Do Personality Traits Interfere? Geometric Limitations of Steering in Large Language Models

Pranav Bhandari[1,*], **Usman Naseem**[2], **Mehwish Nasim**[1,*]
[1]Network Analysis and Social Influence Modelling (NASIM) Lab
School of Physics, Mathematics and Computing
The University of Western Australia
[2]School of Computing, Macquarie University
[*]**Correspondence:** firstname.lastname@uwa.edu.au

Abstract

Personality steering in large language models (LLMs) commonly relies on injecting trait-specific steering vectors, implicitly assuming that personality traits can be controlled independently. In this work, we examine whether this assumption holds by analysing the geometric relationships between Big Five personality steering directions. We study steering vectors extracted from two model families (LLaMA-3-8B and Mistral-8B) and apply a range of geometric conditioning schemes, from unconstrained directions to soft and hard orthonormalisation. Our results show that personality steering directions exhibit substantial geometric dependence: steering one trait consistently induces changes in others, even when linear overlap is explicitly removed. While hard orthonormalisation enforces geometric independence, it does not eliminate cross-trait behavioural effects and can reduce steering strength. These findings suggest that personality traits in LLMs occupy a slightly coupled subspace, limiting fully independent trait control.

1 Introduction and Background

Large Language Models (LLMs) have demonstrated significant advances in their ability to exhibit personality traits (Jiang et al., 2024, 2023; Serapio-García et al., 2023), often aligned with the Big Five personality framework. Extensive prior work has explored the evaluation (Bhandari et al., 2025; Pellert et al., 2024), extraction (Jiang et al., 2024), and steering (Zhu et al., 2024; Chen et al., 2025; Bhandari et al., 2026) of personality traits in the literature. Recent advances in activation engineering have enabled the steering of LLM behaviour by injecting *activation vectors* into the model's residual stream to control the strength of trait expression. This approach offers several advantages over fine-tuning (Zhu et al., 2024), which is typically a heavy and resource-intensive process.

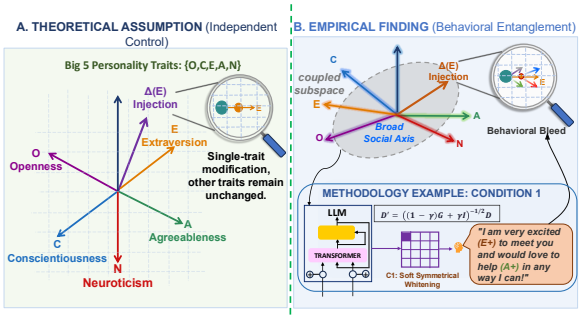

Figure 1: The **Independence Gap**. While theories assume traits can be steered independently (A), our geometric analysis shows LLMs learn coupled representations (B). Enforcing various geometric orthogonality fails to prevent behavioral bleed

Model behaviour can instead be adjusted at inference time using a *precise and controllable "knob"*, allowing for flexible and efficient personality modulation. This paradigm aligns with recent frameworks in mechanistic interpretability, which advocate direct interventions in internal computational structures to ensure that model behaviour remains consistent with intended human values (Naseem, 2026).

However, these studies do not emphasise the inherently entangled nature of personality traits, commonly referred to as the OCEAN traits (Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism), and instead draw conclusions based on the independent manipulation of individual traits. Our findings indicate that models do not learn human psychological constructs as orthogonal basis vectors; rather, they acquire entangled representations shaped by the correlations present in their training data. Empirically, we demonstrate that increasing a single dimension, such as Openness, does not result in an isolated change along that axis. Instead, it simultaneously increases other dimensions, including Agreeableness, Conscientiousness, and Extraversion, while decreasing Neu-

Findings of the Association for Computational Linguistics: ACL 2026, pages 29284–29293

roticism. These observations suggest that, rather than reflecting shifts along a single trait dimension, such changes are better understood as movements along a broader social axis.

