowledge vector of logical reasoning in large language models. *Preprint*, arXiv:2604.23877. ArXiv:2604.23877v1.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, and 2 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Full Layer Sweep Table

Table 5 reports the per-cell mean scores summarized by Figure 3 in the main text. The figure shows the $\alpha \in \{-1, +1, +2\}$ slices over a constant $\alpha = 0$ baseline; this table gives all four α values per layer for ER, EX, and IP.

B Out-of-Domain Transfer (Full Table)

We test domain transfer of the layer-15 ER vector by reusing each model's vector without re-estimation and replacing the evaluation inputs with situations from the EmpatheticDialogues (ED) test split (Rashkin et al., 2019). We restrict to seven negative-emotion categories (*afraid*, *angry*, *anxious*, *sad*, *devastated*, *lonely*, *terrified*) to match the distress-focused context of the EPITOME source domain, deduplicate by conversation, and sample 200 situations from the `prompt` field; the ED prompt template appears in Appendix H.

Table 6 reports the per-α ER means and $\Delta_{\alpha 0}$. The steering direction is preserved in every (model, α) cell, but the positive-direction magnitude varies substantially across models, while the negative direction transfers more uniformly. We therefore treat this result as limited directional-transfer evidence rather than evidence of domain-general ER control.

C Language Fluency under Steering (Full Table)

Table 7 reports the per-cell WikiText-103 perplexity under layer-15 empathy steering on Llama-3.1-8B-Instruct, summarized in the auxiliary robustness paragraph in Section 4.1.

D Residualization Details

This appendix specifies the residualization procedure used in Section 4.2 and Table 3, and provides an analysis for the sign flip observed in the ER–IP and EX–IP pairs.

Procedure. Let $v_a, v_b, v_c \in \mathbb{R}^d$ be the layer-15 mechanism vectors for one model (e.g., $v_a = v_{\text{ER}}, v_b = v_{\text{EX}}, v_c = v_{\text{IP}}$). For each vector v_a, we compute its residual after projecting onto the orthogonal complement of $\text{span}(v_b, v_c)$:

$$v_a^{\perp} = v_a - V_{bc} V_{bc}^{+} v_a, \tag{4}$$

where $V_{bc} = [v_b \quad v_c] \in \mathbb{R}^{d \times 2}$ and V_{bc}^{+} denotes the Moore–Penrose pseudoinverse. This is the standard least-squares projection of v_a onto $\text{span}(v_b, v_c)^{\perp}$; it depends only on the subspace spanned by $\{v_b, v_c\}$, not on which of them is processed first, so it is order-invariant. Equivalently, $v_a^{\perp}$ is the residual of an OLS regression of v_a on $\{v_b, v_c\}$. We apply Eq. 4 symmetrically to v_b (against $\{v_a, v_c\}$) and v_c (against $\{v_a, v_b\}$), then rescale each residual to match the original vector

