Steering Personas Steers Safety: How Personality Trait Vectors Encode Safety-Critical Directions in LLMs

Anonymous ACL submission

Abstract

Persona-based customization of large language models is widely deployed, but its relationship to safety behavior remains under-characterized. We explore whether persona directions are geometrically coupled with safety-relevant directions, whether this coupling predicts side effects, and whether it is linearly removable. Across five open-weight models, we extract 5 personality and 3 safety vectors, build an 8×8 cosine matrix, evaluate steering side effects from two perspectives (LLM-judge expression scoring and benchmark outcome behavior), and test linearity with Gram-Schmidt, random-ablation, and naive-subtraction controls. We focus on three safety-relevant behavior classes: sycophancy, hallucination, and harmful-content compliance/refusal. We find systematic geometric coupling that predicts expression-level side effects, with Pearson r=0.57–0.73 across five models. We observe outcome-level shifts up to +52pp refusal and +22pp HaluEval failure, with frequent divergence between the two perspectives on the same trait pair. A single safety-orthogonal projection substantially weakens side effects in both perspectives, while random ablation preserves outcome-level shifts. Code and data are available online.[1]

Figure 1: Persona steering couples with safety-relevant behavior through linear geometry in activation space. Above waterline: explicit steering. Below waterline: side effects and geometric analysis not captured by persona-free evaluation.

1 Introduction

Persona-based customization of large language models has become a standard deployment technique. From non-player characters (NPCs) in role-playing games to virtual assistants in affective computing applications, practitioners increasingly use activation steering to induce target personality traits (Turner et al., 2023; Zou et al., 2023). The core operation is to extract a trait direction in activation space and inject it into the residual stream at inference time. Figure 1 illustrates this visible steering path above the waterline, and the hidden safety risks, activation-space geometry, and orthogonalization analysis below. Personality steering has shown stable control over persona expression across models, explaining its adoption in systems (Chen et al., 2025; Ma et al., 2025; Bhandari et al., 2026; Siu et al., 2025).

This practical success rests on an assumption that is rarely tested in deployment settings: semantic containment. If persona directions are contained, steering one trait should not systematically perturb safety behavior. Existing work already shows this assumption is fragile even within personality itself, where steering one trait can perturb others (Bhandari et al., 2026). A more consequential question

[1] https://anonymous.4open.science/r/persona-steering-anonymous-release-2E52

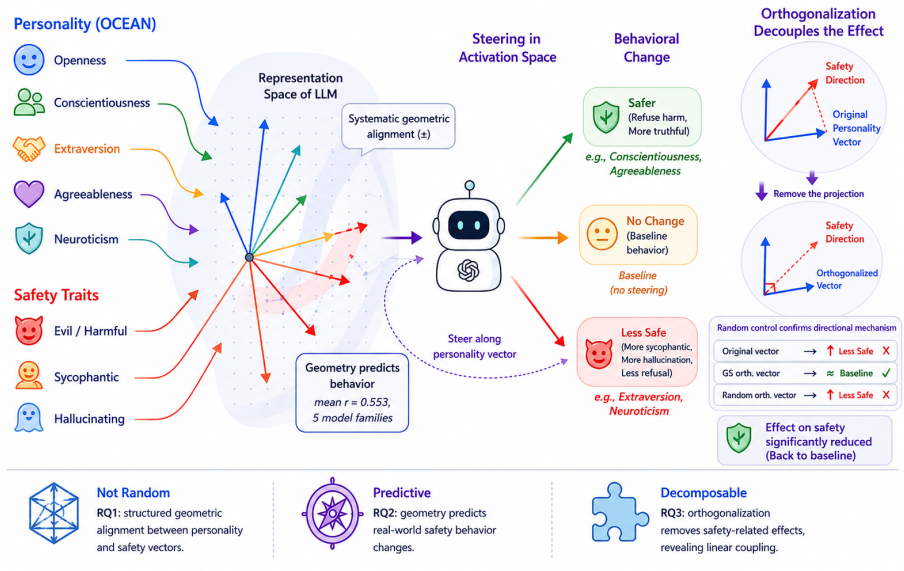

Figure 2: Overview of our framework. OCEAN and safety vectors are geometrically aligned in activation space, and this structure predicts expression-level coupling under persona steering (RQ1). Benchmarks show substantial outcome-level shifts that can diverge from LLM-judge expression-level shifts (RQ2). Gram-Schmidt weakens side effects across both evaluation perspectives, while random ablation preserves outcome-level effects, supporting linear decomposability (RQ3).

remains under-examined: whether persona directions are geometrically coupled with safety directions in the same activation space. If such coupling exists, persona steering is not only style control; it is also a latent safety intervention whose effects may be invisible to persona-free evaluations.

We therefore organize the paper around three research questions. **(RQ1)** Are persona and safety directions geometrically co-aligned across models, and does this structure predict observed behavioral coupling? **(RQ2)** Does persona steering induce substantial safety side effects on standardized benchmarks, and how do those effects relate to shifts measured by an LLM judge? **(RQ3)** Is the coupling linearly decomposable, such that a single intervention can reduce side effects across evaluation perspectives? Figure 2 gives the RQ-level roadmap; Sections 4.2–4.4 report the corresponding findings (F1–F3).

Following Zou et al. (2023) and Chen et al. (2025), we extract five persona and three safety vectors as mean activation differences between trait-positive and trait-negative response sets at a fixed layer, from five open-weight models spanning four families: LLaMA-3.1-8B-Instruct and LLaMA-3-8B-Instruct (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct (Jiang et al., 2023a), and Gemma-2-9B-Instruct (Gemma Team, 2024). We then construct a full 8×8 cosine matrix over trait directions, focus analysis on the OCEAN→Safety 5×3 submatrix, and evaluate steering-induced side effects with two complementary evaluation perspectives: an expression-level perspective (LLM-judge scores) and an outcome-level perspective (standardized benchmark behavior). Following steering-time safety evaluation practice, we focus on three safety-relevant behavior classes—sycophancy, hallucination, and harmful-content compliance/refusal (Siu et al., 2025; Chen et al., 2025)—and instantiate them with established benchmark protocols (Perez et al., 2022; Sharma et al., 2023; Li et al., 2023; Arditi et al., 2024). To test mechanism rather than correlation alone, we compare Gram-Schmidt orthogonalization with a norm-matched random-direction ablation control and a naive mean-subtraction baseline.

This design is mechanism-oriented. The geometric matrix tests whether coupling exists as a representational property before behavioral interpretation. The expression-level perspective provides a dense, cell-aligned signal for correspondence tests with cosine structure, while the outcome-level perspective tests whether the same steering perturbations survive under deployment-facing metrics. The orthogonalization controls separate targeted subspace removal from generic activation perturbation: if GS reduces side effects but random ablation does not, the evidence supports a directional

mechanism rather than a norm-only effect. This sequencing lets us move from structure, to behavior, to mechanism in one empirical pipeline.

Our analyses yield three findings. **(F1)** OCEAN personality vectors and safety directions show systematic geometric co-alignment in activation space, and this structure predicts the magnitude of expression-level coupling, with cross-model Pearson r=0.57–0.73 against LLM-judge measurements. **(F2)** Side effects are not limited to expression: standardized benchmarks reveal large outcome-level perturbations (up to +52pp refusal and +22pp HaluEval failure), and the two perspectives can diverge in direction or magnitude on the same trait pair. **(F3)** A single Gram-Schmidt projection against the safety subspace substantially weakens side effects in both perspectives, including divergent baseline cases; random ablation preserves outcome-level shifts, while a naive subtraction baseline introduces new off-target stylistic biases. A complementary check further shows that the geometric coupling is not explained by prompt-level lexical similarity.

2 Related Work

Activation steering and representation control. LLM behaviors can often be manipulated through directions in activation space. Early work extracted latent steering vectors from pretrained models (Subramani et al., 2022); activation addition showed that contrastive activation differences injected into the residual stream can steer behavior without parameter updates (Turner et al., 2023). Representation Engineering generalized this paradigm to model-internal concept monitoring and control (Zou et al., 2023), and CAA extended it to complex alignment-relevant behaviors such as sycophancy (Rimsky et al., 2024). These methods are consistent with evidence for approximately linear concept representations (Park et al., 2024), including truth and refusal directions (Marks and Tegmark, 2024; Arditi et al., 2024). However, superposition suggests that features may share representational dimensions, making steering directions entangled or polysemantic (Elhage et al., 2022).

