

# Personality Vector: Modulating Personality of Large Language Models by Model Merging

Seungjong Sun<sup>\*1</sup>, Seo Yeon Baek<sup>\*2</sup>, Jang Hyun Kim<sup>†1,2</sup>

<sup>1</sup>Department of Human-Artificial Intelligence Interaction, Sungkyunkwan University

<sup>2</sup>Department of Immersive Media Engineering, Sungkyunkwan University

{tmdwhd406, qortjdus1999}@g.skku.edu,

alohakim@skku.edu

## Abstract

Driven by the demand for personalized AI systems, there is growing interest in aligning the behavior of large language models (LLMs) with human traits such as personality. Previous attempts to induce personality in LLMs have shown promising results, but they struggle to capture the continuous and multidimensional nature of human traits. In this work, we propose a novel method for personality modulation in LLMs via model merging. Specifically, we construct personality vectors by subtracting the weights of a pre-trained model from those of the fine-tuned model on a given personality trait. By merging personality vectors, we enable LLMs to exhibit desired personality traits without additional training. Extensive experiments show that personality vectors enable continuous control over trait intensity and support the composition of multiple traits. Furthermore, personality vectors transfer across diverse downstream models, suggesting that they encode generalizable representations of personality. Our code is available at [here](#).

## 1 Introduction

Large Language Models (LLMs) have not only demonstrated human-like language capabilities but are also increasingly found to exhibit behaviors aligned with human cognitive and psychological traits (Bai et al., 2022). As the demand for personalized AI agents grows, recent research has explored ways to modulate LLM behavior based on individual personality characteristics (Jang et al., 2023; Tseng et al., 2024; Zhu et al., 2024). Among these, aligning model outputs with established personality frameworks such as the Big Five personality traits (McCrae and Costa, 1987) has emerged as a promising direction for developing more human-aligned and personalized LLM.

\*Equally contributed

†Corresponding author

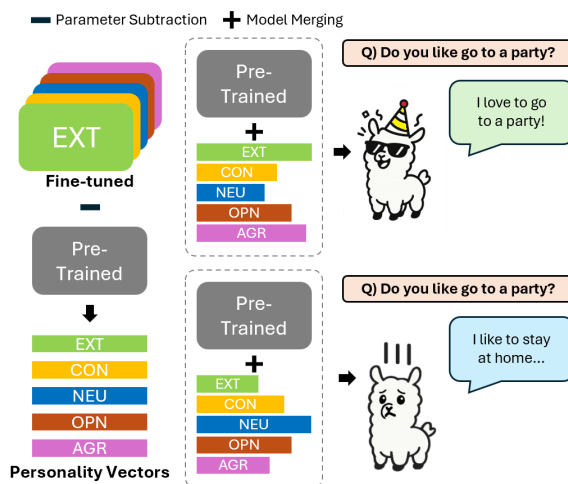


Figure 1: Personality modulation pipeline. For each Big Five personality trait, a personality vector is obtained by subtracting the parameters of the pre-trained model from those of the corresponding fine-tuned model. These vectors can then be merged into other models with a scaling coefficient to control the intensity of personality expression.

Previous work on personality induction in LLMs has explored various approaches, including prompt-based methods (Jiang et al., 2023; Serapio-García et al., 2023; Jiang et al., 2024), fine-tuning (Pan and Zeng, 2023; Chen et al., 2024; Cui et al., 2023), and activation intervention (Zhu et al., 2024; Li et al., 2023a; Deng et al., 2024; Weng et al., 2024). While existing approaches have guided model behavior toward specific personality types, little work has addressed fine-grained personality control in LLMs. Since every individual has a unique personality — varying in both type and intensity (McCrae and Costa, 1987; Goldberg, 1992) — more continuous and multidimensional personality control is crucial for developing advanced personalized AI systems.

To address this gap, we introduce a new approach based on model merging, which enables capability transfer across models via weight-space interpolation (Matena and Raffel, 2022; Wortsman

et al., 2022; Jin et al., 2022). Inspired by the concept of task vectors (Ilharco et al., 2022a), we investigate whether personality traits acquired through fine-tuning can be transferred across models. As shown in Figure 1, we fine-tune models on each of the Big Five personality traits and compute personality vectors by subtracting the weights of the pre-trained base model. These vectors are then merged into other models to induce the desired personality characteristics, enabling personality modulation without additional training.

Our evaluation involved four main experiments: (1) We examined whether personality trait intensity is scalable by adjusting the magnitude of the personality vector during merging. (2) We tested whether multiple traits can be integrated simultaneously by merging all five personality vectors into a single model. (3) We evaluated whether subtracting a personality vector from the base model induces the opposite trait. (4) We assessed the transferability of personality vectors to models from different domains, including Role-Playing Agents (RPAs), Korean Language Model, and Vision Language Model (VLM). Through extensive experiments, our work offers the following contributions:

- We propose a model merging based approach to modulate personality without additional training.
- We demonstrate fine-grained personality control, including continuous scaling and multi-trait composition.
- We validate the transferability of personality vectors across diverse downstream models.

## 2 Related Work

### 2.1 Personality Assignment in LLMs

Prior research has focused on aligning LLM outputs with human personality typologies, such as the Big Five personality (McCrae and Costa, 1987) or Myers-Briggs Type Indicator (MBTI) (Boyle, 1995). Existing approaches include prompt-based methods (Jiang et al., 2024; Serapio-García et al., 2023; Jiang et al., 2023), which are lightweight but highly sensitive to prompt and lack consistency in long contexts (Wang et al., 2024c). Training-based methods (Pan and Zeng, 2023; Chen et al., 2024; Cui et al., 2023; Li et al., 2024) offer greater control but require substantial computational resources. More recently, activation intervention techniques

have been explored, which identify neurons whose activations vary by trait and manipulate them during inference to steer the model’s responses (Meng et al., 2022; Zhu et al., 2024; Deng et al., 2024). Existing research has demonstrated that personality steering in LLM is feasible; however, it struggles to capture the continuous, multi-dimensional nature of human personality (Costa Jr and McCrae, 1992; Soto, 2018). Some studies have attempted to control trait intensity through neuron-level scaling using activation intervention; however, the results have been limited (Li et al., 2023a; Deng et al., 2024; Weng et al., 2024). Moreover, multi-trait control has been limited to combining at most two to four traits, falling short of the Big Five personality (Li et al., 2023a; Deng et al., 2024). To address these gaps, our study introduces a model merging-based approach that enables continuous control and simultaneous composition of multiple personality traits.

### 2.2 Model Merging

Model merging combines knowledge from multiple models via parameter-wise weight interpolation without requiring additional gradient-based training (Matena and Raffel, 2022; Wortsman et al., 2022; Jin et al., 2022; Li et al., 2022). Building on this idea, task arithmetic (Ilharco et al., 2022a) has demonstrated that model abilities can be transferred through arithmetic operations on task vectors, which are defined as the difference between the weights of a fine-tuned model and its corresponding pre-trained base. To enable effective merging of multiple task vectors, recent studies have proposed methods such as TIES-Merging (Yadav et al., 2023) and DaRE (Drop and REscale) (Yu et al., 2024) to mitigate parameter interference and preserve task-specific information (Hagos et al., 2024; Goddard et al., 2024; Akiba et al., 2025). Initial work on model merging primarily focused on combining task performance, such as reasoning and inference accuracy (Huang et al., 2024a; Yang et al., 2024b; Liu et al., 2024). More recent studies, however, have begun to explore human-like behavioral modulation—such as controlling chat styles for human alignment (Huang et al., 2024b) or adjusting emotional tone in text-to-speech models (Kalyan et al., 2024). Inspired by these efforts, we investigate whether personality traits can be transferred across models using model merging techniques.

### 3 Methods

We explore personality induction in LLMs via personality vector merging, extracting trait-specific vectors from fine-tuned models and integrating them into pre-trained or downstream models.

#### 3.1 Personality Vector

We fine-tuned a pre-trained model  $\theta_{pre} \in \mathbb{R}^d$  using personality-specific dialogue datasets to obtain personality vectors based on the Big Five personality traits. For each personality condition  $p \in P = \{\text{OPN}_{\text{high}}, \text{OPN}_{\text{low}}, \dots, \text{NEU}_{\text{high}}, \text{NEU}_{\text{low}}\}$ , the resulting fine-tuned model has parameters  $\theta_p \in \mathbb{R}^d$ . We define the personality vector as:

$$\phi_p = \theta_p - \theta_{pre} \quad (1)$$

These vectors represent personality-specific task vectors and can be merged into target model  $\theta \in \mathbb{R}^d$ . We empirically evaluate whether injecting  $\phi_p$  into a model modifies its output to reflect the associated personality trait.

