Orthogonalizing Personality Vectors for Interference-Free Multi-Trait Modulation in LLMs

Anonymous Authors

Abstract

Personality vectors—parameter-space offsets derived by subtracting pre-trained model weights from fine-tuned counterparts—enable continuous and transferable personality modulation in large language models (LLMs). However, when multiple personality vectors are merged simultaneously, cross-trait interference degrades modulation fidelity: vectors targeting one Big Five trait inadvertently perturb others, reducing the trait-score correlation from $\rho \approx 0.9$ (single trait) to $\rho \approx 0.58$ (five traits simultaneously) (Sun et al., 2025). We introduce the **Cross-trait Interference Matrix** (C-matrix) as a quantitative evaluation framework, where diagonal entries measure target-trait effects and off-diagonal entries capture unwanted interference in BFI score units. Using this framework, we systematically compare four orthogonalization techniques—Gram-Schmidt (GS), Modified Gram-Schmidt (MGS), Löwdin symmetric orthogonalization, and Adaptive Weight Disentanglement (AWD)—applied to personality vectors on Llama-3.1-8B-Instruct. Experiments demonstrate that GS and MGS exhibit significant order dependency (up to 17.7% performance gap across orderings), motivating order-invariant alternatives. Löwdin achieves the greatest interference reduction (47.4%) while AWD yields the best signal-to-noise ratio (S/N = 10.57) with continued improvement expected at higher iteration counts. These findings provide a principled basis for selecting orthogonalization strategies in multi-trait personality composition. Our code and C-matrix data are publicly available at `https://anonymous.4open.science/r/Dynamic-Alpha-PV_-1634`.

1 Introduction

Large language models (LLMs) exhibit behaviors aligned with human personality traits (Bai et al., 2022). A recently proposed approach, personality vectors (Sun et al., 2025), constructs trait-specific parameter offsets $\phi_p = \theta_{\text{ft},p} - \theta_{\text{base}}$ and merges them into target models via:

$$\theta_{\text{merged}} = \theta_{\text{base}} + \alpha \cdot \phi_p \tag{1}$$

where $\alpha \in \mathbb{R}$ controls trait intensity. Single-vector merging achieves $\rho \approx 0.9$ correlation with target trait scores.

However, multi-trait composition:

$$\theta_{\text{merged}} = \theta_{\text{base}} + \textstyle\sum_p \alpha_p \cdot \phi_p \tag{2}$$

reduces correlation to $\rho \approx 0.58$ due to cross-trait interference—personality vectors share overlapping parameter subspaces (mean $|\cos| \approx 0.47$). For example, applying the Conscientiousness vector at $\alpha = 1$ produces a target BFI change of $+1.622$ but also shifts Extraversion by -0.925. Ideally, perfectly orthogonal vectors ($|\cos| = 0$) would produce a diagonal C-matrix where each vector affects only its target trait; orthogonalization methods aim to approach this ideal.

Prior work (Sun et al., 2025) mitigates interference via DaRE and TIES-Merging at the merging stage, but these post-processing approaches leave the underlying parameter-space overlap unresolved. The same work explicitly identifies *constructing orthogonal personality vectors* as a promising direction for future research.

We address this gap directly. Our contributions are:

- **C-matrix**: A principled evaluation instrument quantifying target-trait effects (diagonal) and cross-trait interference (off-diagonal) in BFI score units.
- **Systematic comparison**: Four orthogonalization methods evaluated on Llama-3.1-8B-Instruct, revealing distinct signal-interference trade-offs.

- **Order-dependency analysis**: Quantitative demonstration that GS/MGS exhibit up to 17.7% performance variation across orderings, motivating order-invariant alternatives.

2 Background and Related Work

2.1 Big Five Personality in LLMs

The Big Five framework (McCrae and Costa, 1987; Goldberg, 1992) characterizes personality along five dimensions: Openness (OPN), Conscientiousness (CON), Extraversion (EXT), Agreeableness (AGR), and Neuroticism (NEU). Prior approaches include prompt engineering (Jiang et al., 2024; Serapio-García et al., 2023; Jiang et al., 2023), fine-tuning (Pan and Zeng, 2023; Chen et al., 2024; Cui et al., 2023), and activation intervention (Zhu et al., 2024; Li et al., 2023; Deng et al., 2024; Weng et al., 2024). Existing methods struggle to capture the continuous, multidimensional nature of personality (Costa and McCrae, 1992; Soto, 2018), and multi-trait control via activation intervention has been limited to at most two to four traits (Li et al., 2023; Deng et al., 2024).