Personality and psychological steering. LLMs can exhibit measurable and controllable personality traits. Psychometric studies find stable Big Five and HEXACO profiles in GPT-3 and other LLMs (Miotto et al., 2022; Jiang et al., 2023b, 2024; Serapio-García et al., 2023). Recent work moves from prompt-level personality expression to representation-level control: persona vectors monitor and steer character traits such as sycophancy and hallucination propensity (Chen et al., 2025), while DRESS edits style-related subspaces for stylized question answering (Ma et al., 2025). Most relevantly, Big Five steering directions have been shown to geometrically interfere with one another, even after orthogonalization (Bhandari et al., 2026), suggesting that personality traits may not be independent control knobs.

Safety behaviors and interference. Sycophancy, hallucination, and refusal are central safety-relevant behaviors. Prior work shows that RLHF and preference models can encourage sycophantic responses (Perez et al., 2022; Sharma et al., 2023); HaluEval benchmarks hallucination detection (Li et al., 2023); and refusal can be mediated by a single activation direction (Arditi et al., 2024). More recent studies show that steering can induce non-target effects: personality shaping affects safety and capability benchmarks (Fitz et al., 2025), PsySET evaluates psychological steering trustworthiness (Banayeeanzade et al., 2025), SteeringSafety finds cross-dimensional side effects of representation steering (Siu et al., 2025), and Persona Non Grata shows that different persona-induction methods expose different vulnerabilities (Li et al., 2026). Building on these findings, we analyze whether OCEAN personality directions are geometrically coupled with safety-critical behavior directions in activation space, whether this coupling explains cross-domain interference at both stylistic and functional evaluation perspectives, and whether the coupling can be linearly decomposed and removed. Relative to Siu et al. (2025) and Li et al. (2026), we additionally quantify the activation-level coupling and show that a single linear projection works across both evaluation perspectives.

3 Method

3.1 Persona Vector Extraction

Under the linear representation hypothesis, trait t corresponds to an approximately linear direction $\mathbf{v}_t$ in activation space. The contrastive mean-difference method provides an estimate of this direction:

$$\mathbf{v}_t = \mathbb{E}[\mathbf{h}_t^+] - \mathbb{E}[\mathbf{h}_t^-] \quad (1)$$

where $\mathbf{h}_t^+$ and $\mathbf{h}_t^-$ are the hidden states of high-trait and low-trait responses respectively, extracted at

layer 16 and averaged over response tokens. At inference time, $\alpha \cdot \mathbf{v}_t$ is injected into the residual stream with a uniform steering coefficient of $\alpha = 1.5$. All five models are operated at layer$\tilde{1}$6, a mid-network region across architectures (50% depth for LLaMA-3/3.1 and Mistral, 44% for Qwen3, 38% for Gemma-2) that is consistently associated with abstract semantic representations rather than surface syntactic features (Park et al., 2024; Rimsky et al., 2024).

3.2 Geometric–Behavioral Coupling Analysis

If personality direction $\mathbf{v}_i$ and safety direction $\mathbf{v}_j$ have non-zero cosine similarity in activation space, a linear intervention along $\mathbf{v}_i$ will produce a non-zero component along $\mathbf{v}_j$, inducing behavioral side effects. This geometric–behavioral correspondence constitutes our core hypothesis, examined at two levels.

Geometric level. We compute the full 8×8 cosine similarity matrix $C \in \mathbb{R}^{8\times 8}$, where $C_{ij} = \cos(\mathbf{v}_i, \mathbf{v}_j)$, quantifying the geometric alignment between all pairs of trait directions, and focus on the OCEAN→Safety submatrix.

Behavioral level: two complementary evaluation perspectives. We measure steering-induced behavioral side effects at two complementary evaluation perspectives. The *expression-level perspective* uses an LLM judge to score the dispositional and stylistic shifts in how the model expresses safety-relevant content, providing a continuous behavioral signal at matching granularity to the cosine matrix:

$$\Delta_{\text{behav}}(i \to j) = \text{Score}_j(\text{steer}_i) - \text{Score}_j(\text{baseline}). \tag{2}$$

The *outcome-level perspective* uses standardized safety benchmarks covering sycophancy compliance, hallucination identification, and harmful-content refusal, measuring the model's actual behavior on test inputs at granularity comparable to prior work. We compute the Pearson correlation between the cosine matrix and the LLM-judge matrix across all 15 cells per model; the relationship between geometric alignment and the *functional* benchmark matrix is treated as a separate two-perspective characterization in Section 4.3.

3.3 Orthogonalization as an Analytical Instrument

If persona–safety coupling is linear, decomposing $\mathbf{v}_t$ into a component within the safety subspace and a component in its orthogonal complement, and removing the former, should substantially reduce the vector's influence on safety directions. We define the safety subspace $\mathcal{S} = \text{span}(\mathbf{v}_{\text{evil}}, \mathbf{v}_{\text{syco}}, \mathbf{v}_{\text{hallu}})$ and the orthogonalized vector as:

$$\mathbf{v}_t^{\perp} = \mathbf{v}_t - \text{Proj}_{\mathcal{S}}(\mathbf{v}_t). \tag{3}$$

We compare three implementations: (i) *Gram-Schmidt (GS)* constructs an orthonormal basis $\{e_1, e_2, e_3\}$ via Gram-Schmidt on $\{\mathbf{v}_{\text{evil}}, \mathbf{v}_{\text{syco}}, \mathbf{v}_{\text{hallu}}\}$ with $\text{Proj}_{\mathcal{S}}(\mathbf{v}_t) = \sum_{k=1}^{3} \langle \mathbf{v}_t, e_k \rangle e_k$; (ii) *SVD* constructs a low-rank approximate basis via singular value decomposition; and (iii) a *norm-matched random ablation* replaces the safety vectors with random unit vectors as a null baseline. The random ablation control is the critical condition for testing linearity: if behavioral side effects are substantially reduced after GS/SVD but preserved after random ablation, the differential pattern is consistent with a linear decomposition account. The same logic applies across evaluation perspectives; if a single linear projection reduces side effects at both the stylistic and outcome-level perspectives, even on cases where the two perspectives diverge at baseline, both layers' perturbations flow through the same projection. Orthogonalization is used here as an analytical instrument for probing the geometric structure of the coupling, not as a deployment-ready safety solution.

4 Experiments

Section 4.1 describes the experimental setup; Section 4.2 verifies geometric coupling and its prediction of stylistic behavioral coupling (RQ1); Section 4.3 quantifies behavioral side effects at both evaluation perspectives and characterizes their divergence (RQ2); Section 4.4 probes the linear decomposability of the coupling (RQ3).

4.1 Experimental Setup

Data and vectors. For each trait, we use LLaMA-3.1-8B-Instruct to generate a contrastive dataset of five positive/negative steering instruction pairs and 40 test questions, split into 20 for vector extraction and 20 for behavioral evaluation. Model responses are generated under both positive and negative instructions with greedy decoding, and the mean hidden states across response tokens are extracted at layer 16. A quality filter retains positive samples with trait score ≥ 50, negative samples with trait score ≤ 50, and both with coherence score ≥ 50.

Behavioral evaluation. The LLM judge uses LLaMA-3.1-8B-Instruct (served locally via vLLM), with trait-specific evaluation prompts and 0–100 scores obtained via logprob aggregation over numeric tokens (responses are treated as invalid when the total probability mass on numeric tokens falls below 0.25). We use an open-source judge to ensure full reproducibility and to avoid version drift of proprietary APIs; since the analysis relies on relative differences before and after steering, conclusions are invariant under monotonic rescaling of the judge's absolute scale. Standardized benchmarks cover three dimensions: sycophancy-eval (200 pressure questions, compliance rate), HaluEval (200 questions, undetected hallucination rate), and AdvBench/HarmBench (50 harmful prompts each, refusal rate). All responses use greedy decoding.