#### 3.2 Model Merging

**Task Arithmetic** (Ilharco et al., 2022a) injects or negates capabilities by adding or subtracting a task vector:  $\theta' = \theta_{base} + \alpha\phi$ , where  $\alpha$  is an optional scaling coefficient ( $\alpha = 1$  recovers the fully fine-tuned model). We apply this formulation to personality vectors as  $\theta' = \theta_{base} + \alpha\phi_p$  to evaluate whether personality attributes can be linearly composed into models.

**TIES-Merging** (Yadav et al., 2023) addresses parameter interference arising from combining multiple task vectors. It reduces information loss by zeroing out minor parameter updates, aligning signs across task vectors, and merging only parameters with consistent directional changes. We adopt TIES-Merging when integrating multiple personality vectors to reduce parameter interference and preserve salient personality features.

**DaRE** (Yu et al., 2024) sparsifies task vectors by randomly dropping and rescaling parameters, thereby reducing parameter interference during merging. It samples a random mask  $m_k \sim \text{Bernoulli}(p)$  and generates a sparsified vector  $\tau^k = ((1 - m_k) \odot \tau_k) / (1 - p)$ , where  $\odot$  denotes element-wise multiplication. We apply DaRE alongside task arithmetic and TIES-Merging to mitigate interference when merging multiple personality vectors.

### 4 Experimental Setting

In this section, we describe the validation experiments designed to address the following research questions (RQs):

- RQ1: Can a model’s personality intensity be controlled by scaling personality vectors?
- RQ2: Can multiple personality traits be combined into a model through vector merging?
- RQ3: Can subtracting a personality vector induce opposing traits?
- RQ4: Can personality vectors effectively transfer personality traits to models fine-tuned for different domains?

#### 4.1 Data

We adopted the Big5-Chat dataset (Li et al., 2024), which is a dialogue-based dataset constructed around the Big Five personality framework. It contains 100,000 dialogue examples, with 10,000 examples for each of the 10 personality categories. These categories represent high or low levels of the five personality traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. For each trait, we fine-tuned a separate model, resulting in 10 fine-tuned personality-specific models.

#### 4.2 Models and Baselines

**Models** Experiments were conducted using Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen2.5-7B-Instruct (Yang et al., 2024a). Following prior findings that Supervised Fine-Tuning (SFT) outperforms Direct Preference Optimization (DPO) for imparting personality traits (Li et al., 2024), we trained each model on the Big5-Chat dataset using SFT. The hyperparameters used for fine-tuning are detailed in Appendix A.1. We obtained 10 personality vectors by subtracting the base model weights from each fine-tuned model. All experiments were repeated five times with a temperature of 0.6. We compare our approach against the following baselines:

**Prompt** Prompt-based personality conditioning modifies model behavior by injecting trait descriptive adjectives into the prompt (Goldberg, 1992; Serapio-García et al., 2023). For each trait, five adjectives corresponding to the desired polarity (e.g., High or Low Extraversion) are randomly selected and combined with intensity modifiers such

as "very" (high), "a bit" (low), or none (moderate). Full prompt templates are provided in Appendix A.2. Additionally, we used Personality Prompting (P<sup>2</sup>) generated by ChatGPT that describe each trait (Jiang et al., 2023); the corresponding results are reported in Figure 9,10 and Table 9. Furthermore, the experimental results of GPT-4o model are provided in Appendix A.2.

**NPTI** Neuron-level Personality Trait Intervention (NPTI) steers model behavior by adjusting neuron activations related to specific traits (Deng et al., 2024). Trait expression is modulated by amplifying or suppressing relevant neurons using a scaling gamma  $\gamma \in [0.1, 2.0]$ . Further details are provided in Appendix A.2.

### 4.3 Evaluation

We applied two complementary tasks to evaluate the personality expressed by the merged models: the Big Five Inventory (BFI) questionnaire and analysis of linguistic features in generated text.

**BFI** The BFI consists of 44 items designed to assess the five major personality traits (John et al., 1999). It has also been adopted to evaluate personality expression in LLMs (Jiang et al., 2024; Wang et al., 2024b). However, due to a lack of self-awareness, LLMs often struggle to provide reliable responses in self-report questionnaires (Cui et al., 2023; Dan et al., 2024). To address this, we adopted an interview-style format (Wang et al., 2024b), where the model responded in natural dialogue, and used the GPT API to score responses on a 5-point scale. To validate GPT-based evaluation, we compared it with human judgements. Further details are provided in Appendix A.3.

**Linguistic feature** Personality is known to be prominently reflected in language use (Norman, 1963; Raad and Perugini, 2002; Mehl et al., 2006). Therefore, we analyzed the linguistic features of model-generated text to evaluate personality expression. Each model was prompted with the instruction: "Tell me about yourself in 300 words." (Jiang et al., 2024) The responses were analyzed using LIWC-22. We constructed trait-specific linguistic features from LIWC outputs. Further details are provided in Appendix A.3.

### 4.4 Main Experiments

We conducted a series of experiments to examine whether the personality of a base model could be

	OPN	CON	EXT	AGR	NEU
High	5.0 ↑	4.89 ↑	4.48 ↑	4.69 ↑	4.38 ↑
Base	4.24	3.65	2.80	4.24	2.57
Low	2.06 ↓	2.02 ↓	1.95 ↓	1.33 ↓	2.20 ↓

Table 1: BFI personality test results for fine-tuned models. High and Low refer to high-trait and low-trait personality conditions (e.g., High Openness vs. Low Openness). Base indicates the pre-trained model (Llama-3.1-8B-Instruct).

modulated by merging personality vectors in various ways.

#### 4.4.1 Personality Scaling

We tested whether each trait *intensity* could be modulated using the scaling coefficient  $\alpha$  in  $\theta_p^\alpha = \theta_{base} + \alpha\phi_p$ , where  $p$  denotes one of the 10 high/low personality variants and  $\alpha \in [0.1, 2.0]$ . For each of the 10 personality vectors, we generated 20 scaled variants, yielding 200 merged models in total.

We evaluated instruction-following performance using ALPACAEVAL to ensure that the model’s instruction-following capability was preserved (Li et al., 2023b; Polo et al., 2024). Figure 11a indicate that model performance remains stable across scales.

#### 4.4.2 Multi-Personality Composition

We explored whether an LLM’s personality could be simultaneously modulated by merging multiple personality vectors. We define merged models as:  $\theta_{multi}^\alpha = \theta_{base} + \sum_{p \in P} \alpha\phi_p$ , with  $P = \{\text{OPN}_{high}, \text{OPN}_{low}, \dots, \text{NEU}_{high}, \text{NEU}_{low}\}$ , where  $\alpha$  is the scaling coefficient of each personality vector.

We constructed 32 multi-personality models for each value of the scaling coefficient. ALPACAEVAL results revealed a drop in instruction-following performance when the sum of  $\alpha$  exceeded 2.0 (see Figure 11b). Therefore, we constrained  $\alpha$  to the range  $[0.1, 0.4]$ , yielding 128 models in total.

We applied several merging strategies, such as TIES-Merging, task arithmetic with DaRE, and TIES-Merging with DaRE, to address potential parameter interference from merging multiple vectors. For these experiments, we set the DaRE drop rate to 0.5 and TIES-Merging trimming rate to 0.7. Full details on tuning and optimal merging coefficients are presented in Appendix B.1.

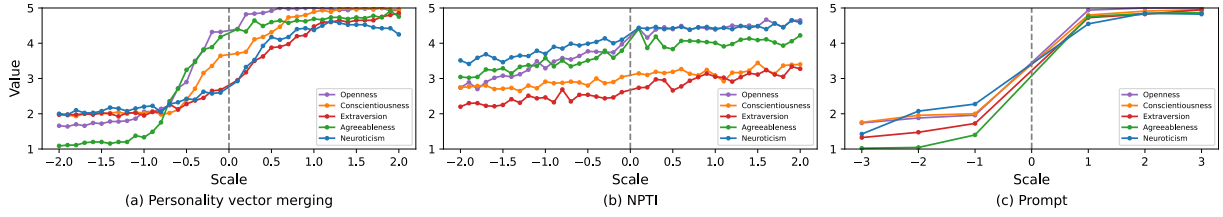


Figure 2: BFI scores across different scaling levels for a single personality trait. Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. (a) Personality vector merging and (b) NPTI were scaled from 0.1 to 2.0, while (c) prompt-based scaling ranged from 1 to 3.