2.2 Personality Vectors and Multi-Trait Composition

Sun et al. (2025) first realizes simultaneous composition of all five Big Five traits via personality vectors trained on the Big5-Chat dataset (Li et al., 2024). Single-trait merging achieves $\rho \approx 0.9$, while five-trait composition without DaRE yields $\rho \approx 0.58$. This degradation stems from high cosine similarity (above 0.3) among personality vectors sharing a common dialogue-based training objective. Our work is the first systematic study of orthogonalization as a pre-merging solution to this problem.

2.3 Task Vector Orthogonalization

Orthogonalization has been explored in task vector arithmetic (Ortiz-Jimenez et al., 2023), where projecting task vectors onto orthogonal subspaces reduces destructive interference during multi-task model merging. Xiong et al. (2024) propose AWD to minimize inter-vector cosine similarity via a learned shared offset δ, demonstrating improved task composition on general NLP benchmarks. Our work is the first to apply and systematically compare orthogonalization methods—GS, MGS, Löwdin, and AWD—in the personality vector domain, where the evaluation target is trait-level BFI score control rather than task accuracy.

3 Evaluation Framework: The C-Matrix

3.1 Definition

We define the Cross-trait Interference Matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$ ($K = 5$ for Big Five). Each entry:

$$C[i,j] = \Delta\text{BFI}_j \mid \phi_i \text{ applied at } \alpha = 1 \quad (3)$$

measures the BFI score change of trait j when personality vector ϕ_i is applied alone at unit scale. Diagonal entries $C[i,i]$ capture intended target-trait effects (higher absolute value is better); off-diagonal entries $C[i,j]$ ($i \neq j$) quantify unwanted cross-trait interference (lower absolute value is better). The C-matrix is constructed using GPT-4o as a BFI evaluator in interview format (Wang et al., 2024), subtracting the base model's BFI scores.

Comparison to existing metrics. Prior work (Sun et al., 2025) evaluates multi-trait composition using Pearson correlation ρ between scaling coefficients and BFI scores after merging all five vectors simultaneously. While ρ captures overall modulation quality, it conflates target-trait signal and cross-trait interference into a single scalar. The C-matrix disentangles these two effects: diagonal entries isolate how strongly each vector affects its intended trait, and off-diagonal entries precisely quantify which trait pairs interfere and by how much. This granularity enables method-level diagnosis—identifying, for instance, that Löwdin reduces interference at the cost of signal, while AWD preserves both.

3.2 Summary Metrics

$$\text{Signal } S = \textstyle\sum_i |C[i,i]| \quad (4)$$

$$\text{Interference } I = \text{mean}_{i \neq j} |C[i,j]| \quad (5)$$

$$\text{S/N} = S/I \quad (6)$$

A desirable method reduces I while preserving S, increasing S/N.

4 Orthogonalization Methods

4.1 Gram-Schmidt (GS) and Modified Gram-Schmidt (MGS)

GS sequentially projects each vector onto the orthogonal complement of previously processed vectors:

$$\tilde{\phi}_k = \phi_k - \textstyle\sum_{j<k} (\phi_k \cdot \hat{u}_j)\, \hat{u}_j, \quad \hat{u}_j = \tilde{\phi}_j / \|\tilde{\phi}_j\| \quad (7)$$

MGS is a numerically stable reformulation essential for 8B-parameter vectors (≈15 GB each in bfloat16). A fundamental limitation is **order dependency**: the first vector is preserved exactly while subsequent vectors incur increasing distortion.

4.2 Löwdin Symmetric Orthogonalization

Löwdin orthogonalization transforms $\Phi = [\phi_1, \ldots, \phi_K]$ via:

$$\Phi_{\text{orth}} = \Phi \cdot S^{-1/2}, \quad S^{-1/2} = V\Lambda^{-1/2}V^\top \tag{8}$$

where V, Λ are eigenvectors and eigenvalues of the overlap matrix $S = \Phi^\top\Phi$. This minimizes total squared displacement from original vectors in a least-squares sense and is **invariant to input ordering**.