Models. Experiments cover five open-weight models from four families: LLaMA-3.1-8B-Instruct and LLaMA-3-8B-Instruct (Meta) (Grattafiori et al., 2024), Qwen3-8B (Yang et al., 2025), Mistral-7B-Instruct (Jiang et al., 2023a), and Gemma-2-9B-Instruct (Gemma Team, 2024). Including two LLaMA-family versions allows us to verify that our findings are not artifacts of a single training pipeline. All models extract and steer at layer 16 with coefficient 1.5. LLaMA-3.1-8B-Instruct serves as the primary reporting model; the other four are used for cross-model verification.

4.2 Geometric Coupling and Stylistic Predictability (F1)

Geometric alignment structure. Table 1 reports the OCEAN→Safety cosine similarity submatrix across the five models; the full 8×8 matrix is in Appendix A. Personality and safety directions exhibit structured geometric alignment rather than random distribution. The strongest positive alignments appear in neuroticism→evil (0.534 on LLaMA-3.1) and openness→hallucination (0.487); the strongest negative alignments appear in agreeableness→evil (-0.497) and conscientiousness→sycophancy (-0.376). These alignments are semantically interpretable: the association between neuroticism and negative affect, the opposition between agreeableness and antagonistic behavior, and the inverse relationship between conscientiousness and sycophantic compliance all align with established characterizations of OCEAN traits, indicating that the geometric structure reflects the model's internal encoding of personality–safety relationships.

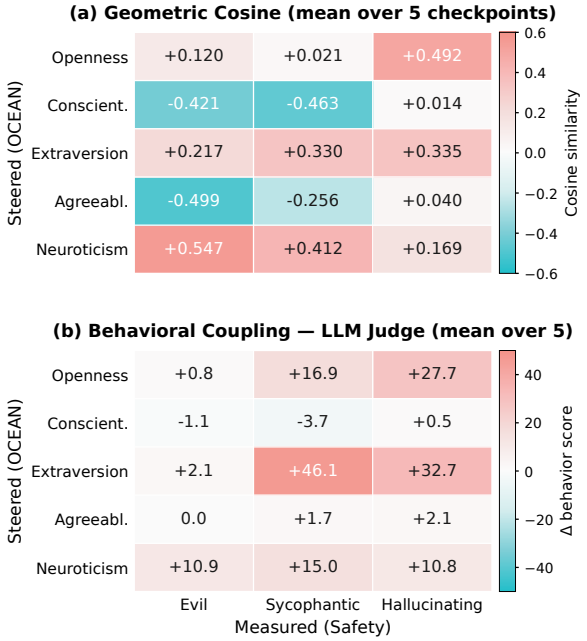

Figure 3: Geometric (a) and behavioral (b) coupling, averaged across all five models. **(a)** OCEAN→Safety cosine similarity; per-cell standard deviation is at most 0.13 (median 0.07). **(b)** LLM-judge OCEAN→Safety behavioral coupling. Per-model matrices in Tables 1 and 12; full data in Appendices A and B.

Cross-model consistency. The geometric pattern is highly consistent across all five models. The two strongest cells in Table 1, neuroticism →evil and agreeableness→evil, appear on every model with consistent sign and similar magnitude (range $+0.489$ to $+0.624$ for the former, -0.535 to -0.445 for the latter). This consistency indicates that the geometric alignment reflects a representational regularity acquired across architectures and training regimes.

Geometric structure predicts stylistic behavioral coupling. To test whether the geometric alignment translates into measurable behavioral coupling, we correlate the OCEAN→Safety cosine matrix with the LLM-judge behavioral coupling matrix (defined in Section 4.3) across all 15 cells, separately for each model. Table 2 reports the Pearson and Spearman coefficients. Across all five models, the geometric matrix significantly predicts stylistic behavioral coupling: Pearson r ranges from 0.570 (Gemma-2, p=2.65e−2) to 0.727 (Mistral, p=2.16e−3); all five models are significant at $p < 0.05$. We restrict this correspondence test to

OCEAN Trait	Safety	LLaMA-3.1	LLaMA-3	Qwen3	Mistral	Gemma-2
Openness	→Evil	0.006	0.069	**0.309**	−0.000	0.217
	→Syco.	−0.021	−0.093	0.157	0.033	0.031
	→Hallu.	**0.487**	**0.534**	**0.457**	**0.436**	**0.544**
Conscient.	→Evil	**−0.339**	**−0.340**	**−0.514**	**−0.373**	**−0.537**
	→Syco.	**−0.376**	**−0.470**	**−0.553**	**−0.270**	**−0.647**
	→Hallu.	0.057	−0.079	0.128	0.037	−0.074
Extraversion	→Evil	0.143	0.187	**0.263**	0.171	**0.323**
	→Syco.	**0.261**	**0.259**	**0.392**	**0.319**	**0.421**
	→Hallu.	**0.359**	**0.317**	**0.255**	**0.395**	**0.347**
Agreeabl.	→Evil	**−0.497**	**−0.445**	**−0.535**	**−0.505**	**−0.513**
	→Syco.	**−0.275**	**−0.349**	−0.146	−0.166	**−0.343**
	→Hallu.	0.071	0.003	0.112	0.012	0.004
Neuroticism	→Evil	**0.534**	**0.489**	**0.624**	**0.514**	**0.573**
	→Syco.	**0.435**	**0.444**	**0.391**	**0.258**	**0.534**
	→Hallu.	0.104	0.158	0.087	**0.253**	0.243

Table 1: Cosine similarity between OCEAN personality vectors and safety-critical behavior vectors across five open-weight models. Boldface marks $|\cos| \geq 0.25$. Full 8×8 matrix in Appendix A.

Model	Pearson r	p-value	Spearman ρ
LLaMA-3.1	0.615	1.47e−2	0.693
LLaMA-3	0.577	2.43e−2	0.706
Qwen3	0.578	2.41e−2	0.719
Mistral	0.727	2.16e−3	0.879
Gemma-2	0.570	2.65e−2	0.743

Table 2: Geometric–stylistic behavioral correspondence on the safety-focused OCEAN→Safety block ($n = 15$): correlations between cosine similarity and LLM-judge behavioral coupling. Full-matrix robustness results are reported in Appendix G.1.

OCEAN	Evil	Syco.	Hallu.
Openness	−0.7	+10.6	+22.3
Conscient.	−1.3	−2.6	−5.4
Extraversion	+1.7	**+53.7**	**+33.8**
Agreeabl.	+0.3	+0.6	−4.8
Neuroticism	**+21.2**	+15.6	+5.1

Table 3: LLM-judge OCEAN→Safety behavioral coupling matrix on LLaMA-3.1. Cells report change in judge score relative to baseline.

the expression-level perspective because the LLM judge yields a continuous signal at matching granularity to the cosine matrix; the relationship between geometric alignment and benchmark behavior is examined separately in Section 4.3.

4.3 Behavioral Side Effects Across Two Evaluation Perspectives(F2)

Stylistic-perspective side effects. Table 3 reports the OCEAN→Safety behavioral coupling matrix on LLaMA-3.1 measured by the LLM judge. Multiple cells exhibit large stylistic coupling: Extraversion→Sycophantic (+53.7), Extraversion→Hallucinating (+33.8), Openness→Hallucinating (+22.3), Neuroticism→Evil (+21.2), and Neuroticism→Sycophantic (+15.6). Across all five models (Appendix B Table 12), the OCEAN→Safety average Δ per OCEAN vector is consistently positive: Extraversion averages +24–+30 across all five models, and Neuroticism averages +9–+15.

Functional-perspective side effects. Table 4 reports OCEAN→Safety effects on LLaMA-3.1 measured by standardized benchmarks (baselines: syco=24.5%, halu=46.0%, evil_refusal=10.5%). Six cells exhibit shifts of at least 10pp on standardized benchmarks. The largest are Neuroticism's +52pp shift in refusal rate, followed by Extraversion's +24pp refusal shift and +22pp HaluEval failure shift. Because standardized benchmarks assume an un-steered model, shifts of this magnitude would not be observed in any safety evaluation that does not actively probe steered configurations, even though they appear under any deployment that applies persona steering.