Previous work has shown that high-level behaviours in large language models can often be approximated as linear directions in the activation space, enabling behaviour steering via difference-of-means or PCA-based vectors (Zou et al., 2023). However, such methods typically assume independence between concept directions. The concept-erasure literature studies how correlated information can be removed from representations, such as with Iterative Nullspace Projection (Ravfogel et al., 2020), introducing greedy, order-dependent projections, and LEACE: Perfect Linear Concept Erasure in Closed Form (Belrose et al., 2023), proposing a global, order-independent solution. Also, interpretability work such as Toy Models of Superposition (Elhage et al., 2022) shows that models often encode correlated concepts in superposition, suggesting that geometric orthogonality alone may not guarantee semantic independence. Such entanglement is a central challenge in mechanistic interpretability, as the superposition of features often leads to unintended interference during causal interventions (Naseem, 2026). Our work builds on these insights by applying both greedy and global Orthogonalisation strategies to personality steering and empirically evaluating their effect on cross-trait interference.

The overarching aim of this work is to systematically understand the personality trait vectors used in activation engineering. Prior activation engineering methods typically apply one trait direction at a time and evaluate whether the target trait shifts as intended (Sun et al., 2025; Yang et al., 2024; Bhandari et al., 2026), while leaving the effects on non-target traits unexamined. We systematically analyse the geometric relationships between Big Five personality directions and test how different constraint strategies, ranging from no constraints to full orthonormalisation, change both trait specificity and downstream behaviour. Rather than assuming that traits should be fully separable, we treat interference as an empirical phenomenon to be measured. Our goal is to understand when enforcing geometric independence helps, when it harms steering, and what this reveals about how personality traits are encoded inside the model.

We propose the following research questions:

1. **RQ1:** We hypothesise that personality steering directions are geometrically independent in large language models.

2. **RQ2:** What happens to steering effectiveness when geometric independence is explicitly enforced?

3. **RQ3:** Are observed trait dependencies consistent across different model families?

2 Methodology

The scope of this work focusses on understanding the behaviour of personality steering vectors and their interactions. We adopt the hybrid layer selection approach from (Bhandari et al., 2026) as our steering baseline, due to its effectiveness while preserving the model capacity. Our methodology then builds on this framework to analyse geometric dependencies between personality traits under different constraint settings.

Steering Mechanism. To control personality traits at inference time without updating model weights, we employ a multi-stage activation engineering framework. The process involves extracting trait directions, projecting them into a low-rank subspace, and applying them via a context-aware hybrid layer selection strategy simialr to (Bhandari et al., 2026).

Activation Extraction and Aggregation: For each Big Five trait $c \in \{O, C, E, A, N\}$, we utilise a subset of 20,000 instances from the Big-5-Chat[1] (Li et al., 2025) dataset, partitioned equally into high and low trait labels. For a given candidate layer L, the last non-pad residual state is extracted $h_L^{(i)}$ for each sequence and compute the normalised mean-difference vector $d_L^{(c)}$ between the high and low class means. Because transformer layers vary in their discriminative power, the learnt non-negative weights is applied $w_L^{(c)}$ to aggregate these layer-wise vectors into a single, robust per-trait direction $d^{(c)}$.

Low-Rank Subspace Projection: To reduce noise and capture shared personality structures, the aggregated directions are stacked to fit a rank-k PCA basis U_k. Each trait vector is then projected into this low-dimensional subspace and renormalised to unit length, yielding the final steering vector $\hat{d}^{(c)}$. This ensures the steering directions remain compact and stable during generation.

[1] Big-5-Chat Dataset

Hybrid Layer Selection: Rather than injecting the steering vector at a fixed middle layer, the hybrid strategy that balances static reliability with prompt-specific adaptability is used consistent with (Bhandari et al., 2026).

- *Offline Prior:* A stable, offline "best" layer is identified L_c^* using neutral probe prompts. This is achieved by applying a micro-perturbation and measuring the distributional shift at the next token using a weighted combination of L_2 distance, KL divergence, and token flip rate.

- *Dynamic Selection:* At runtime, the method evaluates how the current prompt activates different layers, calculating the per-layer representational shift to identify a dynamic candidate layer $\mathcal{R}(p, c)$.

During inference, steering is applied jointly at the verified offline layer and the dynamic layer using fixed mixture weights (0.8 and 0.2, respectively) to ensure stability while retaining context sensitivity.