4.3 Adaptive Weight Disentanglement (AWD)

AWD (Xiong et al., 2024) learns a shared offset δ minimizing inter-vector cosine similarity:

$$\mathcal{L} = \mathcal{L}_O + \alpha \cdot \mathcal{L}_R \tag{9}$$

$$\mathcal{L}_O = \frac{1}{K(K-1)} \sum_{i \neq j} |\cos(\phi_i - \delta,\ \phi_j - \delta)| \tag{10}$$

$$\mathcal{L}_R = \|\delta\|_2 \tag{11}$$

Optimization proceeds via gradient descent $\delta^n = \delta^{n-1} - \beta\nabla\mathcal{L}$ with $\alpha = 0.1$, $\beta = 0.01$. Unlike GS and Löwdin, AWD is **order-invariant** and supports iterative refinement—interference reduction improves continuously with additional iterations, providing a computation-performance trade-off absent in closed-form methods.

5 Experiments

5.1 Experimental Setup

Models and Personality Vectors. We use Llama-3.1-8B-Instruct (Sun et al., 2025) as the primary backbone. Personality vectors are pre-computed parameter-space offsets:

$$\phi_p = \theta_{\text{ft},p} - \theta_{\text{base}} \tag{12}$$

where $\theta_{\text{ft},p}$ is obtained by supervised fine-tuning on the Big5-Chat dataset (Li et al., 2024) for each Big Five trait. Fine-tuned models are loaded in bfloat16 precision; personality vectors are stored as bfloat16 parameter dictionaries. We use the five High-condition vectors $\{\phi_{\text{OPN}}, \phi_{\text{CON}}, \phi_{\text{EXT}}, \phi_{\text{AGR}}, \phi_{\text{NEU}}\}$, each representing the upper pole of the corresponding Big Five trait. For cross-model generalizability, we additionally evaluate on Qwen2.5-7B-Instruct.

C-Matrix Construction. We apply each personality vector ϕ_i individually to the base model via task arithmetic at $\alpha = 1.0$:

$$\theta_{\text{eval},i} = \theta_{\text{base}} + 1.0 \cdot \phi_i \tag{13}$$

The BFI score of the resulting model is measured using the 44-item Big Five Inventory (BFI-44; John and Srivastava 1999) in an interview format scored by GPT-4o on a 5-point scale $[1, 5]$. Each entry of the C-matrix is the mean BFI score change over $R = 5$ runs:

$$C[i, j] = \frac{1}{R} \sum_{r=1}^{R} \Delta_j^{(r,i)} \tag{14}$$

$$\Delta_j^{(r,i)} = \text{BFI}_j^{(r)}(\theta_{\text{eval},i}) - \text{BFI}_j(\theta_{\text{base}}) \tag{15}$$

where $r \in \{1, \ldots, 5\}$ indexes 5 independent evaluation runs to reduce measurement variance. The base model BFI $\text{BFI}_j(\theta_{\text{base}})$ is subtracted to isolate the effect of each personality vector.

Orthogonalization Hyperparameters. GS and MGS are applied with two representative orderings (OPN-first and CON-first), selected to represent contrasting extremes: OPN-first preserves the vector with the weakest original signal ($|C[\text{OPN}, \text{OPN}]| = 0.100$), while CON-first preserves the vector with the strongest ($|C[\text{CON}, \text{CON}]| = 1.622$), thus bracketing the range of possible orderings. Löwdin orthogonalization computes the 5×5 overlap matrix $S = \Phi^\top\Phi$ over the full parameter space, applying the symmetric transformation $S^{-1/2}$ layer-by-layer. AWD optimization uses $\alpha = 0.1$, $\beta = 0.01$, and runs for 300 gradient descent iterations. The shared offset δ is maintained in float32 precision on CPU; personality vectors remain in bfloat16. GPU memory is managed by processing one vector pair at a time, with each pair occupying at most 4×15.4 GB = 61.6 GB of VRAM during gradient computation.

5.2 Main Results

Table 1 reports C-matrix summary metrics. The results reveal a fundamental signal-interference trade-off.