Cross-model functional verification (Appendix C) shows that functional-perspective side effects are also a general phenomenon, though their *direction* is modulated by model-specific safety alignment: Extraversion produces +24pp refusal on LLaMA-3.1 but −24.5pp refusal on Mistral (becoming less safe); Neuroticism produces +52pp refusal on LLaMA-3.1 but −11.5 and −22.5pp refusal on Qwen3 and Mistral. Func-

OCEAN	Δ_{syco}	Δ_{halu}	Δ_{refusal}
Openness	$+3.0$	$+13.0$	-1.5
Conscient.	$+4.5$	$+3.0$	$\mathbf{+18.5}$
Extraversion	$+1.5$	$\mathbf{+22.0}$	$\mathbf{+24.0}$
Agreeabl.	$+4.5$	$+14.0$	-9.0
Neuroticism	-0.5	-5.0	$\mathbf{+52.0}$

Table 4: Standard benchmark OCEAN→Safety effects on LLaMA-3.1. Cells report change relative to persona-free baseline. Δ_{refusal} is the change in refusal rate on harmful prompts; positive values indicate safer behavior.

Metric pair	Pearson r
Per-dimension judge vs benchmark	
judge_evil vs evil_refusal*	-0.657
judge_syco vs syco_answer	-0.315
judge_hallu vs halueval	$+0.556$
Judge average vs each benchmark	
judge_avg vs evil_refusal*	-0.534
judge_avg vs syco_answer	-0.227
judge_avg vs halueval	$+0.400$

Table 5: Stylistic–outcome-level perspective agreement (LLaMA-3.1) measured as Pearson correlation between paired judge and benchmark deltas across OCEAN traits. * Refusal rate is sign-flipped so that positive values consistently indicate "more unsafe."

tional perturbation is universal; its sign depends on how each model has been safety-aligned.

Two evaluation perspectives diverge in direction and magnitude. The expression-level and outcome-level perspectives are not redundant measurements. Table 5 reports the Pearson correlation between paired judge and benchmark deltas: $\text{judge}_{\text{evil}}$ vs evil_refusal is -0.657 (strongly *negative*), $\text{judge}_{\text{syco}}$ vs syco_answer is -0.315 (negative), and $\text{judge}_{\text{hallu}}$ vs halueval is $+0.556$ (positive). Only the hallucination dimension shows agreement; the evil and sycophancy dimensions are anti-correlated across the two layers. The cleanest example of direction divergence is Neuroticism→Evil on LLaMA-3.1: the judge measures a $+21.2$ rise in evil score, while the benchmark measures a $+52$pp rise in refusal rate (i.e., the model becomes *safer*). A plausible mechanism is that steering Neuroticism makes the response style more anxious, dark, and defensive (captured by the judge) while making the model more wary of potentially harmful requests, increasing refusal (captured by the benchmark). A single geometric coupling (neuroticism→evil $\cos = 0.534$, consistent across all five models) manifests as opposite-direction shifts at the two evaluation perspectives. Even when the two perspectives agree in direction, magnitudes can be mismatched: Conscientiousness→Evil registers as near-zero on the judge (-1.3) but the largest entry in its benchmark row ($+18.5$pp refusal); Agreeableness is similar ($+0.3$ judge vs -9pp refusal). This two-perspective divergence implies that any single-axis safety evaluation systematically under- or over-estimates persona-induced risk.

4.4 Linear Decomposability and Removal (F3)

If coupling is linear, removing the persona vector's projection onto the safety subspace should substantially reduce its influence on safety directions. We test this with three orthogonalization variants and focus on three questions: (i) does a single linear projection reduce side effects at both evaluation perspectives? (ii) is this distinguishable from norm-matched random ablation and from a naive subtractive baseline? (iii) does the effect hold across models?

A single projection reduces side effects at both perspectives. Table 6 reports the largest stylistic-perspective and functional-perspective side effects on LLaMA-3.1 under Original, GS, and Random conditions. GS substantially reduces the largest stylistic-perspective effects (Extraversion $+29.7 \rightarrow -3.9$, Neuroticism $+14.0 \rightarrow -9.4$, averaged across the three safety dimensions). On the outcome-level perspective the reduction is sharper: Extraversion → refusal moves from $+24$pp to $+1$pp, and the Neuroticism → refusal shift moves from $+52$pp to -5.5pp, both within a few points of baseline. The same GS projection reduces effects at both layers, including the Neuroticism row, which exhibits strong divergence between the two evaluation perspectives at baseline. That a single linear projection covers both layers indicates that the two perspectives' perturbations flow through the same projection of the persona vector onto the safety subspace.

Random ablation isolates the directional contribution. On the outcome-level perspective, Random preserves the side effects almost completely: Neuroticism → refusal remains at $+52$pp under Random (identical to Original), and Extraversion → refusal remains at $+23.5$pp. This distinguishes GS's targeted subspace removal from any generic activation-space perturbation. On the expression-level perspective, Random partially reduces effects (Extraversion $+29.7 \rightarrow +8.0$), an

expression-level perspective (judge avg across 3 safety dims)			
OCEAN	Original	GS	Random
Extraversion	+29.7	−3.9	+8.0
Neuroticism	+14.0	−9.4	+4.2
outcome-level perspective (benchmark)			
Extr → Δrefusal	+24.0	+1.0	+23.5
Extr → ΔHaluEval	+22.0	+18.0	+22.0
Neuro → Δrefusal	+52.0	−5.5	+52.0

Table 6: Orthogonalization on LLaMA-3.1. GS substantially reduces the largest side effects at both evaluation perspectives; norm-matched random ablation preserves the functional-perspective effects almost entirely. Full table in Appendix D.

asymmetry consistent with the two-perspective divergence: judge scores are sensitive to any sufficiently large activation-space perturbation, while functional benchmarks respond only to perturbations of specific safety-relevant circuits.

Naive subtraction introduces new biases. We compare GS against a naive subtractive baseline $\mathbf{v}_t - \beta\bar{\mathbf{v}}_{\text{safety}}$ with β=1.5 (Appendix E). On Neuroticism → refusal, naive subtraction achieves a comparable reduction (+52 → +3.5pp). However, it introduces a new stylistic bias not present in GS: Extraversion + $\text{Sub}_{\beta=1.5}$ raises Δ_{syco} from +1.5 to +6.0 (a 4× increase), while Extraversion + GS keeps it near baseline (+2.0). GS provides a more surgical removal; naive subtraction trades safety reduction for off-target perturbation.

Cross-model robustness and prompt semantics. GS reduces large stylistic side effects on Extraversion across all five models, from +24 to +30 down to between −12 and +2 (Appendix F, Table 21). The same pattern holds at the outcome-level perspective on the four models with large baseline refusal shifts, where GS reduces the shift magnitude by 57%–110% regardless of sign (Table 22, same appendix). For instance, Mistral's −22.5pp Neuroticism refusal shift drops to −4.5pp under GS yet remains at −22.5pp under random ablation, mirroring the LLaMA-3.1 pattern in the opposite-sign direction. As a final robustness check, we ask whether the geometric coupling can be explained by prompt-level lexical similarity. We compute the semantic embedding similarity of the steering instructions used to generate each contrastive trait pair and correlate it with the persona-vector cosine similarity. On LLaMA-3.1, the OCEAN→Safety correlation is $r = -0.547$ (p=0.035): prompt-similar trait pairs occupy *more divergent* geometric directions in activation space, not more aligned ones. This rules out surface lexical co-occurrence as the source of the geometric entanglement (full methodology in Appendix G) and indicates that the coupling reflects representational structure acquired during training.

Single-dimension refinement. Replacing the full-subspace projection with single-dimension orthogonalization yields a finer-grained picture. For Extraversion, removing only $\mathbf{v}_{\text{syco}}$ collapses Δ_{syco} from +53.7 to +2.9 while leaving Δ_{hallu} near baseline and preserving an additional 10 points of on-target persona expression relative to full-subspace orthogonalization. This selectivity does not generalize uniformly: for weakly-coupled OCEAN rows (Conscientiousness, Agreeableness) all orthogonalization variants erode persona expression, and for Neuroticism single-dimension orthogonalization against most safety directions overshoots into the opposite-trait direction. We therefore treat linear disentanglement as *conditional on coupling strength*. Appendix H reports the full 5 × 5 table and discloses a sign-flip artifact on the Conscientiousness→ Hallucinating row that qualifies the surgical-removal interpretation.