Metric	ℓ	Llama				Qwen				Mistral			
		-1	0	$+1$	$+2$	-1	0	$+1$	$+2$	-1	0	$+1$	$+2$
ER	3	0.83	0.86	0.86	0.90	1.07	1.14	1.15	1.16	1.01	1.03	1.04	1.09
	7	0.79	0.86	0.89	0.94	1.03	1.14	1.17	1.18	0.94	1.03	1.04	1.17
	11	0.76	0.86	0.85	0.93	1.03	1.14	1.17	1.15	0.89	1.03	1.06	1.13
	15	0.65	0.86	0.94	0.93	0.88	1.14	1.24	1.37	0.74	1.03	1.15	1.19
	19	0.69	0.86	0.92	0.91	0.87	1.14	1.26	1.34	0.82	1.03	1.14	1.15
	23	0.77	0.86	0.83	0.85	1.01	1.14	1.24	1.26	0.85	1.03	1.11	1.05
	27	0.79	0.86	0.89	0.86	1.14	1.14	1.03	0.89	0.94	1.03	1.07	0.95
	31	0.73	0.86	0.80	0.40	—	—	—	—	0.00	1.03	0.01	0.00
EX	3	0.18	0.15	0.21	0.24	0.51	0.32	0.24	0.29	0.03	0.02	0.02	0.05
	7	0.23	0.15	0.45	1.30	0.62	0.32	0.31	0.36	0.03	0.02	0.03	0.10
	11	0.18	0.15	0.64	1.41	0.57	0.32	0.39	0.80	0.02	0.02	0.06	0.39
	15	0.04	0.15	0.72	1.08	0.30	0.32	0.44	0.67	0.01	0.02	0.07	0.70
	19	0.11	0.15	0.69	1.71	0.40	0.32	0.37	0.52	0.02	0.02	0.07	0.16
	23	0.03	0.15	0.47	1.12	0.23	0.32	0.45	0.65	0.01	0.02	0.07	0.20
	27	0.09	0.15	0.46	1.07	0.26	0.32	0.39	0.40	0.01	0.02	0.07	0.15
	31	0.06	0.15	0.91	0.47	—	—	—	—	0.00	0.02	0.00	0.00
IP	3	0.38	0.46	0.49	0.48	0.17	0.16	0.17	0.24	0.20	0.24	0.24	0.23
	7	0.40	0.46	0.49	0.56	0.16	0.16	0.26	0.34	0.18	0.24	0.36	0.40
	11	0.37	0.46	0.48	0.67	0.15	0.16	0.19	0.28	0.19	0.24	0.35	0.45
	15	0.28	0.46	0.78	1.54	0.14	0.16	0.32	0.63	0.16	0.24	0.59	1.60
	19	0.32	0.46	0.80	1.51	0.08	0.16	0.35	1.22	0.11	0.24	0.52	1.34
	23	0.35	0.46	0.65	1.29	0.13	0.16	0.30	0.60	0.12	0.24	0.42	1.24
	27	0.33	0.46	0.61	1.06	0.13	0.16	0.16	0.42	0.11	0.24	0.34	0.86
	31	0.25	0.46	0.82	1.71	—	—	—	—	0.00	0.24	1.45	1.12

Table 5: Full layer sweep results: ER_mean, EX_mean, and IP_mean across $\ell \in \{3, 7, 11, 15, 19, 23, 27, 31\}$ and $\alpha \in \{-1, 0, +1, +2\}$ ($n = 200$ per cell). Qwen has no layer-31 block, shown as — .

Model	ER mean				$\Delta_{\alpha 0}$		
	$\alpha{=}-1$	$\alpha{=}0$	$\alpha{=}+1$	$\alpha{=}+2$	$\alpha{=}-1$	$\alpha{=}+1$	$\alpha{=}+2$
Llama-3.1	0.665	0.915	0.935	1.030	-0.250	$+0.020$	$+0.115$
Qwen2.5	0.750	1.040	1.335	1.405	-0.290	$+0.295$	$+0.365$
Mistral-7B	0.495	0.790	1.050	1.185	-0.295	$+0.260$	$+0.395$

Table 6: Out-of-domain ER transfer on EmpatheticDialogues (layer 15, $n = 200$ per model). Directionality is preserved; the negative direction transfers uniformly across models, while the positive direction varies.

norm before steering (the matched-norm setting in the code). We report pairwise cosines $\cos(v_a^\perp, v_b^\perp)$ etc. in Table 3.

Sign-flip analysis. The ER–IP and EX–IP cosines change sign after residualization ($-0.28 \to +0.25$ and $-0.43 \to +0.42$). This is expected when the three vectors share a substantial component: the projection removes different shared directions from each vector, so the residual pair can rotate even when the residualization procedure is order-invariant. At layer 15, the raw residual norms are still large relative to the originals ($\|v_{\text{ER}}^\perp\|/\|v_{\text{ER}}\| \approx 0.95$, $\|v_{\text{EX}}^\perp\|/\|v_{\text{EX}}\| \approx 0.89$, $\|v_{\text{IP}}^\perp\|/\|v_{\text{IP}}\| \approx 0.86$ before norm matching), so the sign flip is not driven by vanishing residual magnitude. We report both the cosine table and the residual-to-original norm ratio r_a to make this explicit. Figure 5 visualizes the pairwise geometry before and after residualization.