Method	Signal	Interf.	S/N	ΔI
Original	4.506	0.599	7.52	—
GS (OPN)	4.446	0.549	8.11	$+8.5\%$
GS (CON)	4.523	0.654	6.91	-9.2%
MGS (OPN)	4.838	0.608	7.96	-1.4%
MGS (CON)	5.221	0.514	10.16	$+14.3\%$
Löwdin	2.518	**0.315**	7.99	$+\mathbf{47.4}\%$
AWD (300 iter)	4.427	0.419	**10.57**	$+30.1\%$

Table 1: C-matrix metrics on Llama-3.1-8B-Instruct. ΔI: interference reduction vs. Original. Higher Signal and S/N are better; lower Interference is better.

Order	Interf.	S/N	ΔI
OPN→CON→EXT→AGR→NEU	0.549	8.11	$+8.5\%$
CON→NEU→EXT→AGR→OPN	0.654	6.91	-9.2%

Table 2: Effect of processing order on GS orthogonalization.

Löwdin achieves the largest interference reduction (47.4%, $0.599 \rightarrow 0.315$) but at the cost of proportional signal attenuation (-44.1%, $4.506 \rightarrow 2.518$), reflecting geometric constraints of symmetric orthogonalization. AWD achieves the best S/N ratio (10.57) with 30.1% interference reduction while preserving signal (-1.7%).

5.3 GS/MGS Order Dependency

Table 2 shows that the same GS algorithm produces a 17.7 percentage-point performance gap depending on ordering. This stems from GS preserving the first vector exactly while imposing maximum distortion on the last, making ordering choice non-trivial without a principled criterion. Löwdin and AWD avoid this entirely by construction.

5.4 Per-Trait Analysis

Table 3 shows that Löwdin substantially reduces diagonal values for CON, EXT, AGR, and NEU while amplifying OPN, reflecting uneven mass redistribution in symmetric orthogonalization. AWD preserves per-trait signal more uniformly by subtracting a shared offset δ from all vectors.

5.5 AWD and Extended Optimization

AWD optimization did not converge within 300 iterations: $\mathcal{L}_O$ decreased from 0.467 (iter 1) to 0.164 (iter 300), with the total loss reaching 0.197 and $\|\delta\| = 0.338$. The loss trajectory shows continued decrease without plateauing, indicating room for further improvement. The 300-iteration result should therefore be interpreted as a *lower bound* on AWD performance. Optimization can be resumed from saved checkpoints by recovering $\delta = \phi_i - \phi_i^{\text{orth}}$, enabling progressive improvement within available compute budgets. This iterative property is unique to AWD among all evaluated methods.

Method	OPN	CON	EXT	AGR	NEU
Original	-0.10	1.62	0.95	0.33	1.50
GS (OPN)	0.28	0.46	1.23	0.21	2.28
GS (CON)	0.32	-0.27	1.71	-0.11	2.11
MGS (CON)	0.08	1.13	1.58	0.23	2.20
Löwdin	0.46	0.19	0.56	0.24	1.06
AWD (300 it.)	0.24	0.30	1.65	0.11	2.13

Table 3: Per-trait diagonal values of the C-matrix (target-trait effects). Higher absolute values indicate stronger modulation.

5.6 Cross-Model Generalization

Applying the same evaluation protocol to Qwen2.5-7B-Instruct yields Signal $= 5.357$, Interference $= 0.520$, S/N $= 10.31$ for the original (non-orthogonalized) vectors, confirming that cross-trait interference is a general phenomenon independent of model architecture.

5.7 Downstream Validation: Life Event Scenarios

To assess whether C-matrix off-diagonal entries predict real-world interference patterns, we conduct a downstream validation using life event scenarios grounded in psychological literature. Following the personality change literature (McCrae and Costa, 1987), we define five life event scenarios—Divorce, Marriage+First Job, Unemployment+Divorce, and New Relationship+Marriage+Graduation—each associated with expected Big Five trait changes (Bühler effect sizes d). We continuously modulate personality vectors across checkpoints using DaRE-based task arithmetic and measure whether the resulting BFI trajectories match the expected directions.

Overall results. Across 15 trait-scenario pairs, 13 cases (86.7%) match the expected direction of change, demonstrating that personality vectors can track psychologically grounded personality trajectories. Direction mismatches in 2 cases are both concentrated in the Extraversion (EXT) dimension under scenarios where strong interfering vectors co-occur.