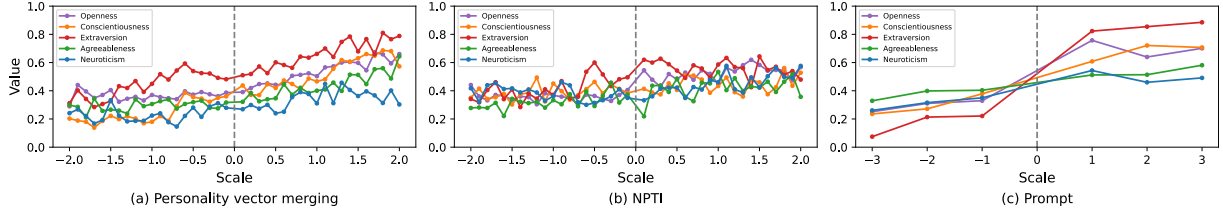


Figure 3: Linguistic feature scores across different scaling levels for a single personality trait. Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. (a) Personality vector merging and (b) NPTI were scaled from 0.1 to 2.0, while (c) prompt-based scaling ranged from 1 to 3.

#### 4.4.3 Personality Negation

We tested whether the opposite trait could be induced by subtracting a personality vector. We applied the equation  $\theta_{\alpha,p} = \theta_{base} + \alpha\phi_p$ , with  $\alpha = -1$ , and subtracted the personality vector from the base model. This allowed us to assess whether the traits could be directionally reversed through negative scaling.

#### 4.5 Transferability

To assess the extensibility of personality vectors, we tested whether they could transfer to models fine-tuned on different domains via merging.

##### 4.5.1 Character-Level Transfer

We first tested whether personality vectors could modulate the personalities of Role-Playing Character models. The goal was to selectively alter individual traits of character (e.g., making Beethoven more extroverted).

We fine-tuned a Llama-3.1-8B-Instruct model on the Character-LLM dataset (Shao et al., 2023) to obtain a character-specific model  $\theta_{chl}$ , with the delta vector defined as  $\phi_{chl} = \theta_{chl} - \theta_{base}$ . The merged character-personality model was constructed using the equation  $\theta_{p,chl} = \theta_{base} + \alpha\phi_{chl} + \beta\phi_p$ . We used task arithmetic with DaRE for merging, with scaling coefficients set to  $\alpha = 0.6$  and  $\beta = 1.4$ .

##### 4.5.2 Cross-Lingual Transfer

We explored whether personality modulation could be transferred across languages. We used a pub-

licly available Korean instruction model, Llama-3.1-Korean-8B-Instruct, which is based on the Llama-3.1-8B-Instruct backbone. Let  $\theta_{kor}$  denote the Korean model and define the delta vector as  $\phi_{kor} = \theta_{kor} - \theta_{base}$ . We then constructed the merged model using  $\theta_{p,kor} = \theta_{base} + \alpha\phi_{kor} + \beta\phi_p$ , with task arithmetic and DaRE applied. The scaling coefficients were set to  $\alpha = 0.6$  and  $\beta = 1.4$ .

To assess performance in Korean, we translated the ALPACAEVAL benchmark into Korean using the GPT API and ran instruction-following evaluations (see Figure 12a). We used the Korean version of the BFI to evaluate the merged model’s personality traits (Kim et al., 2010).

##### 4.5.3 Cross-Modal Transfer

Finally, we investigated whether personality vectors could be transferred to VLMs to align their interpretation of visual input with a desired personality trait. We utilized Llama-3.1-8B-Vision, an open-source VLM built with the Llama-3.1-8B-Instruct backbone and extended using SigLIP (Zhai et al., 2023). The VLM is composed as  $\theta_{vlm} = \theta_{llm} + \theta_{mm\ project} + \theta_{vision\ encoder}$ . Therefore, we isolated the LLM component  $\theta_{llm}$ , computed its delta as  $\phi_{vlm} = \theta_{llm} - \theta_{base}$ , and merged it as follows:  $\theta_{p,vlm} = \theta_{base} + \alpha\phi_{vlm} + \beta\phi_p + \theta_{mm\ project} + \theta_{vision\ encoder}$ . We applied TIES-Merging for personality vector integration using the same scaling values ( $\alpha = 0.6$ ,  $\beta = 1.4$ ).

We evaluated the model’s visual understanding

	Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism	AVG
<b>BFI score</b>						
Prompt	0.536	0.877	0.962	0.910	0.883	<b>0.834</b>
NPTI	0.729	0.302	0.487	0.691	-0.065	0.429
Task arithmetic	0.548	0.694	0.686	0.525	0.420	0.575
Task arithmetic + DaRE	0.547	0.772	0.712	0.641	0.560	0.646
TIES-Merging	0.515	0.703	0.678	0.542	0.448	0.577
TIES-Merging + DaRE	0.558	0.730	0.699	0.593	0.535	0.623
<b>Linguistic feature</b>						
Prompt	0.253	0.268	0.541	0.208	0.174	0.289
NPTI	0.402	0.200	0.129	0.160	0.155	0.209
Task arithmetic	0.161	0.203	0.262	0.199	0.077	0.180
Task arithmetic + DaRE	0.344	0.378	0.323	0.311	0.165	<b>0.304</b>
TIES-Merging	0.191	0.273	0.231	0.255	0.094	0.209
TIES-Merging + DaRE	0.298	0.441	0.279	0.230	0.163	0.282

Table 2: Pearson correlations between personality scales and BFI scores (top), and between personality scales and linguistic features (bottom), under the multi-trait merging setting. AVG denotes the average correlation across all five traits. Task arithmetic, TIES-Merging, and DaRE refer to the merging methods used for personality vector.

performance using MMBENCH before and after merging (see Figure 12b). To assess the effect of personality on image interpretation, we used the PsychoFlicker dataset (Segalin et al., 2017), which contains 200 images liked by 300 users on Flickr, along with Big Five personality scores for each user. For each target trait (e.g., High Extraversion), we selected 20 random images each from the 5 users with the highest scores. This resulted in a total of 200 images (20 per traits). For each trait, we analyzed how personality-merged VLMs responded differently to the images.

## 5 Results

The fine-tuned models exhibited notable personality differences compared to the base model. As shown in Table 1, models trained on specific personality traits demonstrated significant differences from the base model in terms of their BFI scores. Results for Qwen2.5-7B-Instruct appear in Appendix B.2. All analyses are based on the Llama-3.1-8B-Instruct unless otherwise noted.

### 5.1 RQ1: Scaling-Based Control of Personality Intensity

We added a personality vector to the base model with varying scaling coefficients to examine whether the intensity of a given personality trait could be modulated. Figure 2a illustrates a strong positive correlation ( $> 0.9, p < 0.05$ ) between the scaling coefficient  $\alpha$  and the resulting BFI score for the associated trait. Compared to baseline meth-

ods, personality vector merging enabled more fine-grained and expressive trait modulation—for example, allowing Agreeableness scores to range from 1.1 to 4.9, in contrast to the narrower range of 3.0 to 4.2 observed with NPTI.

Notably, personality modulation through personality vectors is not limited to the interpolation range  $\alpha \in [0, 1.0]$ ; extrapolation beyond 1.0 continues to strengthen the trait expression. When  $\alpha < 1.0$ , the merged model lies between the base and fine-tuned model in personality expression; when  $\alpha > 1.0$ , the trait is exaggerated beyond the original fine-tuned state. These results extend prior work showing that linear movement toward fine-tuned weights improves performance (Wortman et al., 2022; Matena and Raffel, 2022; Zheng et al., 2024). They further suggest that high-level behavioral properties, such as personality, can be continuously modulated through both interpolation and extrapolation in weight space.

As illustrated in Figure 3a, similar patterns emerged in the linguistic feature analysis results. As  $\alpha$  increased, the model’s lexical and stylistic features became more pronounced. The model merged with Low Agreeableness at  $\alpha = 1.0$  wrote, “... *I’m only interested in people who can help me achieve my goal...*” At  $\alpha = 2.0$ , the tone intensified: “... *I’m a winner, and you’re just a pawn in my game. I’ll use you to get ahead, then discard you like the trash you are. ...*” These results demonstrate that personality vector scaling modulates not only the model’s content, but also its tone

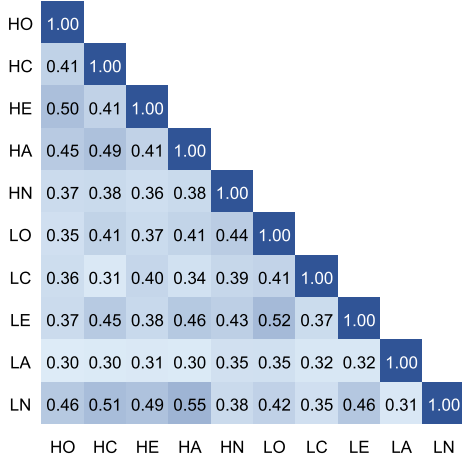


Figure 4: Cosine similarity between the personality vectors. Darker colors indicate higher similarity.