5 Conclusion

Persona steering and safety are not representationally independent in LLMs. Personality–safety coupling is geometrically structured, behaviorally consequential under both expression-level and outcome-level evaluations, and linearly decomposable via a single subspace projection.

These results imply three practical requirements for personalized LLM deployment. *First*, persona-free safety evaluation is insufficient: for example, Neuroticism steering on LLaMA-3.1 yields a +52pp refusal shift that is invisible in unsteered testing. Safety reporting should therefore include persona-injected configurations. *Second*, expression-level and outcome-level evaluations are complementary and may disagree in direction; LLM-judge scoring should therefore be paired with benchmarks. *Third*, because coupling is linearly decomposable, pre-deployment vector-level intervention may be a tractable direction for reducing persona-induced safety effects.

Limitations

Our study focuses on a controlled comparison across five 7–9B open-weight models, with a uniform extraction layer chosen for cross-model comparability and a fully open-source evaluation stack to ensure reproducibility. The analytical framework we develop is not specific to OCEAN personality or to the three safety dimensions studied here. Promising directions for future work include extending the geometric–behavioral correspondence test to larger model scales and additional architectures, exploring alternative trait taxonomies beyond OCEAN, and probing whether the same linear-decomposability structure governs other concept pairings (e.g., persona–capability, persona–bias). We use the orthogonalization procedure primarily as an analytical instrument for probing the geometric coupling, and the engineering work required to turn it into a fully deployment-ready safety tool—e.g., scheduling, robustness under prompt drift, interaction with RLHF—remains future work.

Our work also has several responsible-use limitations. Because activation steering can shift safety-relevant behaviors, the same techniques studied here could in principle be misused to amplify undesirable tendencies or weaken safeguards; we therefore present orthogonalization as an analytical probe rather than a deployment-ready mitigation. The study uses publicly available open-weight models and established safety benchmarks under their respective terms, and does not collect private user data or newly recruit human subjects. Some benchmark prompts involve harmful or offensive content by design, but they are used only for controlled offline evaluation and reported in aggregate. Finally, although we report the main experimental settings, our results are limited to fixed decoding, a fixed steering layer and coefficient, and local experiments on 7–9B-parameter models.

We adopt $\alpha = 1.5$ as the steering coefficient, $\beta = 1.5$ as the orthogonalization coefficient, and a single mid-network layer (layer 16) per model, following the persona-vector and activation-steering convention established by prior work (Chen et al., 2025; Rimsky et al., 2024; Zou et al., 2023). A coarse layer scan and a coefficient scan on the on-target persona expression were conducted at the pilot stage and informed these choices; a full cross-cell sensitivity sweep coupling these settings to the safety dimensions is left to future work.

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. *Preprint*, arXiv:2406.11717.

Amin Banayeeanzade, Ala N. Tak, Fatemeh Bahrani, Anahita Bolourani, Leonardo Blas, Emilio Ferrara, Jonathan Gratch, and Sai Praneeth Karimireddy. 2025. Psychological steering in llms: An evaluation of effectiveness and trustworthiness. *Preprint*, arXiv:2510.04484.

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

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *Preprint*, arXiv:2507.21509.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Stephen Fitz, Peter Romero, Steven Basart, Sipeng Chen, and Jose Hernandez-Orallo. 2025. Psychometric personality shaping modulates capabilities and safety in language models. *Preprint*, arXiv:2509.16332.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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. 2023a. Mistral 7B. *Preprint*, arXiv:2310.06825.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023b. Evaluating and inducing personality in pre-trained language models. *Preprint*, arXiv:2206.07550. NeurIPS 2023.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. *Preprint*, arXiv:2305.02547. Findings of NAACL 2024.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Wenkai Li, Fan Yang, Shaunak A. Mehta, and Koichi Onoue. 2026. Persona non grata: Single-method safety evaluation is incomplete for persona-imbued llms. *Preprint*, arXiv:2604.11120.

Xinyu Ma, Yifeng Xu, Yang Lin, Tianlong Wang, Xu Chu, Xin Gao, Junfeng Zhao, and Yasha Wang. 2025. DRESSing up LLM: Efficient stylized question-answering via style subspace editing. *Preprint*, arXiv:2501.14371. ICLR 2025.

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

Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. *Preprint*, arXiv:2209.14338.

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

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Dario Amodei, and 6 others. 2022. Discovering language model behaviors with model-written evaluations. *Preprint*, arXiv:2212.09251.

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.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Vincent Siu, Nicholas Crispino, David Park, Nathan W. Henry, Zhun Wang, Yang Liu, Dawn Song, and Chenguang Wang. 2025. SteeringSafety: A systematic safety evaluation framework of representation steering in llms. *Preprint*, arXiv:2509.13450.

Nishant Subramani, Nivedita Suresh, and Matthew 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.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

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. *Preprint*, arXiv:2310.01405.

A Full 8×8 Cosine Matrices

Table 1 reports the OCEAN→Safety submatrix for each model. Here we report the full 8×8 cosine similarity matrix between all eight trait vectors (5 OCEAN + 3 safety). The OCEAN→OCEAN block characterizes within-personality geometric structure; the Safety→Safety block characterizes within-safety structure. Full 8×8 matrices for the remaining four models are reported in Tables 7–11.

The full 8×8 matrix exposes two structural features not visible in the OCEAN→Safety submatrix. *Within the OCEAN→OCEAN block*, the five personality vectors are themselves geometrically structured: Openness–Extraversion ($+0.556$) and Conscientiousness–Agreeableness ($+0.522$) co-align, while Conscientiousness–Neuroticism (-0.528) and Agreeableness–Neuroticism (-0.456) are anti-aligned. This is consistent with prior reports that Big Five steering directions are not mutually independent (Bhandari et al., 2026) and indicates that within-OCEAN interference is a separate phenomenon from the OCEAN–safety coupling that is the focus of this paper. *Within the Safety→Safety block*, the three safety vectors are not mutually orthogonal ($\cos \in [0.22, 0.41]$). The Gram-Schmidt construction used in Section 4.4 accounts for this by orthonormalizing within the safety subspace before computing projections, so the analysis does not assume safety-vector independence.

	O	C	E	A	N	Evil	Syco	Hallu
O	1.000	+0.316	+0.556	+0.431	−0.057	+0.006	−0.021	+0.487
C	+0.316	1.000	+0.046	+0.522	−0.528	−0.339	−0.376	+0.057
E	+0.556	+0.046	1.000	+0.217	+0.195	+0.143	+0.261	+0.359
A	+0.431	+0.522	+0.217	1.000	−0.456	−0.497	−0.275	+0.071
N	−0.057	−0.528	+0.195	−0.456	1.000	+0.534	+0.435	+0.104
Evil	+0.006	−0.339	+0.143	−0.497	+0.534	1.000	+0.405	+0.217
Syco	−0.021	−0.376	+0.261	−0.275	+0.435	+0.405	1.000	+0.245
Hallu	+0.487	+0.057	+0.359	+0.071	+0.104	+0.217	+0.245	1.000

Table 7: Full 8×8 cosine similarity matrix on LLaMA-3.1-8B-Instruct.

	O	C	E	A	N	Evil	Syco	Hallu
O	1.000	+0.216	+0.476	+0.368	−0.098	+0.069	−0.093	+0.534
C	+0.216	1.000	−0.097	+0.582	−0.560	−0.340	−0.470	−0.079
E	+0.476	−0.097	1.000	+0.090	+0.223	+0.187	+0.259	+0.317
A	+0.368	+0.582	+0.090	1.000	−0.527	−0.445	−0.349	+0.003
N	−0.098	−0.560	+0.223	−0.527	1.000	+0.489	+0.444	+0.158
Evil	+0.069	−0.340	+0.187	−0.445	+0.489	1.000	+0.422	+0.301
Syco	−0.093	−0.470	+0.259	−0.349	+0.444	+0.422	1.000	+0.219
Hallu	+0.534	−0.079	+0.317	+0.003	+0.158	+0.301	+0.219	1.000

Table 8: Full 8×8 cosine similarity matrix on LLaMA-3-8B-Instruct.