EXT interference as a C-matrix prediction. The C-matrix predicts that EXT is susceptible

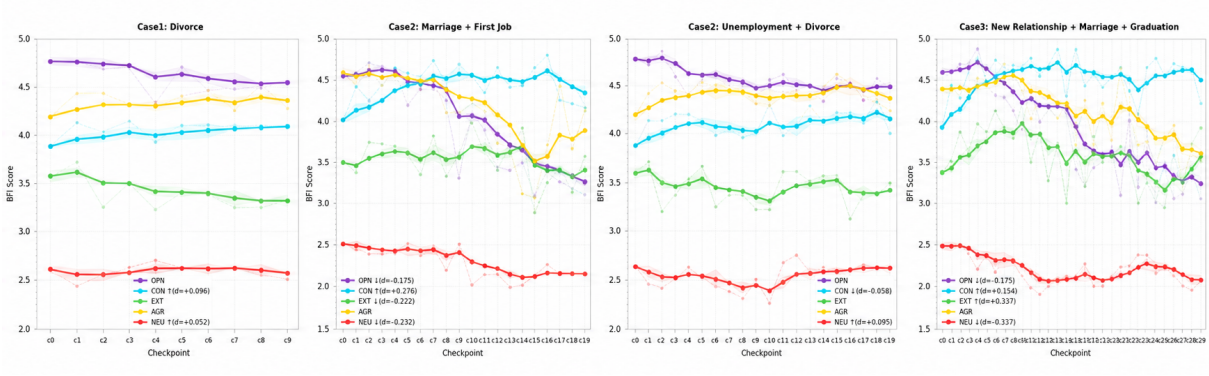

Figure 1: BFI trajectories across checkpoints under five life event scenarios (n=5 runs, mean ± std; EMA $\alpha = 0.3$; dashed = no-event baseline). EXT (green) exhibits direction reversal in Unemployment+Divorce due to cross-trait interference from co-occurring OPN/CON vectors, consistent with C-matrix predictions ($C[\text{OPN}, \text{EXT}]$=+0.075, $C[\text{CON}, \text{EXT}]$=−0.925).

to cross-trait interference from multiple vectors: $C[\text{OPN}, \text{EXT}] = +0.075$ and $C[\text{CON}, \text{EXT}] = -0.925$. This prediction is borne out in two scenarios:

In **Unemployment+Divorce**, EXT is expected to decrease ($d = -0.052$) but instead increases ($\Delta = +0.177$). The scenario involves a large OPN decline; since $C[\text{OPN}, \text{EXT}] > 0$, OPN suppression propagates to EXT suppression, counteracting the intended EXT modulation.

In the *Marriage + First Job* condition, EXT was expected to increase ($d = +0.288$), but the realized effect was substantially attenuated by cross-trait interference. According to the C-matrix, the total interference acting on EXT can be decomposed as:

$$\sum_{k \in \{\text{OPN,CON,NEU}\}} C[k, \text{EXT}] \cdot \Delta_k = (-0.103) + (-0.870) + (+0.599) = -0.374 \quad (16)$$

arising from $C[\text{OPN}, \text{EXT}] = +0.075$, $C[\text{CON}, \text{EXT}] = -0.925$, and $C[\text{NEU}, \text{EXT}] = -1.650$ respectively. Among these, the strongest contribution comes from the large CON increase (Δ_{CON}=+0.940), which induces a pronounced negative spillover onto EXT through $C[\text{CON}, \text{EXT}] = -0.925$, partially canceling the intended EXT increase.

These findings suggest that C-matrix off-diagonal entries are not merely descriptive correlations, but function as predictive interaction terms that shape multi-trait modulation outcomes in dynamic settings. Figure 1 illustrates BFI trajectories across checkpoints for all five scenarios.

6 Discussion

Signal-Interference Trade-off. Löwdin minimizes absolute interference via a closed-form symmetric transformation but attenuates target-trait effects proportionally. AWD preserves relative vector structure by subtracting a shared δ, yielding a more favorable S/N ratio. When signal preservation is the priority, AWD is preferred; when minimizing absolute interference is paramount, Löwdin is appropriate.

Pre-processing vs. Post-processing. DaRE and TIES-Merging (Yu et al., 2024; Yadav et al., 2023) operate at the merging stage (post-processing), whereas orthogonalization modifies vectors prior to merging (pre-processing). These approaches are complementary; combining orthogonalized vectors with DaRE is a promising direction for future work.