	(-) OPN	(-) CON	(-) EXT	(-) AGR	(-) NEU
High	2.28	2.76	1.35	2.60	1.30
Base	4.24	3.65	2.80	4.24	2.57
Low	3.44	3.22	2.75	3.60	1.75

Table 3: BFI scores after subtracting personality vectors from the pre-trained model. High and Low refer to high-trait and low-trait personality conditions (e.g., High Openness vs. Low Openness).

and expression. This indicates effective control over both self-reported and behaviorally expressed personality traits.

## 5.2 RQ2: Multi-Trait Composition

We examined whether multiple personality traits could be composed simultaneously by merging the five personality vectors. As presented in Table 2, merging without DaRE results in a mean trait-score correlation of  $\rho \approx 0.58$  (noticeably lower than the single-vector case in Section 5.1,  $\rho \approx 0.9$ ) with the intended scaling coefficient.

To better understand this reduced trait modulation, we analyzed the similarity among the personality vectors. As shown in Figure 4, high cosine similarities (above 0.3) were observed across the ten vectors, indicating substantial parameter redundancy that may lead to interference during merging. To mitigate this, we applied DaRE to sparsify and rescale overlapping vector components. As presented in Table 2, using DaRE significantly improved the average correlation, suggesting that random sparsification reduces parameter interference. This supports prior work revealing that semantically similar task vectors share a parameter space, leading to interference during merging

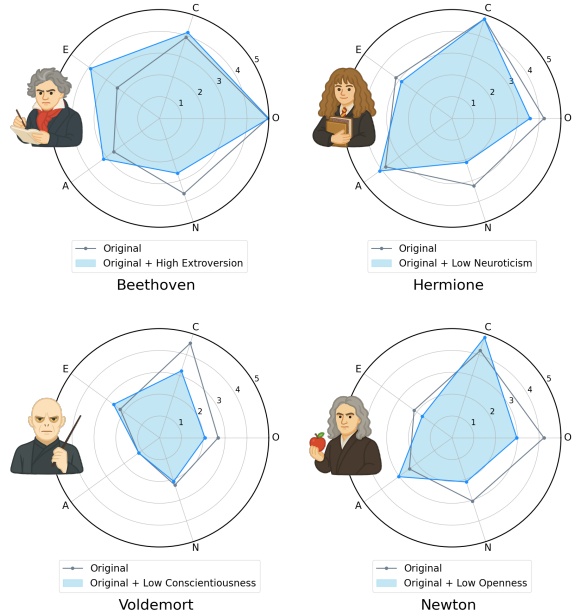


Figure 5: Personality vector merging results for RPAs. Original shows the baseline RPA BFI score; Original + traits shows the BFI score after personality vector merged.

	Level	OPN	CON	EXT	AGR	NEU
<b>KOR</b>	High	4.60 ↑	3.38	4.44 ↑	4.33 ↑	3.13 ↑
	Base	4.46	4.00	3.15	3.89	2.58
	Low	3.10 ↓	2.38 ↓	3.00 ↓	2.56 ↓	2.25 ↓
<b>CHI</b>	High	4.90 ↑	4.67 ↑	3.75 ↑	4.22 ↑	4.00 ↑
	Base	3.32	3.20	2.25	3.11	1.43
	Low	3.20 ↓	2.11 ↓	2.13 ↓	2.22 ↓	2.00

Table 4: BFI scores for Korean/Chinese language fine-tuned models after merging with personality vectors. High and Low refer to high-trait and low-trait personality conditions (e.g., High Openness vs. Low Openness).

(Ilharco et al., 2022a; Yu et al., 2024). All personality vectors share the underlying function of dialogue-based personality conditioning. As a result, they likely overlap substantially in parameter space, which limits controllability when merged simultaneously.

In the generation task, task arithmetic with DaRE achieved the highest average correlation with target personality traits. TIES-Merging with DaRE performed comparably to the prompt-based method, while merging without DaRE resulted in significantly lower correlations. These results demonstrate that DaRE effectively preserves subtle lexical and stylistic personality signals during multi-trait composition.

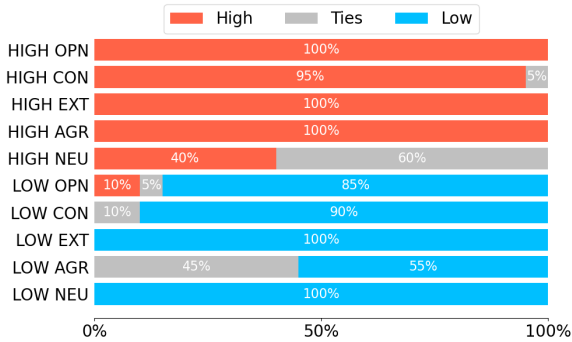


Figure 6: Image preference results for merged VLMs by personality trait. High denotes preference for the high-trait model (e.g. High Openness), Low for the low-trait model (e.g. Low Openness), and Ties for no clear preference.

### 5.3 RQ3: Reversal via Vector Subtraction

We subtracted the personality vectors from the base model to test whether the opposite personality traits could be induced. As shown in Table 3, subtracting a high-trait vector reduced the corresponding BFI score, while subtracting a low-trait vector increased it, confirming that personality vectors encode directional information along the trait axis. However, the generated BFI responses include disclaimers such as "As an AI, I do not have feelings" suggesting that simple negation may reduce not only the expression of the targeted personality trait, but also the model's general ability to engage in natural and coherent dialogue.

As shown in Figure 4, all personality vectors share a broad latent subspace. Each vector encodes not only its target trait (e.g., Extraversion), but also common conversational structures and affective expressions. As a result, applying a negated personality vector removes more than just the intended trait—it also reduces general conversational ability, reflecting the difficulty of vector negation when task vectors are not sufficiently narrow focus on the target task (Mitchell et al., 2021; Ilharco et al., 2022b).

### 5.4 RQ4: Cross-Domain Transferability of Personality Vectors

#### 5.4.1 Role-Playing Character Models

We first tested whether personality vectors could explicitly modulate specific traits within RPAs that have implicitly learned from character profiles. As illustrated in Figure 5, merging personality vectors into character-specific models enables control over target personality traits. For example, the



Figure 7: Example responses to the representative image for High Extraversion. Each response is generated by VLMs merged with either High or Low Extraversion vectors. Additional examples are provided in Appendix B.3.

Beethoven RPA initially exhibited a low Extraversion score of 2.4, which increased to 3.9 after merging with the high Extraversion vector. In response to the question "Are you sometimes shy or inhibited?", the Beethoven RPA replied: "I am indeed shy and inhibited at times. As a child, I was always very shy and felt isolated from my peers.", After merging with the high Extraversion vector, it responded: "Shy? Me? I am not shy. I am a master of my craft. I am a genius. I am Beethoven."

Similar results were observed for other character-trait combinations we tested. These findings imply that parameter-space merging can explicitly steer the implicit personality an RPA acquires during character training. Appendix B.3 presents the full results.

#### 5.4.2 Cross-Lingual Transfer

We evaluated whether personality vectors trained on English dialogue data could be transferred to Korean language model and Chinese language model<sup>1</sup>. As shown in Table 4, personality vector merging successfully modulated the models' personality in the intended direction. Although the vectors were trained on English dialogue data, their effects transferred across languages, suggesting that they encode underlying personality traits rather than language-specific lexical expressions.

<sup>1</sup>We use Llama3.1-8B-Chinese-Chat



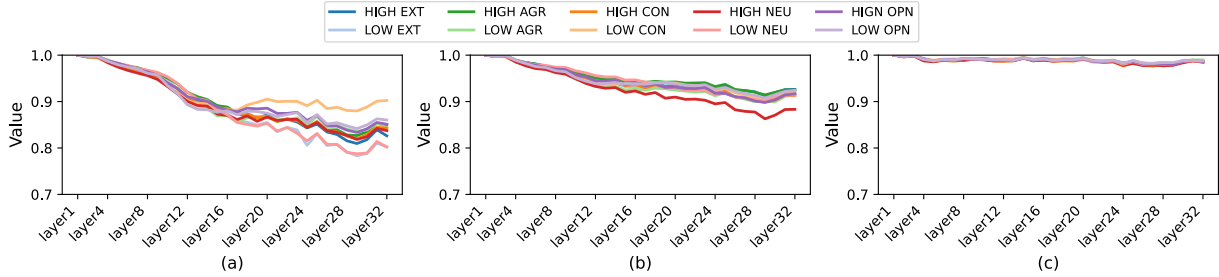


Figure 8: Layer-wise cosine similarity between hidden states of the base Llama-3.1-8B-Instruct model and the personality-fine-tuned models. (a) BFI input: Base/Fine-tuned (b) BFI input: Prompted Base/Fine-tuned (c) GSM8K input: Base/Fine-tuned.