Polarity Calibration and Inference Injection: To ensure the vector $\hat{d}^{(c)}$ accurately drives the desired semantic effect (e.g., high vs. low extraversion), the polarity is calibrated $sign^{(c)}$ by applying a small bidirectional steer on a neutral dataset and selecting the direction that maximizes KL divergence from the baseline distribution. Finally, at each decoding step, the scaled perturbation $\Delta^{(c)}(\alpha) = \alpha \cdot sign^{(c)}\hat{d}^{(c)}$ is added to the residual stream of the selected layers via forward hooks. The intensity parameter α is scaled relative to a global gain to maintain fluency and ensure reproducibility.

Trait Direction Conditioning (C0–C5). Let $\mathbf{d}_c \in \mathbb{R}^D$ denote the normalised weighted steering direction for trait c. Empirically, these directions are not independent and exhibit substantial cosine overlap. To study how geometric constraints affect steering behaviour and cross-trait interference, we construct the following conditioning schemes:

- **C0 (Baseline).** Original trait directions $\mathbf{d}_c$ are used without modification.

- **C1 (Soft Symmetric Whitening).** Directions are stacked into a matrix $\mathbf{D} \in \mathbb{R}^{k \times D}$ (where k is the number of traits) and transformed via a regularised Gram matrix $\mathbf{G} = \mathbf{D}\mathbf{D}^\top$: $\mathbf{D}' = ((1-\gamma)\mathbf{G} + \gamma\mathbf{I})^{-1/2}\mathbf{D}$, where $\gamma \in (0, 1)$ is a shrinkage parameter. This scales down off-diagonal correlations without forcing strict orthogonality, preserving more of the original shared geometry than hard whitening. This method is chosen as a least destructive first step in case the cross-trait bleed is caused by only the shared directional component.

- **C2 (Greedy Orthogonalisation).** A Gram–Schmidt procedure sequentially removes projections $\langle \mathbf{d}_i, \mathbf{d}_j \rangle \mathbf{d}_j$, yielding an orthonormal basis that is order-dependent. This is used to test whether pairwise perpendicularity reduces cross-trait bleed.

- **C3 (Selective Orthogonalisation).** Projection is applied only when the cosine similarity satisfies $|\cos(\mathbf{d}_i, \mathbf{d}_j)| > \tau$, preventing over-disentanglement while suppressing dominant overlaps. Based on the assumption that cross-trait bleed may be driven primarily by strongly aligned pairs, this scheme tests whether correcting only the largest geometric dependencies is sufficient

- **C4 (Soft Projection).** Correlated components are partially attenuated as $\mathbf{d}_i \leftarrow \mathbf{d}_i - \beta\langle \mathbf{d}_i, \mathbf{d}_j \rangle \mathbf{d}_j$ when $|\cos(\mathbf{d}_i, \mathbf{d}_j)| > \tau$, trading off disentanglement strength and retention. This setting tests whether partial correction provides a better balance between cross-trait bleed and overall fluency, as full removal may distort the semantic meaning of trait vectors.

- **C5 (Hard Symmetric Orthonormalisation).** A symmetric Löwdin transformation enforces $\mathbf{D}'\mathbf{D}'^\top = \mathbf{I}$, completely removing linear overlap in an order-independent manner. This transforms the entire set of trait vectors so that they are perfectly orthogonal to each other. If shared directional components are the primary cause of cross-trait bleed, this condition should provide the cleanest separation.

All conditions use the same steering injection and hybrid layer selection mechanism, isolating the impact of geometric constraints on steering efficacy and trait interference.

3 Evaluation

Personality steering is evaluated under three controlled settings: *base* (no steering), *positive steering*, and *negative steering*, where the latter two

	Llama-3-8B						Mistral-8B					
	C0 (Base)		C4 (Soft)		C5 (Hard)		C0 (Base)		C4 (Soft)		C5 (Hard)	
Target Trait	T	B_{max}	T	B_{max}	T	B_{max}	T	B_{max}	T	B_{max}	T	B_{max}
Openness	3.1	-3.5 (Neu)	3.0	-3.4 (Neu)	2.9	-3.0 (Neu)	3.3	3.3 (Agr)	3.2	3.4 (Agr)	3.1	2.8 (Agr)
Conscientiousness	2.9	-2.9 (Neu)	2.9	-2.7 (Neu)	2.9	2.6 (Agr)	2.2	-2.0 (Neu	2.3	2.0 (Agr)	2.4	2.0 (Agr)
Extraversion	3.0	-3.1 (Neu)	2.6	-2.4 (Neu)	3.0	-2.5 (Neu)	3.1	3.3 (Agr)	3.5	3.1 (Agr)	3.3	3.0 (Agr)
Agreeableness	3.3	2.8 (Con)	3.2	2.6 (Opn)	**3.7**	-3.1 (Neu)	2.7	-3.3 (Neu)	3.7	-3.0 (Neu)	2.8	-2.3 (Neu)
Neuroticism	3.1	-3.1 (Agr)	3.1	-3.2 (Agr)	3.2	-3.1 (Agr)	0.7	-1.2 (Ext)	0.0	-1.0 (Agr)	0.1	-1.1 (Ext)