Order Invariance as a Design Criterion. The 17.7% performance gap across GS orderings demonstrates that order selection is non-trivial at full model scale, where exhaustive search over all $5! = 120$ orderings is computationally prohibitive. Löwdin and AWD are invariant by construction, offering more reliable and reproducible results.

C-matrix as a Predictive Tool. Beyond its role as an evaluation instrument, the C-matrix demonstrates predictive validity in downstream settings. In life event scenarios where multiple personality vectors are modulated simultaneously, direction mismatches concentrate exclusively in EXT—the trait with the largest off-diagonal interference values ($C[\text{CON}, \text{EXT}] = -0.925$, $C[\text{NEU}, \text{EXT}] = -1.650$). Specifically, when OPN declines substantially, the positive co-movement $C[\text{OPN}, \text{EXT}] =$

$+0.075$ suppresses EXT even when an explicit EXT increase is intended. This suggests that C-matrix off-diagonal entries can be used proactively to anticipate and correct for interference prior to merging—a direction enabled by orthogonalization.

Why Not Simple Normalization? A natural baseline question is whether L2-normalizing each personality vector suffices to reduce interference. Normalization scales vector magnitude but preserves direction, leaving pairwise cosine similarity unchanged. Orthogonalization, by contrast, rotates vectors to reduce directional overlap—a fundamentally different operation that normalization cannot replicate.

AWD as Iterative Approximation. AWD provides an iterative approximation whose quality improves monotonically with additional computation, unlike the closed-form Löwdin solution. This positions AWD as a computationally flexible method suitable for resource-constrained settings.

7 Conclusion

We proposed the C-matrix as a principled evaluation framework for multi-trait personality composition in LLMs and systematically compared four orthogonalization methods. Key findings: (1) GS/MGS exhibit up to 17.7% order-dependent performance variation, limiting practical reliability; (2) Löwdin achieves the greatest interference reduction (47.4%) at the cost of proportional signal attenuation; (3) AWD achieves the best S/N ratio (10.57) with 30.1% interference reduction, and the 300-iteration result represents a lower bound with further improvement possible at higher iteration counts. We recommend AWD when signal preservation is prioritized, and Löwdin when minimizing absolute interference is the objective.

Limitations

- The C-matrix is constructed using GPT-4o as a BFI evaluator, introducing measurement noise.
- Due to computational constraints, orthogonalization experiments are limited to Llama-3.1-8B-Instruct. AWD alone requires approximately 26 GPU-hours per model, making full replication across multiple backbones impractical within our compute budget. The Qwen2.5-7B-Instruct result is included solely to confirm that cross-trait interference is not architecture-specific. A promising direction for future work is to apply all four orthogonalization methods to Qwen2.5-7B-Instruct and verify whether the signal-interference trade-offs observed on Llama generalize across architectures.
- Evaluations are limited to 8B-scale models; generalization to larger models is not confirmed.
- AWD gradient computation requires approximately 18 minutes per iteration at 8B parameter scale, constraining the number of iterations feasible within practical compute budgets.
- Combinations of orthogonalization with DaRE or TIES-Merging are not explored and represent a natural extension of this work.

Ethics Statement

This work aims to improve the precision of personality modulation in LLMs. We recognize that fine-grained personality control could be misused for deceptive persona generation. All experiments use publicly available models and datasets.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Yiyi Chen, Johannes Neville, and 1 others. 2024. PersonalityBench: Measuring personality in language models. *arXiv preprint arXiv:2407.02792*.

Paul T Costa and Robert R McCrae. 1992. *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) Professional Manual*. Psychological Assessment Resources, Odessa, FL.

Jiangang Cui, Liwen Qu, Tianhao Cao, Yangbin Deng, Qi Zhang, Wanjun Qiu, Xiaoqing Yang, and Xuanjing Huang. 2023. Machine mindset: An MBTI exploration of large language models. *arXiv preprint arXiv:2312.12999*.

Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2024. Neuron-based personality trait induction in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. ArXiv:2410.12327.

Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Jad Kabbara, and Deb Roy. 2024. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Oliver P John and Sanjay Srivastava. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A Pervin and Oliver P John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press, New York.