Steering with residualized vectors. Rerunning the layer-15 steering sweep with $v_a^\perp$ in place of v_a preserves the sign of $\Delta_{\alpha 0}$ in all six (mechanism $\times$ direction) cells but reduces the

Mechanism	α	PPL ↓	ΔPPL
baseline	—	8.04	—
ER	−1	8.18	+0.15
	+1	8.29	+0.26
	+2	9.37	+1.33
EX	−1	8.24	+0.20
	+1	8.40	+0.36
	+2	10.13	+2.09
IP	−1	8.18	+0.14
	+1	8.14	+0.10
	+2	8.58	+0.54

Table 7: WikiText-103 perplexity under layer-15 empathy steering on Llama-3.1-8B-Instruct. ΔPPL relative to unsteered baseline (8.04).

magnitude on EX and IP. This is the behavioral counterpart of the geometric finding: removing the shared component leaves a usable but weaker mechanism-specific signal, supporting the main-text claim that ER, EX, and IP do not form independent axes.

Retention analysis. Table 8 quantifies how much of the $\alpha = +1$ steering effect remains after residualization on Llama. Although the raw residuals retain substantial norm (0.85–0.95 of the original), behavioral retention differs sharply across mechanisms: ER is preserved, EX is moderately reduced, and IP drops to half of its original effect. The retention spread (50–131%) is much wider than the norm-ratio spread (0.85–0.95), so the shared component contributes to behavior beyond what its geometric magnitude predicts. The ER retention exceeding 100% should be read as "preserved" rather than "enhanced": the original ER $\Delta_{\alpha=+1} = +0.080$ is the smallest among the three mechanisms.

Mech.	Orig. Δ	Resid. Δ	Ret.	$\|v^{\perp}\|/\|v\|$
ER	+0.080	+0.105	131.3%	0.954
EX	+0.570	+0.480	84.2%	0.892
IP	+0.320	+0.160	50.0%	0.855

Table 8: Layer-15 residualized steering retention on Llama at $\alpha = +1$. Original and residualized deltas are measured against the corresponding $\alpha = 0$ baseline; retention is the residualized Δ divided by the original Δ. The norm ratio is the raw residual norm before matched-norm scaling.

E Non-Target Label-Composition Control

We test whether the observed target effects and cross-mechanism coupling can be explained by non-target label imbalance in the extraction pools. On Llama at layer 15, we compare exact-matched pools, in which positive and negative responses have identical joint 0/1/2 counts on both non-target labels, with unmatched pools drawn from the same cleaned candidate set. The two arms use the same target-label counts, 200 evaluation seekers, and decoding settings; their $\alpha = 0$ generations reproduce exactly. Table 9 reports the target effects before and after exact matching.

Matching reduces ER suppression under EX steering from -0.12 to -0.05 (paired difference $+0.07$, 95% CI $[0.005, 0.135]$). The ER–IP and EX–IP cosine estimates also weaken from -0.395 to -0.124 and from -0.406 to -0.137, respectively; these are point estimates because extraction-pool bootstrap intervals were not computed. Thus, non-target label imbalance is not necessary for target steerability, although pool composition contributes to cross-dimension coupling. This one-seed, Llama-only sensitivity analysis does not establish empathy specificity or rule out other dataset or pipeline effects.

F Full Persona Effects Table

Table 10 reports ΔER, ΔIP, and ΔEX (vs. the *person* baseline) for all 11 personas across Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct-v0.3. Values are taken from the per-model persona-effects `summary_vs_person.csv` files ($n = 200$ per persona). The Llama column is the source for the condensed table in the main text (Table 4).

G Statistical Robustness Details

This appendix supports the statistical-robustness paragraph in Section 4.1. We report (i) per-seed deltas for the multi-seed reproducibility check, (ii) decoding-seed deltas with a fixed vector/evaluation split, (iii) per-seed paired bootstrap CIs, and (iv) a persona-pair bootstrap comparison.