Table 1: **Trait-Level Steering Contrast under Geometric Constraints.** Comparison of target steering contrast (T, Intended Target) and maximum cross-trait bleed ($B_{\max}$, Unintended Target) for LLaMA-3-8B and Mistral-8B under baseline (C0), soft-constrained (C4), and hard orthonormal (C5) trait vector constructions. Values report the difference between positively and negatively steered generations (High−Low) as measured by judge scores (1–5 scale). Parentheses indicate the trait responsible for $B_{\max}$.
T: Diagonal magnitude (Targeted Trait Steering). B_{max}: Maximum absolute off-diagonal value. **C4 (Soft)** uses $\beta = 0.5$. **C5 (Hard)** uses full symmetric Orthogonalisation. $(-)$ sign suggests the opposite nature of the trait effects.

apply trait-specific steering vectors of equal magnitude and opposite polarity. For each geometric condition (C0–C5), all other factors are held constant, including the learned subspace, layer weights, steering intensity, injection point, and decoding configuration. This isolates the effect of geometric constraints on personality vectors, ensuring that observed differences arise solely from vector structure rather than the steering mechanism itself.

Steering effectiveness is assessed using interview-style Big Five Inventory (BFI) questionnaires (Wang et al., 2024) consistent with (Bhandari et al., 2026). For direct comparison, first and second order statistics are reported. Beyond target trait shifts, we measure cross-trait responses to quantify inter-trait effects. Evaluation uses neutral prompts, ensuring observed personality changes arise solely from internal steering. We use Gpt-4o-mini as a judge, building upon the literature (Jiang et al., 2024; Frisch and Giulianelli, 2024) of using models as judges. Finally, we report fluency scores alongside personality metrics to verify that steering and geometric constraints do not degrade generation quality or general language behaviour.

4 Results

We conduct experiments on two instruction-tuned models from different architectural families – *LLaMA-3-8B-Instruct* and *Ministral-8B-Instruct*. For each model, baseline steering performance is quantified using the difference between positively and negatively steered generations (*high–low*), which serves as a reference point to compare how geometric constraints (C1–C5) alter both target-trait control and cross-trait interactions.

4.1 Geometric Independence of Personality Steering Directions

To evaluate whether personality steering directions are geometrically independent, we compare the *target steering strength* (T, diagonal entries) against the *maximum cross-trait bleed* ($B_{\max}$, largest absolute off-diagonal entry) under different geometric constraints (Table 1). Across both LLaMA-3-8B and Mistral-8B, steering a single trait produces non-negligible changes in at least one other trait, with $B_{\max}$ often comparable in magnitude to T. This pattern persists under both soft disentanglement (C4) and full symmetric orthonormalisation (C5). Notably, while C5 enforces near-zero pairwise cosine similarity between trait directions in activation space, it does not consistently reduce $B_{\max}$ in generation-level evaluations.

These results indicate that eliminating geometric overlap between steering vectors does not guarantee behavioural independence. Even when trait directions are orthonormal by construction, downstream generations continue to exhibit systematic cross-trait interactions. We therefore **reject the hypothesis in RQ1**, implying that personality steering directions in large language models are not geometrically independent in a behaviourally meaningful sense.

Table 1 reports the steering strength for conditions base(C0), C4 and C5. Additionally, we analyze *fluency* and *variance* scores under identical measurement conditions, using LLaMA as the reference model.