	O	C	E	A	N	Evil	Syco	Hallu
O	1.000	−0.132	+0.647	+0.213	+0.376	+0.309	+0.157	+0.457
C	−0.132	1.000	−0.265	+0.472	−0.598	−0.514	−0.553	+0.128
E	+0.647	−0.265	1.000	+0.117	+0.470	+0.263	+0.392	+0.255
A	+0.213	+0.472	+0.117	1.000	−0.435	−0.535	−0.146	+0.112
N	+0.376	−0.598	+0.470	−0.435	1.000	+0.624	+0.391	+0.087
Evil	+0.309	−0.514	+0.263	−0.535	+0.624	1.000	+0.401	+0.192
Syco	+0.157	−0.553	+0.392	−0.146	+0.391	+0.401	1.000	+0.014
Hallu	+0.457	+0.128	+0.255	+0.112	+0.087	+0.192	+0.014	1.000

Table 9: Full 8×8 cosine similarity matrix on Qwen3-8B.

	O	C	E	A	N	Evil	Syco	Hallu
O	1.000	+0.074	+0.477	+0.345	+0.145	−0.000	+0.033	+0.436
C	+0.074	1.000	−0.215	+0.468	−0.462	−0.373	−0.270	+0.037
E	+0.477	−0.215	1.000	+0.091	+0.330	+0.171	+0.319	+0.395
A	+0.345	+0.468	+0.091	1.000	−0.340	−0.505	−0.166	+0.012
N	+0.145	−0.462	+0.330	−0.340	1.000	+0.514	+0.258	+0.253
Evil	−0.000	−0.373	+0.171	−0.505	+0.514	1.000	+0.201	+0.191
Syco	+0.033	−0.270	+0.319	−0.166	+0.258	+0.201	1.000	+0.268
Hallu	+0.436	+0.037	+0.395	+0.012	+0.253	+0.191	+0.268	1.000

Table 10: Full 8×8 cosine similarity matrix on Mistral-7B-Instruct.

	O	C	E	A	N	Evil	Syco	Hallu
O	1.000	+0.046	+0.598	+0.324	+0.190	+0.217	+0.031	+0.544
C	+0.046	1.000	−0.297	+0.534	−0.585	−0.537	−0.647	−0.074
E	+0.598	−0.297	1.000	+0.052	+0.438	+0.323	+0.421	+0.347
A	+0.324	+0.534	+0.052	1.000	−0.367	−0.513	−0.343	+0.004
N	+0.190	−0.585	+0.438	−0.367	1.000	+0.573	+0.534	+0.243
Evil	+0.217	−0.537	+0.323	−0.513	+0.573	1.000	+0.488	+0.291
Syco	+0.031	−0.647	+0.421	−0.343	+0.534	+0.488	1.000	+0.134
Hallu	+0.544	−0.074	+0.347	+0.004	+0.243	+0.291	+0.134	1.000

Table 11: Full 8×8 cosine similarity matrix on Gemma-2-9B-Instruct.

OCEAN	Evil	Syco.	Hallu.
Openness	-0.2	$+12.6$	$+24.8$
Conscient.	-1.2	-4.5	$+3.5$
Extraversion	$+1.4$	$\mathbf{+50.1}$	$\mathbf{+25.2}$
Agreeabl.	-0.2	-0.2	$+5.1$
Neuroticism	$+2.8$	$\mathbf{+20.2}$	$+4.3$

Table 13: LLM-judge OCEAN→Safety behavioral coupling matrix on LLaMA-3-8B-Instruct.

OCEAN	Evil	Syco.	Hallu.
Openness	$+4.0$	$\mathbf{+37.0}$	$\mathbf{+42.5}$
Conscient.	-1.0	-9.3	$+2.5$
Extraversion	$+6.2$	$\mathbf{+37.0}$	$\mathbf{+31.9}$
Agreeabl.	-0.4	$+1.8$	$+4.9$
Neuroticism	$+4.9$	$\mathbf{+15.6}$	$+7.9$

Table 14: LLM-judge OCEAN→Safety behavioral coupling matrix on Qwen3-8B.

OCEAN	Evil	Syco.	Hallu.
Openness	$+0.3$	$+2.6$	$\mathbf{+16.9}$
Conscient.	-1.2	$+0.1$	$+2.2$
Extraversion	$+0.9$	$\mathbf{+33.9}$	$\mathbf{+37.8}$
Agreeabl.	-0.1	$+1.9$	$+1.2$
Neuroticism	$\mathbf{+18.2}$	$+9.1$	$\mathbf{+18.6}$

Table 15: LLM-judge OCEAN→Safety behavioral coupling matrix on Mistral-7B-Instruct.

OCEAN	Evil	Syco.	Hallu.
Openness	$+0.7$	$\mathbf{+21.7}$	$\mathbf{+31.9}$
Conscient.	-0.7	-2.4	-0.1
Extraversion	$+0.5$	$\mathbf{+55.8}$	$\mathbf{+34.8}$
Agreeabl.	$+0.4$	$+4.3$	$+3.9$
Neuroticism	$+7.6$	$\mathbf{+14.4}$	$\mathbf{+17.9}$

Table 16: LLM-judge OCEAN→Safety behavioral coupling matrix on Gemma-2-9B-Instruct.

B Per-model LLM-Judge Behavioral Matrices

Table 12 reports the OCEAN→Safety average Δ per OCEAN vector for each of the five models; Tables 13–16 report the full per-model 5×3 matrices.

Three patterns are consistent across all five models. *First*, Extraversion ($+24$ to $+30$) and Neuroticism ($+9$ to $+15$) produce the largest stylistic shifts on every model, matching the two strongest OCEAN→Safety geometric alignments identified in Table 1. The per-cell tables further reveal that the Extraversion effect is carried almost entirely by Sycophancy ($+33.9$ to $+55.8$) and Hallucinating ($+25.2$ to $+37.8$), with Evil shifts remaining small ($\leq +6.2$) on every model. *Second*, Conscientiousness is consistently near zero ($\Delta \in [-3.1, +0.4]$), despite its strong *negative* geometric alignment with both Evil and Sycophancy. This asymmetry, where strong negative geometric alignment translates into small stylistic shift, helps explain why the OCEAN→Safety correlation is moderate rather than near-perfect (Table 2, $n = 15$). *Third*, Openness shows the largest cross-model variance ($+6.6$ on Mistral, $+27.8$ on Qwen3), driven primarily by variance in its Sycophancy and Hallucinating couplings (Qwen3 reaches $+37.0$ and $+42.5$ respectively on these dimensions, while Mistral stays at $+2.6$ and $+16.9$).

C Cross-model Standardized-Benchmark Effects

Two patterns are notable. *First*, the existence of a functional shift is robust across all four models for both Extraversion and Neuroticism. Most cells show a non-trivial $|\Delta| \geq 2$pp, and the largest cells exceed 20pp. *Second*, the *sign* of the refusal shift varies across models in a way that the existence does not: Mistral shows the largest negative refusal shifts under both Extraversion (-24.5pp) and Neuroticism (-22.5pp), while LLaMA-3.1 shows the largest positive shifts on these same rows. This is consistent with the interpretation that persona steering reliably perturbs the safety-relevant circuits identified by our geometric analysis, but the direction of the perturbation is modulated by model-specific safety-alignment training. The cross-model universality of *magnitude* combined with model-specific modulation of *sign* is itself an actionable finding for practitioners: persona steering will perturb safety on any model they deploy, but the direction of the perturbation cannot be predicted without per-model testing.