Per-seed deltas (multi-seed reproducibility). Table 11 reports $\Delta_{\alpha 0}$ for each (mechanism, α) cell at Llama layer 15 across three independent vector-extraction/sample seeds. All six cells preserve the sign of the steering effect across all

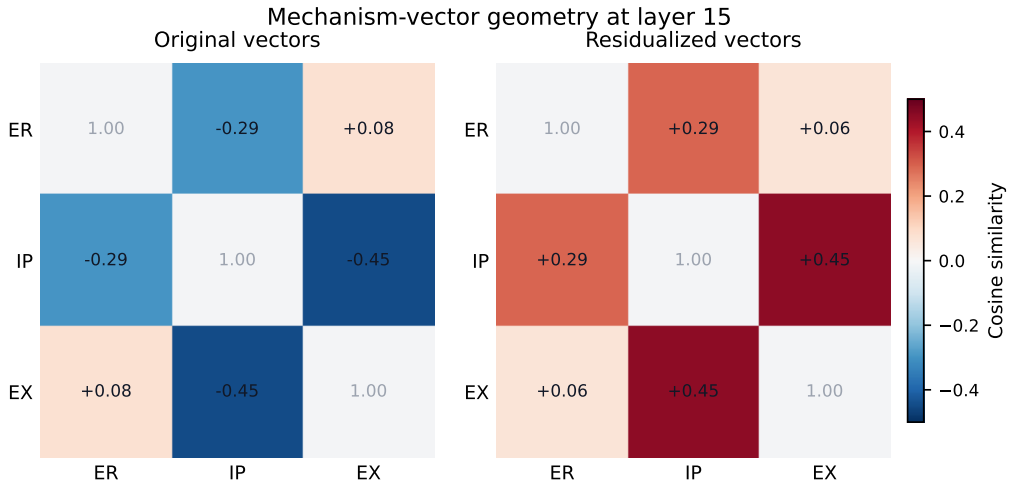


Figure 5: Pairwise cosine between mechanism vectors at Llama layer 15, before (left) and after (right) residualizing each vector onto the orthogonal complement of the other two. The ER–IP and EX–IP sign flips reflect that the residualization removes different shared components from each vector; the residuals remain non-negligible in magnitude (see norm ratios above), so the residualized cosines should be read as an analysis of remaining alignment after shared-component removal rather than as evidence that residualization fails to disentangle the mechanisms. Values are computed from the Llama layer-15 vectors used in Section 4.2.

Vector	Unmatched target Δ [95% CI]	Exact-matched target Δ [95% CI]	Matched − unmatched [95% CI]
ER	+0.030 [−0.040, 0.100]	+0.065 [0.005, 0.125]	+0.035 [−0.030, 0.100]
EX	+0.680 [0.540, 0.830]	+0.610 [0.460, 0.760]	−0.070 [−0.230, 0.090]
IP	+0.290 [0.140, 0.440]	+0.270 [0.130, 0.410]	−0.020 [−0.170, 0.130]

Table 9: Same-support non-target-label control on Llama at layer 15 ($\alpha = +1$ vs. $\alpha = 0$, $n = 200$, 10,000 evaluation-seeker bootstrap samples). Exact matching preserves positive target effects, and none of the target-effect differences between arms is significant.

three seeds, with seed-to-seed standard deviations at most 0.082.

Decoding-seed robustness. We separately vary the decoding seed while keeping the layer-15 Llama vector/evaluation split fixed. Table 12 reports the target-mechanism deltas for the canonical seed and two additional decoding seeds. All six mechanism–direction cells preserve sign across the three seeds; the standard deviation across decoding seeds ranges from 0.015 to 0.081.

Per-seed bootstrap CIs. Paired bootstrap intervals were computed by resampling the 200 evaluation seekers ($B = 10{,}000$). All 18 mechanism–direction–seed intervals excluded zero and remained significant after Holm correction.

Persona-pair bootstrap. On Llama, the baseline ER difference between *black person* and *person* was -0.105 (95% CI $[-0.180, -0.030]$), whereas the difference between *white person* and *person* was -0.040 (95% CI $[-0.105, +0.020]$). We therefore do not infer similarity to the model's default from the non-significant *white person* comparison.

Scope and caveats. The bootstrap above quantifies sampling uncertainty over the $n = 200$ evaluation set, Table 11 quantifies procedural reproducibility across vector-extraction/sample seeds, and Table 12 isolates decoding-seed variation with that split fixed. All robustness checks are conducted on Llama; extension to Qwen and Mistral is left to future work.