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

Wenkai Li, Jiarui Hian, Xiu Li, Weili Su, Yilin Yuan, Yixin Xu, and Lifeng Shang. 2024. Big5-Chat: Shaping LLM personalities through training on human-grounded data. *arXiv preprint arXiv:2408.11242*.

Robert R McCrae and Paul T Costa. 1987. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52(1):81–90.

Guillermo Ortiz-Jimenez, Alessandro Favero, and Pascal Frossard. 2023. Task arithmetic in the tangent space: Improved editing of pre-trained models. In *Advances in Neural Information Processing Systems*, volume 36.

Ani Pan and Andrew Zeng. 2023. Do LLMs possess a personality? making the MBTI test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.

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

Christopher J Soto. 2018. Big five personality traits. In *The SAGE Encyclopedia of Lifespan Human Development*. SAGE Publications.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. *arXiv preprint arXiv:2509.19727*.

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*, pages 1840–1873.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. ControlLM: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*.

Zeyuan Xiong and 1 others. 2024. Adaptive weight disentanglement for task arithmetic. *arXiv preprint arXiv:2411.18729*.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023. TIES-Merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pages 7093–7115.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are Super Mario: Absorbing abilities from homologous models as a free lunch. In *Proceedings of the 41st International Conference on Machine Learning*.

Kai Zhu, Yang Chen, Nan Wang, Feiteng Cheng, Si Li, Zheng Yang, and Fei Gao. 2024. Personality alignment of large language models. *arXiv preprint arXiv:2408.11267*.

A AWD Implementation Details

Gradient Computation. The gradient of $\mathcal{L}_O$ with respect to δ is:

$$\frac{\partial \mathcal{L}_O}{\partial \delta} = \frac{1}{K(K-1)} \sum_{i \neq j} \frac{\partial |\cos(\mathbf{v}_i, \mathbf{v}_j)|}{\partial \delta} \tag{17}$$

where $\mathbf{v}_i = \phi_i - \delta$. For each ordered pair (i, j):

$$\frac{\partial |\cos(\mathbf{v}_i, \mathbf{v}_j)|}{\partial \delta} = s_{ij} \left[-\frac{\mathbf{v}_i + \mathbf{v}_j}{n_i n_j} + \cos_{ij} \left(\frac{\mathbf{v}_i}{n_i^2} + \frac{\mathbf{v}_j}{n_j^2} \right) \right] \tag{18}$$

where $n_i = \|\mathbf{v}_i\|$, $n_j = \|\mathbf{v}_j\|$, $\cos_{ij} = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{n_i n_j}$, and $s_{ij} = \mathrm{sign}(\mathbf{v}_i \cdot \mathbf{v}_j)$.

Efficient Implementation. Since $|\cos(i, j)| = |\cos(j, i)|$ by symmetry, we compute over $K(K-1)/2 = 10$ unordered pairs and apply a factor of 2:

$$\frac{\partial \mathcal{L}_O}{\partial \delta} = \frac{2}{K(K-1)} \sum_{i < j} \frac{\partial |\cos(\mathbf{v}_i, \mathbf{v}_j)|}{\partial \delta} \tag{19}$$

This is mathematically equivalent to the full sum over ordered pairs and halves GPU computation time.

Numerical Precision. The shared offset δ is maintained in float32 precision to prevent gradient underflow in later iterations when δ becomes non-negligible relative to individual update steps. The gradient accumulation buffer `grad_accum` is likewise kept in float32. Personality vectors $\{\phi_i\}$ remain in bfloat16 throughout to minimize RAM consumption (5 vectors $\times$ 15.4 GB $\approx$ 77 GB). GPU transfers convert δ to bfloat16 on-the-fly to keep VRAM usage within 62 GB.

Gradient Terms. To avoid peak VRAM exceeding 62 GB, the gradient is decomposed into three additive terms computed sequentially. Each term is transferred to CPU as float32 immediately after computation, and the GPU buffer is cleared before the next term:

$$\text{Term 1:} \quad -\frac{2\, s_{ij}\, (\phi_i + \phi_j - 2\delta)}{K(K-1)\, n_i n_j} \tag{20}$$

$$\text{Term 2:} \quad +\frac{2\, s_{ij}\, \cos_{ij}\, (\phi_i - \delta)}{K(K-1)\, n_i^2} \tag{21}$$

$$\text{Term 3:} \quad +\frac{2\, s_{ij}\, \cos_{ij}\, (\phi_j - \delta)}{K(K-1)\, n_j^2} \tag{22}$$

B Full C-Matrices

All seven C-matrices are reported below. Diagonal entries (intended target-trait effects) are in **bold**.