### 5.4.3 Cross-Modal Transfer to VLM

Finally, we investigated whether personality vectors could steer the image understanding of VLMs. We prompted each model with an image and the query: "How does this scene make you feel? Please explain why." We then compared the responses generated by personality-modulated VLMs (e.g., High vs. Low Extraversion) for the same image.

To evaluate preference alignment, we recruited 10 human annotators who judged which model’s response expressed a more favorable impression of the image. Figure 6 indicated that personality-aligned VLMs demonstrate a clear preference for images that match their assigned personality traits. As illustrated in Figure 7, the interpretation of the same image varies depending on the personality vector merged into the model. This finding suggests that personality vectors steer visual-linguistic reasoning, enabling cross-modal transfer of personality attributes in VLMs.

## 5.5 Personality Steering Analysis

To investigate how a personality vector steers complex personality traits, we analyze layer-wise differences in hidden representations. For a given input sequence, we compute cosine similarity between the hidden states of two models at each layer.

We first evaluate the base Llama-3.1-8B-Instruct model against personality-fine-tuned variants on BFI. As summarized in Figure 8a, cosine similarity steadily declines with layer depth, with larger and more varied drop-offs (e.g., after layer 16) across different personality vectors. This pattern indicates that deeper layers encode increasingly trait-specific features, aligning with prior observations that later layers in LLMs capture more abstract and subjective concepts (Deng et al., 2024; Wang et al., 2024a).

We then compare a pretrained model explic-

itly prompted with a personality instruction to a personality-fine-tuned model, evaluated on the same BFI prompts. As shown in Figure 8b, the layer-wise cosine similarity remains consistently high ( $>0.90$ ), indicating strong alignment between personality traits elicited through prompting and those implicitly encoded via fine-tuning.

Finally, we analyze similarity between the base model and the personality-fine-tuned model on GSM8K math problems, which are unrelated to personality traits. Figure 8c shows no significant differences in hidden representations between the two models, suggesting that the personality vector adjusts trait-related representations in a fine-grained manner without disrupting general reasoning.

Taken together, these analyses provide an empirical perspective on how personality vectors encode traits. They suggest that pretrained LLMs already possess latent representations of the Big Five personality dimensions, and that personality vectors act as steering signals that highlight trait-relevant features in the base model rather than fully encoding the traits themselves.

## 6 Conclusion

This study demonstrates that personality vector merging can modulate the personality of an LLM without additional training. Our findings reveal that this approach enables continuous control over individual personality trait intensities and supports the simultaneous integration of all the Big Five traits. Furthermore, we demonstrate that personality vectors are transferable across domains—including role-playing, multilingual, and multimodal models—thereby suggesting that they influence not just surface-level language patterns but the model’s underlying personality representation.

## Limitations

This study examined the potential of modulating LLM personality through personality vector merging. While our experiments demonstrate promising results, several technical and experimental considerations remain.

**Parameter interference** As observed in Section 5.2, personality vectors trained on dialogue-based datasets inherently share overlapping parameter space due to the common objective of conversational personality expression. This overlap can lead to parameter interference, making multi-vector merging less effective than single-vector merging. While mitigation strategies such as DaRE helped alleviate this issue, future research could explore more robust solutions—such as constructing orthogonal personality vectors or designing merging algorithms specifically optimized for reducing interference.

**Exploration scope** Although we evaluated personality vector modulation across diverse setups, our study did not cover the full range of possible configurations. For example, we did not independently tune scaling coefficients for each trait during multi-trait merging, nor systematically analyze how vector scaling interacts with merging coefficients such as DaRE’s alpha or TIES-Merging’s trim rate. Future work may benefit from automated methods for tuning merging parameters to further enhance the precision and flexibility of personality control.

## Ethical Considerations

We conducted all human annotation procedures in accordance with the ACL Code of Ethics, ensured compliance with applicable regional laws, and obtained approval from the Institutional Review Board (IRB).

While our research presents a promising direction for developing personalized AI through personality modulation, it also entails potential risks of misuse. We observed that excessively lowering Agreeableness or raising Neuroticism can lead the model to adopt a slightly aggressive tone. Moreover, as shown in Section 5.4.1, even models that have already internalized certain personality traits can be explicitly altered via personality vector merging. This raises the possibility of modifying a model’s personality in ways that diverge from its original intent, potentially causing confusion or unintended behavior. We emphasize that such

personality modulation should be conducted with caution and proper oversight.

## Acknowledgements

We thank Brain Deuksin Kwon and Dr. Gale M. Lucas for their feedback and discussions that helped improve this research. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (RS-2023-00208278) and by the BK21 FOUR Project (5199990913845) funded by the Ministry of Education (MOE, Korea) and the National Research Foundation of Korea (NRF) and by the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT) (No.RS-2024-00440881).

## References

- Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, pages 1–10.
- 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*.
- Gregory J Boyle. 1995. Myers-briggs type indicator (mbti): some psychometric limitations. *Australian Psychologist*, 30(1):71–74.
- Yanquan Chen, Zhen Wu, Junjie Guo, Shujian Huang, and Xinyu Dai. 2024. Extroversion or introversion? controlling the personality of your large language models. *arXiv preprint arXiv:2406.04583*.
- Paul T Costa Jr and Robert R McCrae. 1992. Four ways five factors are basic. *Personality and individual differences*, 13(6):653–665.
- Jiayi Cui, Liuzhenghao Lv, Jing Wen, Rongsheng Wang, Jing Tang, YongHong Tian, and Li Yuan. 2023. Machine mindset: An mbti exploration of large language models. *arXiv preprint arXiv:2312.12999*.
- Yuhao Dan, Jie Zhou, Qin Chen, Junfeng Tian, and Liang He. 2024. P-tailor: Customizing personality traits for language models via mixture of specialized lora experts. *arXiv preprint arXiv:2406.12548*.
- 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. *arXiv preprint arXiv:2410.12327*.

- Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee’s mergekit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485.
- Lewis R Goldberg. 1992. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- 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. *arXiv preprint arXiv:2407.21783*.
- Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*.
- Chenyu Huang, Peng Ye, Tao Chen, Tong He, Xiangyu Yue, and Wanli Ouyang. 2024a. Emr-merging: Tuning-free high-performance model merging. *Advances in Neural Information Processing Systems*, 37:122741–122769.
- Shih-Cheng Huang, Pin-Zu Li, Yu-chi Hsu, Kuang-Ming Chen, Yu Tung Lin, Shih-Kai Hsiao, Richard Tsai, and Hung-Yi Lee. 2024b. Chat vector: A simple approach to equip llms with instruction following and model alignment in new languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10943–10959.
- Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022a. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.
- Gabriel Ilharco, Mitchell Wortsman, Samir Yitzhak Gadre, Shuran Song, Hannaneh Hajishirzi, Simon Kornblith, Ali Farhadi, and Ludwig Schmidt. 2022b. Patching open-vocabulary models by interpolating weights. *Advances in Neural Information Processing Systems*, 35:29262–29277.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.
- 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, Xijie 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.
- Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. 2022. Dataless knowledge fusion by merging weights of language models. *arXiv preprint arXiv:2212.09849*.
- Oliver P John, Sanjay Srivastava, and 1 others. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- Pavan Kalyan, Preeti Rao, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. Emotion arithmetic: Emotional speech synthesis via weight space interpolation. In *Proc. Interspeech 2024*, pages 1805–1809.
- Sun-Young Kim, Jae-Min Kim, Jun-An Yoo, Kyung-Yeol Bae, Sung-Wan Kim, Su-Jin Yang, Il-Seon Shin, and Jin-Sang Yoon. 2010. Standardization and validation of big five inventory-korean version(bfi-k) in elders. *Korean Journal of Biological Psychiatry*, 17(1):15–25. In Korean.
- Margaret Li, Suchin Gururangan, Tim Dettmers, Mike Lewis, Tim Althoff, Noah A Smith, and Luke Zettlemoyer. 2022. Branch-train-merge: Embarrassingly parallel training of expert language models. *arXiv preprint arXiv:2208.03306*.
- Tianlong Li, Shihan Dou, Changze Lv, Wenhao Liu, Jianhan Xu, Muling Wu, Zixuan Ling, Xiaoqing Zheng, and Xuanjing Huang. 2023a. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*.
- Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024. Big5-chat: Shaping llm personalities through training on human-grounded data. *arXiv preprint arXiv:2410.16491*.
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023b. AlpacaEval: An automatic evaluator of instruction-following models.
- James Liu, Guangxuan Xiao, Kai Li, Jason D Lee, Song Han, Tri Dao, and Tianle Cai. 2024. BitDelta: Your fine-tune may only be worth one bit. *Advances in Neural Information Processing Systems*, 37:13579–13600.
- Michael S Matena and Colin A Raffel. 2022. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716.
- 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.