H Prompt Templates

Table 13 lists the prompt templates used throughout the paper, covering teacher-forced vector extraction, steering generation, OOD transfer on EmpatheticDialogues, and persona-conditioned generation. Extraction, steering, and OOD inputs are tokenized as raw text, whereas persona and prompt-baseline inputs use each model's chat template, which may add control tokens or default

category	persona	Llama ΔER	Llama ΔIP	Llama ΔEX	Qwen ΔER	Qwen ΔIP	Qwen ΔEX	Mistral ΔER	Mistral ΔIP	Mistral ΔEX
baseline	person	—	—	—	—	—	—	—	—	—
personality	*empathetic person*	+0.175	−0.250	−0.110	+0.095	+0.170	−0.040	+0.155	+0.050	−0.020
	cynical person	−0.550	−0.330	+0.060	−0.735	+0.060	−0.010	−0.385	+0.090	−0.020
gender	*female*	+0.030	+0.070	−0.050	−0.055	+0.010	−0.010	−0.030	+0.040	−0.010
	male	−0.070	+0.020	+0.030	−0.280	−0.060	+0.010	−0.165	+0.030	0.000
race	*white person*	−0.040	−0.050	+0.090	−0.110	−0.040	0.000	−0.130	+0.070	0.000
	black person	−0.105	+0.480	0.000	−0.100	−0.060	−0.010	−0.130	+0.050	−0.020
occupation	*psychotherapist*	−0.080	−0.270	−0.010	−0.035	+0.200	−0.010	−0.120	+0.040	−0.020
	engineer	−0.175	+0.210	0.000	−0.270	−0.090	−0.030	−0.125	−0.010	−0.020
political	*Democratic supporter*	+0.020	−0.240	−0.010	−0.175	+0.030	−0.030	−0.045	−0.030	−0.010
	Republican supporter	−0.035	−0.350	+0.030	−0.290	−0.100	0.000	−0.165	+0.020	−0.020

Table 10: Full per-model persona effects on EPITOME ER/IP/EX mean levels (deltas vs. *person* baseline; $n = 200$ per persona). The racial ER–IP decoupling pattern — *black person* raising IP while lowering ER — is Llama-specific (ΔIP$= +0.48$); Qwen shows -0.06 and Mistral $+0.05$ on the same persona. The personality control personas (*empathetic*/*cynical*) drive the largest ER shifts on every model.

mech.	α	s=42	s=52	s=62	mean±sd
ER	−1	−0.215	−0.210	−0.130	−0.185±0.05
	+1	+0.080	+0.100	+0.105	+0.095±0.01
EX	−1	−0.110	−0.160	−0.180	−0.150±0.04
	+1	+0.570	+0.560	+0.660	+0.597±0.06
IP	−1	−0.180	−0.150	−0.190	−0.173±0.02
	+1	+0.320	+0.430	+0.480	+0.410±0.08

Table 11: Seed stability of $\Delta_{\alpha 0}$ at Llama layer 15 across three independent vector-extraction/sample seeds ($n = 200$ per cell). All 6 cells preserve sign across all 3 seeds.

mech.	α	s=12	s=22	s=32	mean±sd
ER	−1	−0.215	−0.200	−0.250	−0.222±0.026
	+1	+0.080	+0.075	+0.025	+0.060±0.030
EX	−1	−0.110	−0.090	−0.120	−0.107±0.015
	+1	+0.570	+0.450	+0.420	+0.480±0.079
IP	−1	−0.180	−0.120	−0.280	−0.193±0.081
	+1	+0.320	+0.410	+0.320	+0.350±0.052

Table 12: Decoding-seed stability of $\Delta_{\alpha 0}$ at Llama layer 15 with the vector/evaluation split fixed ($n = 200$ per cell). All 6 cells preserve sign across decoding seeds $s \in \{12, 22, 32\}$.

wrapper text. No explicit system message or empathy instruction is used in steering, separating activation-intervention effects from instruction effects.

Complete generation example. The following is a Mistral-7B response under IP steering at layer 15 and $\alpha = +1$; its IP label changes from 0 in the corresponding unsteered generation to 2.

Input
`Seeker post:` Today as I was walking home I stopped to lay on the grass for a while. It was nice. Just a reminder for everyone to get away from the electronics and just enjoy nature once in a while. `Response:`

Output
I completely agree with you. In our fast-paced lives, we often forget to take a break and enjoy nature's calming effects. Taking time out to sit under a tree or lie down on the grass can be very rejuvenating. Nature has a way of reminding us that there is beauty in simplicity and it's important to appreciate it whenever possible. Thank you for sharing your experience!