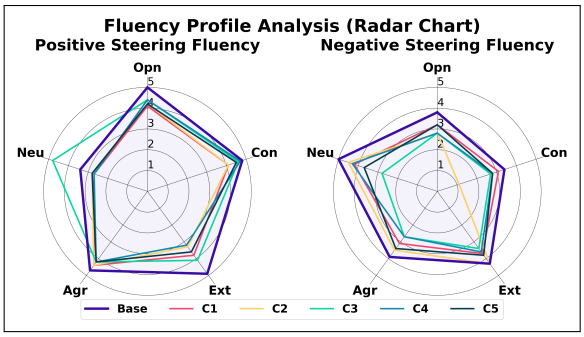

Figure 2: **Fluency profile analysis for conditions C0-C5** compared against Base steering. *Fluency degradation for both Positive and Negative steering can be observed across all traits for all the Conditional Methods used.* Although trait shifts that were comparable to the base values, the significant degradation of fluency suggests the need to use orthogonalised vectors carefully for steering purpose.

Steering strength remains largely conserved under progressive orthonormalisation (C1–C5)(Table 1). However, Figure 2 demonstrates how fluency scores consistently degrade as geometric constraints are enforced. Comparing to baseline fluency values reported in (Bhandari et al., 2026) (High/Low), Openness drops from $5.0/3.8$ to $4.1/3.2$ (C1), $4.3/2.8$ (C2), and remains around $4.4/2.8$ through C3–C5. Similarly, conscientiousness decreases from $4.8/3.5$ to $4.3/3.1$ in C1 following identical trends to Openness for other conditions, while Extraversion exhibits the largest decline, from $4.9/4.3$ to $3.8/3.6$.

These results indicate that although orthonormalisation preserves directional steering magnitude, it removes shared components necessary for fluent and expressive generation, leading to reduced variance and degraded output quality. Hence, to explain our **RQ2**: we conclude that explicitly enforcing geometric independence does not improve steering effectiveness and instead introduces a quality–control trade-off.

4.2 RQ3: Cross-Model Consistency of Trait Dependencies

To assess whether observed trait dependencies generalise across model families, we compare steering behaviour between *LLaMA-3-8B-Instruct* and *Mistral-8B-Instruct* under identical conditions (C0, C4, C5) using the same extraction, constraint, and evaluation pipeline. Across both models, we observe consistent geometric patterns: several traits exhibit substantial cross-trait bleed even after enforcing geometric constraints. For example, **Openness** shows high maximum bleed values in both models, with $B_{\max}$ remaining large under hard orthonormalisation (C5), e.g., ≈ 3.0 in LLaMA-3 and ≈ 2.8 in Mistral-8B. Similarly, **Extraversion** and **Agreeableness** continue to induce strong off-diagonal effects across conditions, indicating that these dependencies are not artifacts of a single model but reflect shared structure in personality representations.

At the same time, we observe clear model-specific modulation in steering responsiveness. Most notably, **Neuroticism** exhibits strong target steering in LLaMA-3 ($T \approx 3.1$–3.2 across C0–C5), whereas Mistral-8B shows low target response ($T \approx 0.0$–0.7) under the same conditions, despite comparable geometric treatment. Importantly, this suppression persists even when geometric independence is enforced (C5), suggesting that the absence of behavioural response cannot be attributed solely to vector entanglement. Together, these results indicate that while cross-trait dependencies are largely consistent across model families, their behavioural expression is shaped by model-specific training and alignment constraints rather than geometry alone. A detailed table for all the observations (C1-C5) is provided in Appendix A.

5 Conclusion

Given the popularity of steering methods in the literature, we systematically analysed the behaviour of steering vectors under various conditions. Our method considered the Big Five Personality traits, and we investigated whether personality steering directions in large language models are geometrically independent, and how enforcing geometric constraints affects the steering behaviours. Through the analysis across two model families (Llama and Mistral), we show that personality traits are not independent directions in activation space. Even when strong global constraints such as symmetric orthonormalisation are applied, steering one trait consistently induces measurable changes in other unintended traits, indicating persistent cross-trait dependencies.

Acknowledgment

Dr Mehwish Nasim acknowledges JTSI/Defence Science Centre's grant 2223R5CRG002, awarded to her in 2023.