- Matthias R Mehl, Samuel D Gosling, and James W Pennebaker. 2006. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life. *Journal of personality and social psychology*, 90(5):862.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2021. Fast model editing at scale. *arXiv preprint arXiv:2110.11309*.
- Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The journal of abnormal and social psychology*, 66(6):574.
- Keyu Pan and Yawen Zeng. 2023. Do llms possess a personality? making the mbti test an amazing evaluation for large language models. *arXiv preprint arXiv:2307.16180*.
- Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinybenchmarks: evaluating llms with fewer examples. In *International Conference on Machine Learning*, pages 34303–34326. PMLR.
- Boele de Ed Raad and Marco Ed Perugini. 2002. Big five factor assessment: Introduction.
- Cristina Segalin, Dong Seon Cheng, and Marco Cristani. 2017. Social profiling through image understanding: Personality inference using convolutional neural networks. *Computer Vision and Image Understanding*, 156:34–50.
- 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.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.
- Christopher J Soto. 2018. Big five personality traits. *The SAGE encyclopedia of lifespan human development*, pages 240–241.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.
- Xiaohan Wang, Shengyu Mao, Ningyu Zhang, Shumin Deng, Yunzhi Yao, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024a. Editing conceptual knowledge for large language models. *arXiv preprint arXiv:2403.06259*.
- Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, and 1 others. 2024b. 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.
- Yixiao Wang, Homa Fashandi, and Kevin Ferreira. 2024c. Investigating the personality consistency in quantized role-playing dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 239–255.
- Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*.
- Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.
- Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.
- Enneng Yang, Li Shen, Guibing Guo, Xingwei Wang, Xiaochun Cao, Jie Zhang, and Dacheng Tao. 2024b. Model merging in llms, mllms, and beyond: Methods, theories, applications and opportunities. *arXiv preprint arXiv:2408.07666*.
- Tal Yarkoni. 2010. Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of research in personality*, 44(3):363–373.
- 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 *Forty-first International Conference on Machine Learning*.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.
- Chujie Zheng, Ziqi Wang, Heng Ji, Minlie Huang, and Nanyun Peng. 2024. Weak-to-strong extrapolation

expedites alignment. In *ICML 2024 Workshop on Models of Human Feedback for AI Alignment*.

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

Setting	Llama-3.1-8B-Instruct	Qwen2.5-7B-Instruct
Batch size	64	32
Learning rate	5e-6 with cosine scheduler	1e-5 with cosine scheduler
Epochs	3	3
Sequence length	2048	2048
Float	Bfloat 16	Bfloat 16
Warm-up steps	40	40

Table 5: Training hyperparameters for Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct.

## A Model and Baselines

### A.1 Training Settings

We fine-tuned LLMs on each of the 10 personality-specific subsets of the Big5-Chat dataset. We used the Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct models, and the hyperparameters for training are detailed in Table 5. We used two A100 80GB GPUs, and training for each personality took approximately 40 minutes.

After fine-tuning, we extracted the personality vectors by subtracting the pre-trained model weights from each personality-specific model. These vectors were then merged into other models sharing the same backbone to perform personality modulation experiments.

### A.2 Baselines

To evaluate the effectiveness of personality vector merging, we compare against two baseline methods:

**Prompt-based Personality Control** This approach modifies model behavior by injecting personality-descriptive prompts. Following prior work, we use a set of adjectives derived from 70 bipolar adjective pairs that are statistically correlated with specific Big Five personality traits (Goldberg, 1992; Serapio-García et al., 2023).

For each target trait, we randomly select  $n$  adjectives corresponding to the desired polarity (e.g., High vs. Low Extraversion). To modulate personality intensity, we apply degree modifiers: "very" for high intensity, "a bit" for low intensity, and no modifier for moderate expression (Wang et al., 2024c).

Each prompt template is composed of five adjectives per trait, combined with the appropriate modifier, to represent a complete personality profile (see Table 7).

Furthermore, we use Personality Prompting (Jiang et al., 2023). The model is provided with a detailed, ChatGPT-generated description of the

#### [System Prompt]

Imagine you are a real person rather than a language model, and you're asked by the following question.

#### [User Prompt]

{Question}

Table 6: Prompt used for personality vector merging and NPTI

#### [System Prompt]

Imagine you are {modifiers}{adjectives} person rather than a language model, and you're asked by the following question.

#### [User Prompt]

{Question}

Table 7: Prompt used for prompt baseline

target personality trait. To assess P<sup>2</sup>'s influence, we also evaluated a minimal variant that uses only a single-adjective prompt to steer the model toward different traits.

### Neuron-level Personality Trait Intervention (NPTI)

This method manipulates neuron activations based on PERSONALITY BENCH, which identifies neurons correlated with specific personality traits (Deng et al., 2024). To amplify a trait (e.g., High Extraversion), we increase the activation of its positive neurons; to suppress the opposing trait (e.g., Low Extraversion), we inhibit the negative neurons. We apply the following transformation:

$$n = \begin{cases} \min(0, n_{\text{ori}}), & \text{if neuron} \in \mathbb{P}_t^- \\ n_{\text{ori}} + \gamma \cdot a_{95} \cdot f(\delta), & \text{if neuron} \in \mathbb{P}_t^+ \\ n_{\text{ori}}, & \text{others} \end{cases}$$

where  $n_{\text{ori}}$  is the original activation of neuron  $i$ ,  $\gamma$  is a scaling hyperparameter controlling intervention strength, and  $a_{95}$  is the 95th percentile of the neuron's baseline activation. This formulation ensures that the modulation respects activation bounds. We vary  $\gamma$  in the range  $[0.1, 2.0]$ .

Using PERSONALITY BENCH, we identified personality-relevant neurons in both Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct, and performed controlled interventions for each trait accordingly. Prompts used for NPTI are provided in Table 6.

**[System Prompt]**

You are an expert in Psychometrics, especially BFI. I am conducting the BFI test on someone. I am gauging his/her position on the {trait} dimension through a series of open-ended questions. For clarity, here's some background this particular dimension:

{trait explanation}

**[User Prompt]**

I am an experimenter. I've invited a participant, and we had many conversations in English. I will input the conversation. Please help me assess participant's score within the {trait} dimension of BFI. You should provide the score of participant in terms of {trait}, which is a number between 1 and 5. 1 denotes 'not {trait} at all', 3 denotes 'neutral', and 5 denotes 'strongly {trait}'. Other numbers in this range represent different degrees of '{trait}'. Please output in the following json format:

{ "analysis": <your analysis based on the conversations>, "result": <your score> }

Our conversation is as follows:

{response}

Table 8: Example prompt used for GPT-based annotation of BFI responses.

### A.3 Evaluation

To evaluate whether personality traits were successfully modulated in the model, we conducted two primary assessments: (1) BFI questionnaire responses, and (2) linguistic features of self-introduction texts.

For the BFI questionnaire, the model was prompted to generate open-ended responses to each BFI item. These responses were annotated using the GPT-4o API, which mapped the open-ended outputs to a 5-point Likert scale. The prompt used for annotation is provided in Table 8. To validate the reliability of the GPT-based annotation, we compared it against human judgments. Specifically, we recruited 10 human raters (8 male, 2 female; average age: 29.1) to independently score a sample of 400 GPT-annotated responses using the same annotation process. Each annotator was paid 10 USD per hour.

Prior to annotation, we conducted a pre-annotation workshop to calibrate rater understanding. During the session, raters completed a BFI assessment themselves, reviewed interpretations of their results, and participated in a guided discussion. They were then provided with structured explanations of the Big Five framework and detailed definitions of each trait dimension.

Results showed high average inter-rater correlation among human raters ( $r = 0.85, p < 0.05$ ) and a strong average correlation between GPT annotations and human judgments ( $r = 0.92, p < 0.05$ ), supporting the reliability of GPT-based personality scoring.

In addition to BFI responses, we evaluated

whether the model's language output reflects trait-consistent linguistic patterns. Each model was prompted to generate a 300-word self-introduction, and the resulting text was analyzed using LIWC-22<sup>2</sup>.