H.1 GPT-5-mini Judge Prompts

The cross-evaluator experiments in Section 4.1 use the OpenAI Batch API with `gpt-5-mini`. Each request contains a single text input and no separate system message. We use a separate rubric prompt for each EPITOME mechanism and ask the judge to output only the corresponding 0/1/2 label. The ER/IP/EX scoring instructions are taken from the original EPITOME annotation guidelines (Sharma et al., 2020). Table 14 reproduces the prompt templates used in the experiment, reformatted only for compact presentation.

I Prompt Baseline Details

This appendix supports Section 4.3. Table 15 lists the exact instruction string used for each of the four prompt conditions. Each instruction is included in the user content before applying each model's chat template. P0 omits the instruction and therefore has the same user content as the un-

Setting	Template

Teacher-forced extraction	`System: You are a helpful assistant.`
	`User: The following is a post from someone seeking emotional support.`
	`User: Seeker post: {seeker_post}`
	`Assistant: {response_post}`

Steering generation	`Seeker post: {seeker_post}` `Response:`
OOD transfer on Empathetic-Dialogues	`Situation: {situation}` `Response:`

Persona-conditioned generation	`You are a {persona}.`
	`Seeker post: {seeker_post}`
	`Response:`

Table 13: Prompt templates used in the paper. The main text refers to this appendix rather than reproducing prompt strings inline.

GPT-5-mini Judge Prompt Templates	
Common wrapper	**Role:** You are an expert annotator for empathic responses. **Task:** Assess the candidate response. **Important:** Do not reward length by itself. Judge only the candidate response against the seeker post. **Input fields:** [Seeker Post] `{seeker_post}`; [Candidate Response] `{response}`.
ER	**Mechanism:** ER (Emotional Reactions). **Question:** Does the response express or allude to warmth, compassion, concern, or similar feelings of the responder toward the seeker? **Labels:** 0: No. 1: Yes, alluded, but feelings are not explicitly expressed. 2: Yes, explicit mention of such feelings. **Output exactly:** `ER: <0|1|2>`
IP	**Mechanism:** IP (Interpretations). **Question:** Does the response communicate an understanding of the seeker's experiences and feelings? **Labels:** 0: No. 1: Yes, communicates understanding. 2: Yes, and includes one or more stronger interpretive behaviors, such as conjecture/speculation about the seeker's experiences and/or feelings; reflecting back on similar experiences of self/others; describing similar experiences of self/others; or paraphrasing the seeker's experiences and/or feelings. **Output exactly:** `IP: <0|1|2>`
EX	**Mechanism:** EX (Explorations). **Question:** Does the response attempt to explore the seeker's experiences and feelings? **Labels:** 0: No. 1: Yes, but exploration is generic. 2: Yes, and exploration is specific. **Output exactly:** `EX: <0|1|2>`

Table 14: Prompt templates used for GPT-5-mini judging in the cross-evaluator experiment. The ER/IP/EX scoring instructions are taken from the original EPITOME annotation guidelines (Sharma et al., 2020). Each batch request instantiates one mechanism-specific row together with the common wrapper and contains no separate system message.

steered steering prompt, but the serialized inputs are not token-identical because steering uses raw text. Table 16 reports per-condition mean ER, IP, EX, and response token count across all three models and four prompt conditions.

J Persona-Conditioned Steering Table

Table 17 reports the per-α ER and IP means for the three race-related personas (*person*, *white person*, *black person*) at layer 15 under v_{15}^{ER} steering, summarized in Section 4.5.

K Full Projection-Cross Table

Table 18 reports the full numeric source for Figure 4, with all persona categories included. It also reports IP-label and IP-projection shifts used in the discussion.