Limitations

This work studies personality steering behaviour using a limited set of large language models and personality datasets, and future work could extend the analysis to a broader range of model families and trait representations. While we focus on Big Five traits and judge-based evaluation, additional datasets and alternative evaluation frameworks may reveal further structure in trait interactions. Our analysis relies on linear geometric constraints; exploring other orthogonalisation or projection methods could provide a more comprehensive understanding of trait disentanglement. Finally, we use LLMs as judges for behavioural assessment, and incorporating human evaluation or complementary metrics is left for future investigation.

Ethical Considerations

Steering large language models using latent vectors introduces ethical considerations, particularly when such steering is applied in uncontrolled or unsupervised settings. Steering vectors are learned approximations of complex behavioural traits and do not provide transparent or complete representations of the values they encode; as a result, unintended attributes or hidden information may be co-activated during steering. This raises concerns about value misalignment, especially in real-world deployments where subtle behavioural shifts could have downstream social or psychological impacts. Additionally, aggressive or poorly understood steering may bypass safety mechanisms or distort model behaviour in ways that are difficult to detect or reverse. These risks highlight the importance of careful evaluation, interpretability, and constraint-aware steering when modifying model behaviour.

References

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063.

Pranav Bhandari, Nicolas Fay, Sanjeevan Selvaganapathy, Amitava Datta, Usman Naseem, and Mehwish Nasim. 2026. Activation-space personality steering: Hybrid layer selection for stable trait control in LLMs. In *19th Conference of the European Chapter of the Association for Computational Linguistics (EACL'26))*.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 868–872.

Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. Persona vectors: Monitoring and controlling character traits in language models. *arXiv 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, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Ivar Frisch and Mario Giulianelli. 2024. Llm agents in interaction: Measuring personality consistency and linguistic alignment in interacting populations of large language models. *arXiv preprint arXiv:2402.02896*.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

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. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2025. Big5-chat: Shaping llm personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471.

Usman Naseem. 2026. Mechanistic interpretability for large language model alignment: Progress, challenges, and future directions. *arXiv preprint arXiv:2602.11180*.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2024. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, 19(5):808–826.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models.

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.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

Shu Yang, Shenzhe Zhu, Liang Liu, Lijie Hu, Mengdi Li, and Di Wang. 2024. Exploring the personality traits of llms through latent features steering. *arXiv preprint arXiv:2410.10863*.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2024. Personality alignment of large language models. *arXiv preprint arXiv:2408.11779*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A All detailed tables for C0-C5 for trait values

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	2.80	1.90	2.90	2.70	-3.00
Conscientiousness	1.44	3.11	-1.22	2.44	-2.00
Extraversion	2.00	-0.38	2.75	1.50	-2.75
Agreeableness	3.11	2.56	2.67	3.67	-3.22
Neuroticism	-1.50	-2.50	-1.63	-3.13	3.25

Table 2: **Condition C1 (Soft Symmetric Whitening).** Cross-trait impact of steering vectors on Llama-3-8B. The rows represent the *Targeted Trait* (steering vector applied), and columns represent the *Measured Trait* (judged output). Values indicate the shift in Likert score from High to Low values(High – Low). Note the high off-diagonal bleed, particularly between Openness and Extraversion.

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	2.50	2.20	2.30	1.80	-2.50
Conscientiousness	0.11	-3.00	2.33	0.22	-0.11
Extraversion	-2.00	0.63	-3.13	0.88	2.88
Agreeableness	1.89	0.67	0.67	2.67	-1.00
Neuroticism	-1.63	-2.88	-2.13	-3.38	3.63

Table 3: **Condition C2 (Greedy Orthogonalisation).**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.10	2.50	2.90	3.30	-3.40
Conscientiousness	1.56	2.89	-1.22	2.44	-2.75
Extraversion	2.00	-1.75	2.75	1.38	0.13
Agreeableness	2.56	2.67	1.67	3.22	-3.00
Neuroticism	-1.50	-2.88	-0.50	-3.63	2.75

Table 4: **Condition C3 (Selective Orthogonalisation).**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.00	2.60	2.90	3.30	-3.40
Conscientiousness	1.78	2.88	-1.22	2.44	-2.75
Extraversion	2.00	-0.88	2.63	1.38	-2.38
Agreeableness	2.56	2.33	1.67	3.22	-2.33
Neuroticism	-1.38	-2.63	-2.25	-3.25	3.13