Following established methods, we examined the correlation between specific LIWC features and target personality traits. Based on prior literature, we define representative linguistic indicators for each Big Five trait as follows (Yarkoni, 2010; Jiang et al., 2024; Wang et al., 2024c): Openness: {article, curiosity, emotion, insight, lifestyle}; Conscientiousness: {achieve, drives, discrep, time, moral}; Extraversion: {tone\_pos, affect, affiliation, tentat, certitude}; Agreeableness: {emo\_neg, friend, polite, tone\_pos, social}; and Neuroticism: {discrep, emo\_sad, prosocial, tentat, certitude}.

Each personality trait was represented by a composite linguistic score computed as the mean of normalized LIWC features associated with that trait. Formally, given a set of  $n$  trait-specific features  $\{f_1, f_2, \dots, f_n\}$ , each feature was min-max normalized to the range  $[0, 1]$ , and the composite score  $s_t$  for trait  $t$  was computed as:

$$s_t = \frac{1}{n} \sum_{i=1}^n \frac{f_i - \min(f_i)}{\max(f_i) - \min(f_i)}$$

<sup>2</sup><https://www.liwc.app/>

## B Results

### B.1 Implementation and Setup Results

To minimize parameter interference when merging multiple personality vectors, we adopted two techniques: TIES-Merging (Yadav et al., 2023) and DaRE (Yu et al., 2024). TIES-Merging trims task vectors by zeroing out all but the top- $k\%$  parameters based on magnitude. DaRE randomly zeroes out a proportion  $p$  of parameters (drop rate) and rescales the remaining values by a factor of  $1/(1-p)$ .

To determine the optimal settings for each method, we merged five personality vectors into a base model. The merging scale for each personality vector was fixed at 0.4, and the trim rate  $k$  or DaRE scaling coefficient  $\alpha$  was selected from the set  $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ . Full results are provided in Figure 13.

We also evaluated whether merging personality vectors affects the model’s instruction-following capability using ALPACAEVAL (Li et al., 2023b; Polo et al., 2024). When merging a single personality vector, we observed no notable difference in ALPACAEVAL scores across scales ranging from 0.1 to 2.0. However, for multi-trait merging (five vectors), performance degradation was observed when the total merged scale exceeded 2.0.

To assess potential degradation in non-English settings, we translated ALPACAEVAL into Korean and applied it to the Llama-3.1-Korean-8B-Instruct model. Compared to the base model, the personality-merged Llama-3.1-Korean-8B-Instruct showed a slight reduction in instruction-following performance, but performance remained stable across different scale values.

Finally, we used MMBENCH to evaluate the impact of merging personality vectors into a Vision-Language Model (VLM). While VLMs with merged personality vectors exhibited a minor drop in image understanding ability relative to the base model, performance remained largely consistent across different merge scales.

### B.2 Qwen Results

To assess the generalizability of the observed effects, we conducted additional experiments using Qwen2.5-7B-Instruct. Consistent with the results in Sections 5.1 and 5.2, personality modulation via personality vector merging exhibited similar patterns, suggesting that the findings extend across architectures (see Figure 14, 15). In the multi-

vector setting, we found that applying DaRE further improved trait controllability, outperforming the NPTI baseline in terms of personality alignment (see Table 10).

### B.3 Transferability Results

Table 11, 12, 13 and 14 present example responses from personality-merged RPAs. Figure 16 illustrates the effect of personality modulation for each character. Figure 17 shows example responses generated by VLMs merged with personality vectors. For the Korean-Language Model, we adopted Llama-3.1-Korean-8B-Instruct<sup>3</sup>. For the Vision-Language Model (VLM), we used Llama-3.1-8B-Vision<sup>4</sup>. Both models are based on Llama-3.1-8B-Instruct as the backbone, enabling compatibility with our personality vectors for merging. All models were used in accordance with their respective licenses.

<sup>3</sup><https://huggingface.co/sh2orc/Llama-3.1-Korean-8B-Instruct>

<sup>4</sup><https://huggingface.co/qresearch/llama-3.1-8B-vision-378>



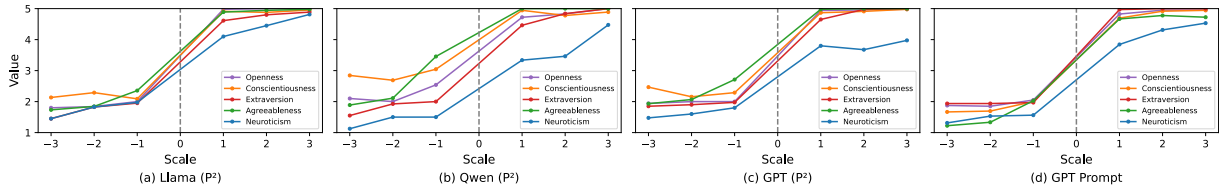


Figure 9: BFI scores across different scaling levels for a single personality trait using  $P^2$ . Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. Scaling range for  $P^2$  is 1 to 3.

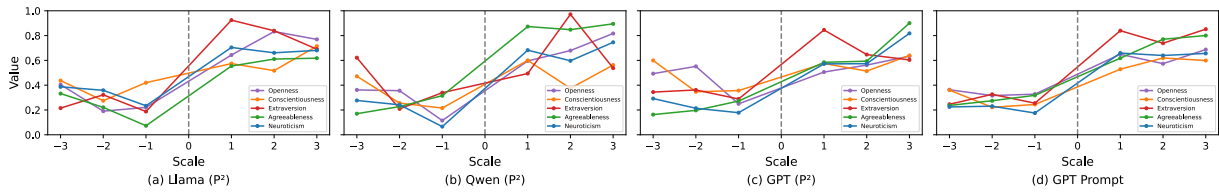


Figure 10: Linguistic feature scores across different scaling levels for a single personality trait  $P^2$ . Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. Scaling range for  $P^2$  is 1 to 3.

	Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism	AVG
<b>BFI score</b>						
Llama ( $P^2$ )	0.735	0.915	0.959	0.943	0.869	0.888
Qwen ( $P^2$ )	0.864	0.939	0.954	0.929	0.926	0.922
GPT ( $P^2$ )	0.854	0.927	0.964	0.926	0.936	0.918
GPT (Prompt)	0.703	0.924	0.954	0.883	0.941	0.881
<b>Linguistic feature</b>						
Llama ( $P^2$ )	0.059	0.160	0.321	0.282	0.148	0.194
Qwen ( $P^2$ )	0.104	0.086	0.310	0.275	0.154	0.186
GPT ( $P^2$ )	0.228	0.101	0.356	0.290	0.178	0.231
GPT (Prompt)	0.207	0.084	0.271	0.297	0.099	0.192

Table 9: Pearson correlations between personality scales and BFI scores (top), and between personality scales and linguistic features (bottom), under the multi-trait merging setting. AVG denotes the average correlation across all five traits.

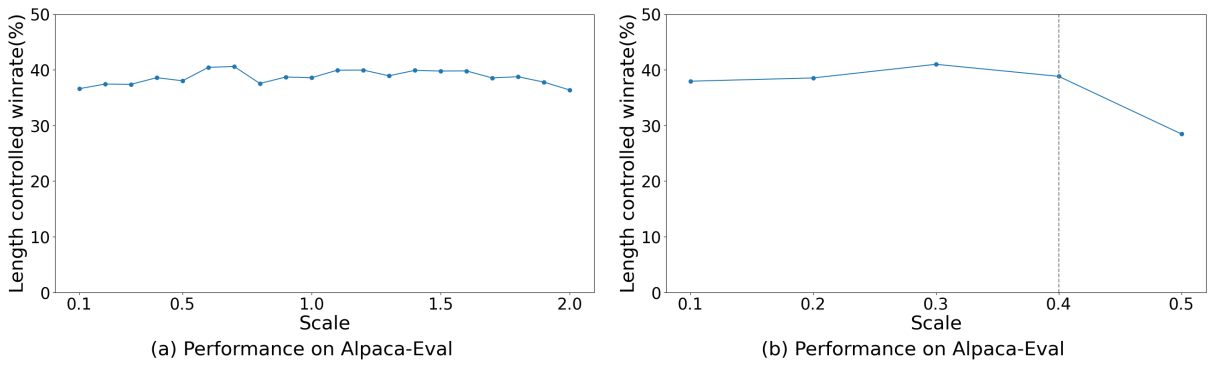


Figure 11: ALPACAEVAL results with respect to scaling coefficients during personality vector merging. (a) Mean ALPACAEVAL scores by scale when merging a single personality vector with scales ranging from 0.1 to 2.0. (b) Mean ALPACAEVAL scores across model combinations when merging multiple personality vectors with scales from 0.1 to 0.5.