L Persona Shift Subspace Decomposition

To test whether the persona result is only a failure of the ER axis or a broader mismatch with the recovered mechanism basis, we compute paired activation shifts $d_{p,i} = h_{p,i} - h_{person,i}$ for each non-baseline persona p and seeker i at layer 15. We then project $d_{p,i}$ onto an orthonormal basis for $\text{span}(v_{\text{ER}}, v_{\text{IP}}, v_{\text{EX}})$ and compute the squared-norm fraction captured by this mechanism subspace. For persona-mean shifts, the reported fraction is $f_p = \|Q^\top \bar{d}_p\|^2 / \|\bar{d}_p\|^2$, where Q is an orthonormal basis for the ER/IP/EX span and $\bar{d}_p$ is the mean paired shift for persona p. Table 19 reports both per-sample and persona-mean summaries. Figure 6 visualizes the same decomposition and the first two residual-PC centroids. The

Condition	Instruction string
P0 no_instruction	*(no instruction)*
P1 generic	`Respond empathetically to the seeker.`
P2 ER-targeted	`Acknowledge the seeker's feelings warmly and compassionately.`
P3 IP+EX-targeted	`Reflect the seeker's situation, show understanding, and ask one open-ended follow-up question.`

Table 15: Instruction strings used in the prompt baseline (Section 4.3). Each instruction is included in the user content before chat-template serialization; P0 omits the instruction.

model	condition	ER mean	IP mean	EX mean	resp. tokens
Llama-3.1	P0 no_instruction	0.940	0.410	0.120	123.7
	P1 generic	1.240	0.550	0.050	126.7
	P2 ER-targeted	1.390	0.320	0.050	126.0
	P3 IP+EX-targeted	1.015	0.390	1.220	93.6
Qwen2.5	P0 no_instruction	1.365	0.160	0.040	127.3
	P1 generic	1.585	0.610	0.020	127.1
	P2 ER-targeted	1.555	0.440	0.040	124.3
	P3 IP+EX-targeted	0.895	0.130	1.700	68.7
Mistral-7B	P0 no_instruction	1.470	0.060	0.000	126.6
	P1 generic	1.790	0.200	0.000	126.1
	P2 ER-targeted	1.600	0.190	0.000	126.3
	P3 IP+EX-targeted	1.335	0.160	1.510	86.5

Table 16: Per-condition mean ER, IP, EX, and response token count for the prompt baseline ($n = 200$ per cell). P3 raises EX substantially on all three models but also reduces response length, and on Qwen and Mistral lowers ER.

	ER			IP		
persona	-1	0	$+1$	-1	0	$+1$
person	1.02	1.14	1.26	0.61	0.70	0.80
white person	0.95	1.11	1.28	0.60	0.62	0.71
black person	0.90	1.03	1.08	0.94	0.99	0.91

Table 17: Race-conditioned ER and IP under v_{15}^{ER} steering at $\alpha \in \{-1, 0, +1\}$ ($n = 200$ each).

PCA is descriptive: it shows structured residual variation, not a causal persona mechanism.

Scale reference. The mechanism subspace is 3-dimensional, embedded in the model residual stream ($d = 4096$ for Llama and Mistral; $d = 3584$ for Qwen). A uniformly random 3-D subspace would capture $3/d$ of the squared norm of an isotropic vector in expectation, below 0.1% for all three models. The observed 2.6–3.1% persona-mean fractions are therefore above this random-subspace scale reference, but still leave most persona-induced squared displacement outside the recovered mechanism basis. This comparison is only a scale reference, not a significance test: hidden states and persona shifts are not isotropic random vectors.