Table 5: **Condition C4 (Soft Greedy Projection, $\beta = 0.5$).**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	2.90	1.90	3.00	2.70	-3.00
Conscientiousness	1.44	2.89	-1.11	2.56	-2.00
Extraversion	1.88	-0.38	3.00	1.25	-2.50
Agreeableness	2.89	2.78	2.44	3.67	-3.11
Neuroticism	-1.38	-2.50	-1.63	-3.13	3.25

Table 6: **Condition C5 (Hard Symmetric Orthonormalisation).**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.10	2.40	2.80	2.80	-2.50
Conscientiousness	1.11	2.44	0.44	2.00	-1.89
Extraversion	2.38	2.00	3.38	2.75	-1.75
Agreeableness	2.00	0.56	2.22	2.78	-2.22
Neuroticism	-0.50	-1.13	-1.13	-0.75	-0.13

Table 7: **Condition C1 (Soft Symmetric Whitening) on Mistral-8B.**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	2.50	2.10	2.40	2.10	-0.40
Conscientiousness	0.89	2.22	0.11	2.22	-1.67
Extraversion	2.13	1.75	3.13	2.88	-2.25
Agreeableness	1.56	0.44	2.00	2.89	-2.22
Neuroticism	-1.13	-1.50	-1.25	-1.00	-1.00

Table 8: **Condition C2 (Greedy Orthogonalisation) on Mistral-8B.**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.30	2.60	3.00	3.30	-2.30
Conscientiousness	1.11	2.22	0.89	2.00	-1.89
Extraversion	2.63	1.50	3.63	3.25	-2.13
Agreeableness	2.11	1.33	2.11	3.33	-2.89
Neuroticism	-0.25	-1.25	-0.38	0.13	-0.25

Table 9: **Condition C3 (Selective Orthogonalisation) on Mistral-8B.**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.20	2.40	3.10	3.40	-2.40
Conscientiousness	1.00	2.33	1.00	2.00	-1.67
Extraversion	2.50	1.88	3.50	3.13	-2.00
Agreeableness	2.11	1.89	2.44	3.67	-3.00
Neuroticism	-0.50	-0.88	-0.50	-1.00	0.00

Table 10: **Condition C4 (Soft Greedy Projection, $\beta = 0.5$) on Mistral-8B.**

	Measured Trait				
Targeted Trait	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**
Openness	3.10	2.30	2.70	2.80	-2.50
Conscientiousness	1.11	2.44	0.67	2.00	-1.56
Extraversion	2.38	2.13	3.25	3.00	-1.63
Agreeableness	2.00	0.56	2.22	2.78	-2.33
Neuroticism	-0.50	-0.88	-1.13	-0.50	-0.13

Table 11: **Condition C5 (Hard Symmetric Orthonormalisation) on Mistral-8B.**

C	Method	Geom. Independence	Signal Retention	Key Note
C1	Global Gram whitening	$\max \lvert\cos\rvert < 10^{-8}$	0.83–0.94	Perfect ortho; mild attenuation
C2	Strict QR (order-dependent)	$\max \lvert\cos\rvert < 10^{-8}$	-1.00–0.63	Order effects destroy trait semantics
C3	Selective removal (τ=0.5)	$\max \lvert\cos\rvert = 0.466$	0.63–1.00	Partial decorrelation; E most affected
C4	Soft removal (β=0.5, τ=0.5)	$\max \lvert\cos\rvert = 0.527$	0.85–1.00	Best trade-off; semantics largely preserved
C5	Hard orthonormal (global)	$\max \lvert\cos\rvert < 10^{-8}$	0.83–0.94	Perfect ortho; signal uniformly reduced

Table 12: Diagnostics of progressive orthonormalization constraints (C1–C5) applied to personality steering directions in LLaMA-3-8B. For each constraint, we report the achieved geometric independence (maximum absolute off-diagonal cosine similarity between trait directions), the range of signal retention relative to the unconstrained baseline, and a brief qualitative summary. C1 and C5 enforce global, order-independent orthonormality and achieve near-zero cosine overlap, but uniformly attenuate trait signal. C2 also enforces strict orthonormality, but its greedy, order-dependent construction leads to severe semantic degradation. C3 and C4 relax hard orthogonality by selectively or softly removing projections, preserving substantially more trait signal at the cost of residual geometric entanglement, with C4 exhibiting the best semantic–geometry trade-off.