OCEAN avg Δ	LLaMA-3.1	LLaMA-3	Qwen3	Mistral	Gemma-2
Openness	+10.7	+12.4	+27.8	+6.6	+18.1
Conscient.	−3.1	−0.8	−2.6	+0.4	−1.1
Extraversion	+29.7	+25.6	+25.0	+24.2	+30.4
Agreeabl.	−1.3	+1.6	+2.1	+1.0	+2.8
Neuroticism	+14.0	+9.1	+9.5	+15.3	+13.3

Table 12: Cross-model LLM-judge OCEAN→Safety average Δ per OCEAN vector, averaged over the three safety dimensions. Extraversion and Neuroticism are the two consistently large rows; Conscientiousness is consistently near zero.

D Full Orthogonalization Results on LLaMA-3.1

OCEAN	Original	GS	SVD	Random
Openness	+10.7	−5.8	−7.2	−3.0
Conscient.	−3.1	−7.7	−7.9	−8.4
Extraversion	+29.7	−3.9	−2.6	+8.0
Agreeabl.	−1.3	−6.3	−5.9	−6.6
Neuroticism	+14.0	−9.4	−9.3	+4.2

Table 18: Full orthogonalization on the stylistic (LLM-judge) layer for LLaMA-3.1. Values are average Δ across the three safety dimensions. GS and SVD produce nearly identical reductions; Random preserves the large positive rows.

On the three weakly-coupled rows (Openness, Conscientiousness, Agreeableness), Random and GS produce comparable shifts of similar magnitude ($|\Delta| \leq 8.4$ in both conditions), so the GS-versus-Random distinction cannot be observed where there is no large coupling to remove. On the two strongly-coupled rows (Extraversion, Neuroticism), Random preserves substantial positive magnitude ($+8.0$ and $+4.2$ respectively), in contrast to -3.9 and -9.4 under GS. The differential pattern is therefore visible only on the rows with non-trivial baseline coupling, which is the correct location to expect a targeted subspace removal to distinguish itself from generic norm-matched perturbation.

Condition	Δ_{syco}	Δ_{halu}	Δ_{refusal}
Extraversion (Original)	+1.5	+22.0	+24.0
Extraversion + GS	+2.0	+18.0	+1.0
Extraversion + Random	−1.0	+22.0	+23.5
Neuroticism (Original)	−0.5	−5.0	+52.0
Neuroticism + GS	−1.0	−8.0	−5.5
Neuroticism + Random	+0.5	−7.0	+52.0

Table 19: Standardized-benchmark orthogonalization on LLaMA-3.1. GS reduces the dramatic refusal-rate shifts on both Extraversion and Neuroticism close to baseline; Random ablation preserves the refusal shifts almost entirely.

The functional-perspective table makes the GS-versus-Random distinction sharper. On the standardized benchmarks, Random ablation preserves the $+24$pp Extraversion refusal shift and the $+52$pp Neuroticism refusal shift almost entirely (drops to $+23.5$ and $+52.0$ respectively), while GS brings both shifts within 5pp of baseline. This sharpness is consistent with the two-perspective characterization in Section 4.3: the stylistic-perspective (judge) score reflects general activation-space perturbation and is partially reduced by any sufficiently large intervention, while the functional-perspective (benchmark) behavior reflects only those perturbations that flow through specific safety-relevant circuits.

E Naive Subtractive Baseline

Condition	Δ_{syco}	Δ_{halu}	Δ_{refusal}
Extraversion			
Original	$+1.5$	$+22.0$	$+24.0$
+ Sub(β=1.5)	$+6.0$	$+19.0$	$+4.5$
+ GS	$+2.0$	$+18.0$	$+1.0$
Neuroticism			
Original	-0.5	-5.0	$+52.0$
+ Sub(β=1.5)	-3.5	-11.0	$+3.5$
+ GS	-1.0	-8.0	-5.5

Table 20: Naive subtraction $\mathbf{v}_t - \beta\,\bar{\mathbf{v}}_{\text{safety}}$ at β=1.5 compared with GS on LLaMA-3.1. The Extraversion row shows the key asymmetry: Sub raises Δ_{syco} from $+1.5$ to $+6.0$ ($4\times$), while GS keeps it near baseline at $+2.0$.

We report only β=1.5 in this comparison; smaller β produces correspondingly weaker reduction on the targeted refusal shift without alleviating the off-target stylistic bias. The key asymmetry between naive subtraction and GS is visible in the Extraversion block. Both methods bring the refusal shift close to baseline ($+1.0$ for GS, $+4.5$ for Sub), but naive subtraction raises Δ_{syco} from $+1.5$ to $+6.0$ ($4\times$), while GS keeps it near baseline at $+2.0$. The geometric explanation is that naive subtraction removes a single direction $\bar{\mathbf{v}}_{\text{safety}}$ rather than projecting onto the orthogonal complement of the safety subspace; the residual component of $\mathbf{v}_t$ within the safety subspace, but perpendicular to $\bar{\mathbf{v}}_{\text{safety}}$, is preserved and re-projects onto the orthogonal safety directions (Sycophancy in this case). GS, by construction, removes all such residual components.

F Cross-model Orthogonalization at Both evaluation perspectives

Model	Original	GS	Random
LLaMA-3.1	$+29.7$	-3.9	$+8.0$
LLaMA-3	$+25.6$	$+2.0$	$+11.0$
Qwen3	$+25.0$	-11.9	-8.9
Mistral	$+24.2$	-3.0	$+9.7$
Gemma-2	$+30.4$	-7.0	-5.8

Table 21: Cross-model orthogonalization on Extraversion (average Δ across the three safety dimensions, expression-level perspective). GS reduces the large original shifts to near zero on all five models; Random ablation does not on three of five.

The functional-perspective cross-model evidence reinforces the stylistic result in Table 21. The four informative cells (bold) span both signs of the original shift: LLaMA-3.1 shows the largest positive shifts under Extraversion $\rightarrow$ refusal ($+24.0$pp) and Neuroticism $\rightarrow$ refusal ($+52.0$pp), while Mistral shows the largest negative shifts on the same rows (-24.5pp and -22.5pp). On all four cells, GS reduces the *magnitude* of the shift by 57%–110% ($+24.0 \rightarrow +1.0$, $+52.0 \rightarrow -5.5$, $-24.5 \rightarrow -10.5$, $-22.5 \rightarrow -4.5$), whereas norm-matched random ablation reduces the magnitude by at most 7% on all four ($+23.5$, $+52.0$, -23.0, -22.5 respectively). This is consistent with the interpretation that GS removes the safety-aligned component of the persona steering vector, irrespective of how the model's safety-alignment training routes this perturbation into observed behavior. The five small-magnitude cells (e.g. LLaMA-3's $+1.5$pp Extraversion shift, Gemma-2's -3.0pp Neuroticism shift) do not distinguish GS from random because there is no behavioral shift large enough for either to substantially reduce.

GS reduces the Extraversion stylistic shift to within ± 12 of zero on every model, despite original shifts spanning a tight $+24.2$ to $+30.4$ range. Random ablation produces an inconsistent pattern: it overshoots toward negative on two models (Qwen3 -8.9, Gemma-2 -5.8) and only partially reduces on the other three ($+8.0$ to $+11.0$). The qualitative GS-versus-Random distinction observed on LLaMA-3.1 generalizes to every model in the study. The variance in the magnitude of GS reduction across models (from -11.9 on Qwen3 to $+2.0$ on LLaMA-3) reflects model-specific differences in how tightly the OCEAN representations are coupled to the safety subspace, rather than a failure of the orthogonalization procedure itself.

G Prompt-Semantics Robustness Check

We embed the steering instructions used to generate each contrastive trait pair using a sentence encoder and compute pairwise cosine similarity, then correlate this with the persona-vector cosine similarity from Table 1. A non-zero positive correlation would suggest that geometric coupling is partly an artifact of prompt-level lexical overlap.