category	persona	Llama ΔER	Llama ΔpER	Llama ΔIP	Llama ΔpIP	Qwen ΔER	Qwen ΔpER	Qwen ΔIP	Qwen ΔpIP	Mistral ΔER	Mistral ΔpER	Mistral ΔIP	Mistral ΔpIP
baseline	person	—	—	—	—	—	—	—	—	—	—	—	—
personality	*empathetic person*	+0.175	+0.013	−0.250	−0.013	+0.095	−0.007	+0.170	+0.016	+0.155	+0.008	+0.050	−0.002
	cynical person	−0.550	−0.055	−0.330	+0.025	−0.735	−0.035	+0.060	+0.039	−0.385	−0.035	+0.090	+0.017
gender	*female*	+0.030	+0.007	+0.070	−0.006	−0.055	−0.003	+0.010	+0.002	−0.030	−0.002	+0.040	−0.006
	male	−0.070	+0.001	+0.020	−0.002	−0.280	−0.006	−0.060	−0.001	−0.165	−0.015	+0.030	−0.002
race	*white person*	−0.040	+0.016	−0.050	−0.019	−0.110	+0.003	−0.040	−0.001	−0.130	−0.015	+0.070	−0.001
	black person	−0.105	+0.011	+0.480	+0.007	−0.100	+0.008	−0.060	−0.001	−0.130	−0.011	+0.050	−0.001
occupation	*psychotherapist*	−0.080	−0.009	−0.270	−0.029	−0.035	−0.025	+0.200	+0.005	−0.120	−0.021	+0.040	−0.006
	engineer	−0.175	+0.003	+0.210	−0.024	−0.270	−0.005	−0.090	−0.005	−0.125	−0.006	−0.010	−0.003
political	*Democratic supporter*	+0.020	+0.032	−0.240	−0.027	−0.175	−0.005	+0.030	+0.005	−0.045	+0.002	−0.030	−0.003
	Republican supporter	−0.035	+0.022	−0.350	−0.032	−0.290	−0.004	−0.100	+0.003	−0.165	−0.012	+0.020	−0.004

Table 18: Full cross-model persona effects on ER/IP labels and on $v_{\text{ER}}/v_{\text{IP}}$ cosine projections at layer 15 ($n = 200$ per persona), grouped by category. The ER columns are visualized in Figure 4.

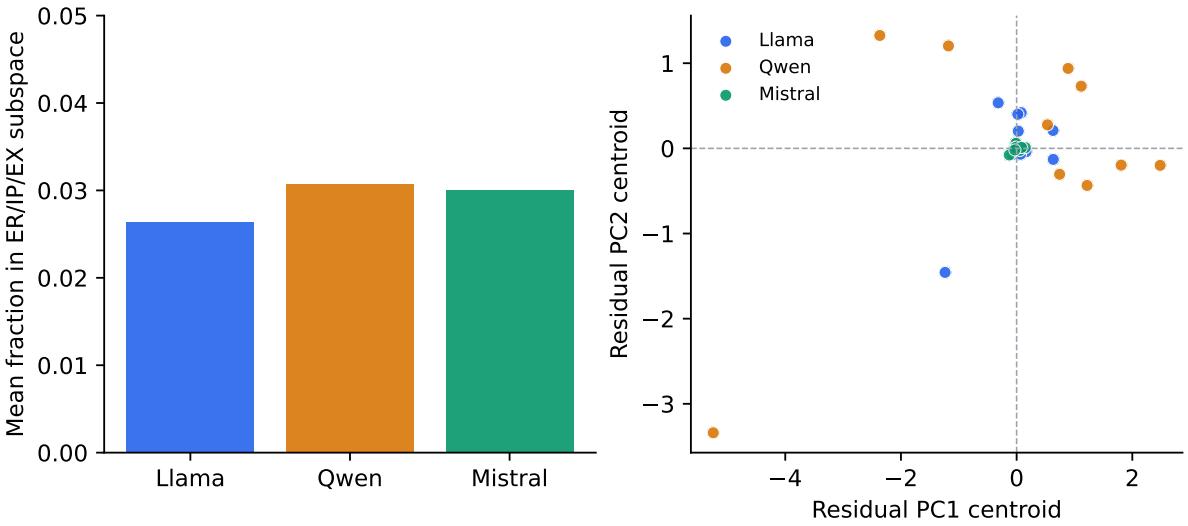


Figure 6: Persona-shift decomposition. Left: mean persona-shift squared-norm fraction captured by the ER/IP/EX mechanism subspace. Right: persona centroids in the first two residual-PC coordinates. Residual PCs are computed after removing the ER/IP/EX mechanism-subspace component.

Model	Sample	Mean	Resid.	Top-5 PCs
Llama	1.85	2.64	97.36	27.49
Qwen	2.63	3.08	96.92	29.60
Mistral	1.80	3.00	97.00	29.86

Table 19: Persona-shift decomposition at layer 15. Columns report the percentage of squared shift norm captured by the orthonormalized ER/IP/EX mechanism subspace, the residual percentage for persona-mean shifts, and the cumulative percentage of residual variance explained by the top five residual PCs.