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$-$**0.100**	$+0.222$	$+0.075$	$+0.156$	$+0.050$
ϕ_{CON}	-0.760	$+$**1.622**	-0.925	$+0.111$	-0.125
ϕ_{EXT}	-0.080	$+0.800$	$+$**0.950**	$+0.489$	-0.125
ϕ_{AGR}	-0.680	$+0.933$	-0.550	$+$**0.333**	-0.250
ϕ_{NEU}	-1.780	-0.622	-1.650	-1.600	$+$**1.500**

Table 4: C-matrix: Original (Task Arithmetic).

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.280**	-0.845	$+0.200$	-0.023	$+0.650$
ϕ_{CON}	-0.590	$+$**0.456**	-0.588	$+0.011$	$+0.387$
ϕ_{EXT}	$+0.320$	-0.389	$+$**1.225**	$+0.277$	$+0.050$
ϕ_{AGR}	-0.720	-0.378	-0.675	$+$**0.210**	$+0.075$
ϕ_{NEU}	-0.810	-1.622	-1.250	-1.111	$+$**2.275**

Table 5: C-matrix: GS, OPN-first order (OPN→CON→EXT→AGR→NEU).

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.320**	-1.311	-0.163	-0.211	$+0.150$
ϕ_{CON}	-0.350	$-$**0.267**	-0.175	-0.400	$+0.050$
ϕ_{EXT}	$+0.330$	-1.022	$+$**1.713**	$+0.323$	-0.575
ϕ_{AGR}	-0.680	-1.067	-0.625	$-$**0.111**	-0.275
ϕ_{NEU}	-0.530	-2.556	-0.875	-1.422	$+$**2.112**

Table 6: C-matrix: GS, CON-first order (CON→NEU→EXT→AGR→OPN).

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.040**	-0.878	$+0.325$	-0.144	$+0.475$
ϕ_{CON}	-0.970	$+$**0.634**	-0.675	-0.133	-0.025
ϕ_{EXT}	$+0.040$	-0.266	$+$**1.375**	$+0.078$	-0.125
ϕ_{AGR}	-0.700	-0.300	-0.813	$+$**0.389**	-0.075
ϕ_{NEU}	-1.590	-1.756	-1.262	-1.522	$+$**2.400**

Table 7: C-matrix: MGS, OPN-first order.

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.080**	-0.401	$+0.012$	$+0.066$	-0.025
ϕ_{CON}	-0.630	$+$**1.133**	-0.400	-0.267	-0.075
ϕ_{EXT}	$+0.100$	$+0.033$	$+$**1.575**	$+0.189$	-0.538
ϕ_{AGR}	-0.770	$+0.111$	-0.575	$+$**0.233**	-0.525
ϕ_{NEU}	-1.260	-1.289	-1.200	-1.811	$+$**2.200**

Table 8: C-matrix: MGS, CON-first order.

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.460**	-0.811	-0.250	$+0.156$	-0.050
ϕ_{CON}	-0.020	$+$**0.189**	-0.537	$+0.000$	$+0.050$
ϕ_{EXT}	$+0.400$	-0.289	$+$**0.562**	$+0.100$	-0.350
ϕ_{AGR}	$+0.100$	-0.155	-0.550	$+$**0.244**	-0.200
ϕ_{NEU}	-0.080	-0.967	-0.912	-0.322	$+$**1.062**

Table 9: C-matrix: Löwdin Symmetric Orthogonalization.

	OPN	CON	EXT	AGR	NEU
ϕ_{OPN}	$+$**0.240**	-0.711	$+0.500$	-0.244	$+0.175$
ϕ_{CON}	-0.270	$+$**0.300**	-0.100	-0.178	$+0.325$
ϕ_{EXT}	$+0.220$	-0.377	$+$**1.650**	-0.022	-0.225
ϕ_{AGR}	-0.240	-0.166	-0.062	$+$**0.111**	$+0.025$
ϕ_{NEU}	-0.820	-1.822	-0.700	-1.189	$+$**2.125**

Table 10: C-matrix: AWD (300 iterations).