On LLaMA-3.1, the OCEAN$\rightarrow$ Safety prompt-vs-persona correlation is $r = -0.547$ (p=0.035, n=15). Prompt-similar trait pairs occupy *more divergent* geometric directions in activation space, not more aligned ones. This is inconsistent with a prompt-similarity

OCEAN	Method	LLaMA-3.1	LLaMA-3	Qwen3	Mistral	Gemma-2
Extraversion	Original	**+24.0**	+1.5	+2.5	**−24.5**	−1.0
	+ GS	**+1.0**	−0.5	+4.5	**−10.5**	−0.5
	+ Rand	**+23.5**	+2.0	+2.0	**−23.0**	−1.0
Neuroticism	Original	**+52.0**	+2.5	−11.5	**−22.5**	−3.0
	+ GS	**−5.5**	0.0	−2.0	**−4.5**	−0.5
	+ Rand	**+52.0**	+2.0	−12.5	**−22.5**	−2.5

Table 22: Cross-model orthogonalization on the *functional* (standardized-benchmark) layer. Values are the steered minus baseline shift in refusal rate on the AdvBench+HarmBench harmful-prompt set (positive = safer). Bold cells mark the four high-magnitude configurations ($|\text{original}| \geq 12$pp); on these, GS reduces the magnitude of the shift to within 5–10pp of baseline regardless of the sign of the original shift, while random ablation preserves the shift almost entirely. The non-bold cells have small baseline shifts ($|\text{original}| < 12$pp), where any intervention produces similarly small values and the GS-vs-random distinction has no headroom to be observed.

account of the coupling.

The negative correlation deserves brief comment. A non-zero positive correlation would have indicated that prompt-similar trait pairs also occupy similar geometric directions, allowing the geometric coupling to be attributed to prompt-level word overlap. The observed negative correlation goes in the opposite direction: prompt-similar trait pairs occupy more divergent geometric directions. This is inconsistent with a prompt-similarity account of the coupling and instead suggests that the model has acquired representations that actively distinguish semantically adjacent traits, placing them in nearly opposite regions of activation space. We do not develop this interpretation further; for the purposes of this paper, it suffices that the geometric coupling is not explained by prompt-level overlap.

G.1 Full-Matrix Correlation Robustness

As a robustness check, we also compute correlations over all off-diagonal directed pairs in the full 8×8 vector-behavior matrix ($n = 56$), rather than only the safety-focused OCEAN→Safety block used in Table 2. The same positive association holds across all five models.

Model	Pearson r	p-value	Spearman ρ
LLaMA-3.1	0.581	2.7e−6	0.625
LLaMA-3	0.516	4.8e−5	0.593
Qwen3	0.480	1.8e−4	0.572
Mistral	0.574	3.8e−6	0.578
Gemma-2	0.616	4.3e−7	0.639

Table 23: Full-matrix robustness check: correlations between cosine similarity and LLM-judge behavioral coupling over all off-diagonal directed pairs in the complete 8×8 matrix ($n = 56$).

H Single-Dimension Orthogonalization

The orthogonalization analysis in Section 4.4 removes the projection of $\mathbf{v}_t$ onto the full safety subspace spanned by $\{\mathbf{v}_{\text{evil}}, \mathbf{v}_{\text{syco}}, \mathbf{v}_{\text{hallu}}\}$. Here we ask a finer question: does the side-effect on each safety direction decompose into a separately removable component, or is the coupling intrinsically multi-dimensional? We answer this by replacing the full-subspace projection with a *single-dimension* projection: for each (OCEAN, safety) pair we orthogonalize $\mathbf{v}_t$ against only one safety vector at a time, then re-evaluate the stylistic (LLM-judge) coupling across all three safety dimensions and the on-target OCEAN trait. All numbers below are LLaMA-3.1-8B-Instruct at layer 16 with $\alpha = 1.5$, matching the main analysis.

Selective decoupling under strong coupling. For Extraversion, the strongest-coupled OCEAN row in our study (+53.7 on syco under unaltered steering), single-dimension orthogonalization against $\mathbf{v}_{\text{syco}}$ reduces Δ_{syco} from +53.7 to +2.9 while leaving Δ_{hallu} at +1.5 and preserving TR(self) at +33.1. The analogous pattern holds for the other two single-dimension targets: orthogonalizing against $\mathbf{v}_{\text{hallu}}$ alone drops Δ_{hallu} from +33.8 to −7.8 while Δ_{syco} remains at +20.3. Compared to full-subspace orthogonalization, which collapses TR(self) from +57.4 to +23.0, single-dimension orthogonalization preserves an additional 10–17 points of persona expression. This is a finer-grained instance of the F3 result: when persona–safety coupling is strong, the coupling can be unwound one safety dimension at a time.

Erosion under weak coupling. For Agreeableness, every orthogonalization variant drives TR(self) negative. For Conscientiousness, orthog-

OCEAN	Orth. target	TR(self)	DE(safety)	Δ_{evil}	Δ_{syco}	Δ_{hallu}
Openness	— (original)	+4.6	+10.7	−0.7	+10.6	+22.3
	full (gs)	+8.9	+3.8	−1.6	+2.8	+10.2
	evil only	+11.7	+6.2	**−1.0**	+3.8	+15.8
	syco only	+15.5	+5.3	−2.7	**−1.7**	+20.3
	hallu only	+9.3	+1.9	−1.8	+0.6	**+7.0**
Conscient.	— (original)	+1.3	−3.1	−1.3	−2.6	−5.4
	full (gs)	+1.0	+2.5	−1.4	+5.0	+4.0
	evil only	+1.2	+1.0	**−1.3**	−1.7	+6.0
	syco only	+0.5	+1.3	−1.3	**+0.0**	+5.3
	hallu only	−8.9	−8.9	−6.5	−9.1	**−11.1**
Extraversion	— (original)	+57.4	+29.7	+1.7	+53.7	+33.8
	full (gs)	+23.0	−3.7	−6.1	+3.9	−8.8
	evil only	+39.6	+10.3	**−4.8**	+24.8	+10.9
	syco only	+33.1	−0.3	−5.5	**+2.9**	+1.5
	hallu only	+33.1	+2.9	−3.8	+20.3	**−7.8**
Agreeabl.	— (original)	+4.3	−1.3	+0.3	+0.6	−4.8
	full (gs)	−8.1	−8.0	−6.5	−7.1	−10.5
	evil only	−6.4	−7.2	**−5.5**	−8.7	−7.5
	syco only	−4.2	−8.0	−5.8	**−7.4**	−10.8
	hallu only	−7.3	−9.2	−4.8	−10.1	**−12.8**
Neuroticism	— (original)	+26.2	+14.0	+21.2	+15.6	+5.1
	full (gs)	−10.3	−9.8	−4.5	−7.7	−17.1
	evil only	−4.9	−8.9	**−4.1**	−1.7	−20.8
	syco only	−6.3	−8.0	−1.8	**−3.7**	−18.5
	hallu only	+4.1	−0.6	+7.5	+8.9	**−18.1**

Table 24: Single-dimension orthogonalization on LLaMA-3.1, stylistic (LLM-judge) layer. Each block reports an OCEAN trait under five conditions: original steering, full-subspace orthogonalization (gs), and three single-dimension variants. TR(self) is the on-target OCEAN trait score change; DE(safety) is the mean change over the three safety dimensions; Δ_* are per-dimension changes. Bolded cells mark the safety dimension that was actually orthogonalized against

onalization consistently erodes TR(self), but only the hallucinating-target condition flips it negative. For Neuroticism, original-steering TR(self) is large ($+26.2$) but all orthogonalization variants except the single-dimension hallucinating target overshoot into the opposite-trait direction (-5 to -10). The selectivity pattern from Extraversion does not generalize: when the persona vector lies sufficiently inside (or close to) the safety subspace, orthogonalization removes not only the safety projection but also components of the persona itself. We therefore characterize linear disentanglement as *conditional on coupling strength* rather than a universal property of the persona vectors.

A sign-flip artifact. One row deserves explicit disclosure. The Conscientiousness→ Hallucinating coupling under original steering is -5.4 (steering Conscientiousness mildly *reduces* the hallucinating score), but under every orthogonalization variant that does not target hallucinating it becomes positive ($+4.0$ to $+6.0$). The reverse, orthogonalizing against hallucinating, drives it back to -11.1 (below baseline). This sign-flip indicates that the orthogonalization operation is not a purely surgical removal of safety projection: removing one component re-routes residual coupling onto the orthogonal safety directions in a way that the linear-decomposability picture does not predict. We report this row because it qualifies the strength of the F3 claim; a full characterization of the residual-coupling geometry is left to future work.