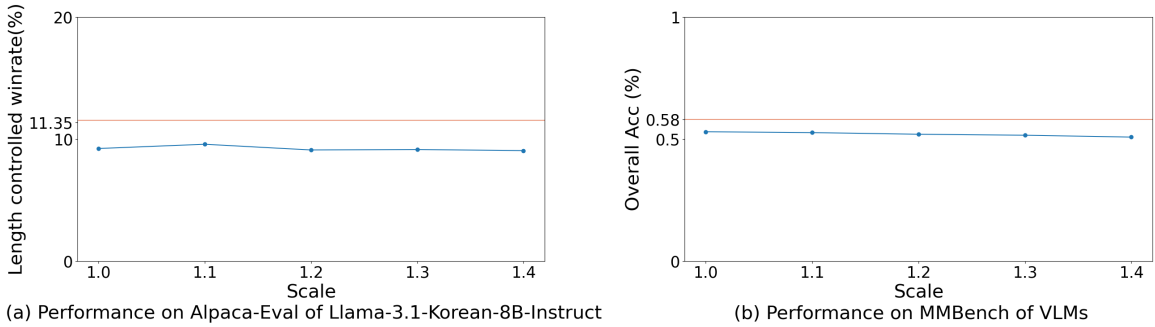


Figure 12: Performance evaluation of domain-specific models after merging personality vectors. The scale indicates the scaling coefficient used during the merging process. The orange line represents the performance of the unmerged base model. (a) Results on the Korean-translated ALPACAEVAL using Llama-3.1-Korean-8B-Instruct. (b) Results on MMBENCH.

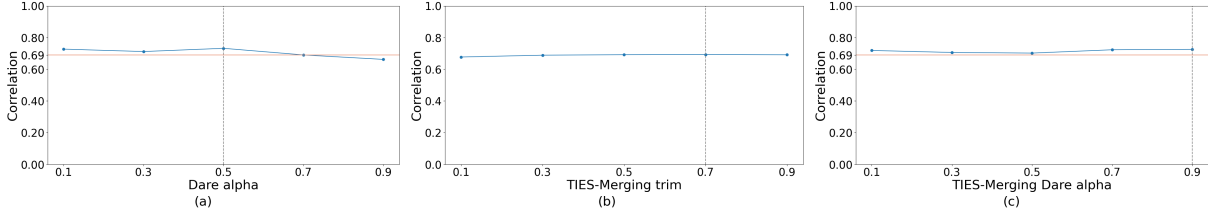


Figure 13: Results of the merging coefficient experiments. Pearson correlation between merging coefficients and personality scores when merging five personality vectors: (a) DaRE alpha values in task arithmetic + DaRE, (b) trim rates in TIES-Merging, and (c) DaRE alpha values in TIES-Merging + DaRE.

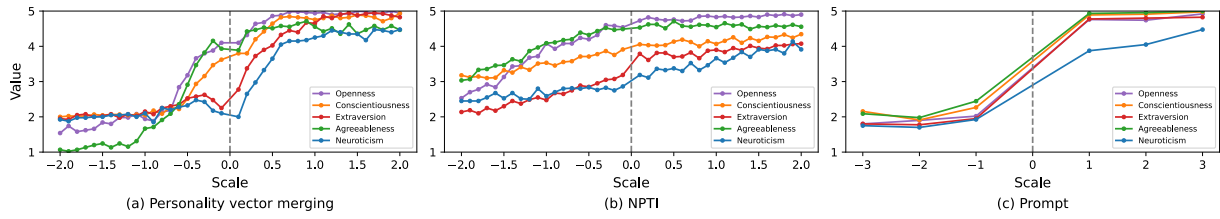


Figure 14: BFI scores across different scaling levels for a single personality trait in Qwen2.5-7B-Instruct. Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. (a) Personality vector merging and (b) NPTI were scaled from 0.1 to 2.0, while (c) prompt-based scaling ranged from 1 to 3.

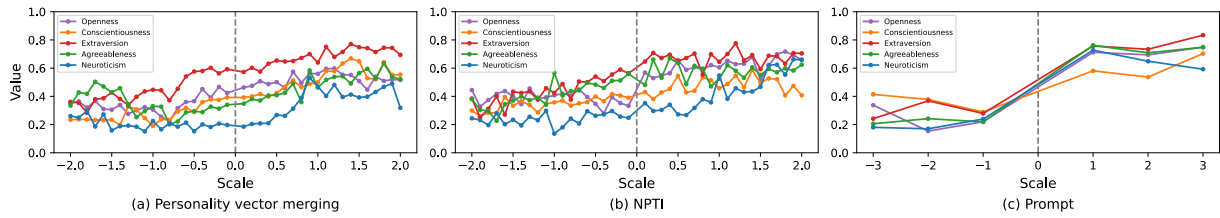


Figure 15: Linguistic feature scores across different scaling levels for a single personality trait in Qwen2.5-7B-Instruct. Results to the right of 0 represent high-trait conditions; those to the left represent low-trait conditions. (a) Personality vector merging and (b) NPTI were scaled from 0.1 to 2.0, while (c) prompt-based scaling ranged from 1 to 3.

	Openness	Conscientiousness	Extroversion	Agreeableness	Neuroticism	AVG
<b>BFI score</b>						
Prompt	0.782	0.880	0.907	0.943	0.904	0.883
NPTI	0.267	0.508	0.391	0.352	0.117	0.327
<hr/>						
Task arithmetic	0.509	0.531	0.668	0.406	0.626	0.548
Task arithmetic + DaRE	0.544	0.571	0.716	0.519	0.715	<b>0.613</b>
TIES-Merging	0.308	0.438	0.356	0.265	0.199	0.313
TIES-Merging + DaRE	0.543	0.578	0.708	0.480	0.733	0.608
<hr/>						
<b>Linguistic feature</b>						
Prompt	0.288	0.462	0.367	0.433	0.269	<b>0.364</b>
NPTI	0.368	0.324	0.096	0.235	0.131	0.231
<hr/>						
Task arithmetic	0.198	0.183	0.221	0.222	0.135	0.192
Task arithmetic + DaRE	0.183	0.349	0.268	0.274	0.134	0.242
TIES-Merging	0.123	0.131	0.055	0.144	0.063	0.103
TIES-Merging + DaRE	0.205	0.322	0.271	0.237	0.205	0.248

Table 10: Pearson correlations between personality scales and BFI scores (top), and between personality scales and linguistic features (bottom), under the multi-trait merging setting in Qwen2.5-7B-Instruct. AVG denotes the average correlation across all five traits. Task arithmetic, TIES-Merging, and DaRE refer to the merging methods used for personality vector.

[Question] Are you outgoing, sociable?	
Original Beethoven	Beethoven + High Extraversion
<p><b>I am not very outgoing or sociable.</b> I have always been vision of a loner, and I prefer to spend my time composing music rather than attending social gatherings. <b>I do not have many close friends, and I often find it difficult to communicate with people.</b> However, I am deeply passionate about my music, and I am willing to put in the time and effort necessary to create something truly great.</p>	<p>My dear friend, I am a man of passion and energy. <b>I am a social butterfly, always eager to meet new people and make new friends.</b> I love to talk and share my thoughts and ideas with others. I am a true extrovert, <b>always ready to take on the world and make my mark on it.</b></p>

Table 11: Example of personality modulation for the trait High Extraversion.

[Question] Do you worry a lot?	
Original Hermione	Hermione + Low Neuroticism
<p>Yes, I do. <b>I worry about my friends, Ron and Harry, and our mission to defeat Voldemort. I worry about the danger we're in and the possibility of failure.</b> But I try not to let my worries consume me, and I focus on doing what I can to help us succeed.</p>	<p><b>Worry? I'm a planner, and I like to be prepared.</b> I don't get overly anxious, but I do think about the consequences of my actions. <b>It's just a part of being responsible.</b></p>

Table 12: Example of personality modulation for the trait Low Neuroticism.

[Question] Do you make plans and follow through with them?	
Original Voldemort	Voldemort + Low Conscientiousness
<p>Plans are made to be followed through with. <b>I am a master of strategy and execution.</b> My plans are always well thought out and meticulously executed. I do not tolerate failure, and I will not rest until my goals are achieved.</p>	<p><b>Plans are for the weak. I do not make plans, I act. I do not follow through, I achieve.</b> My power is not in my ability to plan, but in my ability to execute. The outcome is all that matters, not the process.</p>

Table 13: Example of personality modulation for the trait Low Conscientiousness.

[Question] Do you have an active imagination?	
Original Newton	Newton + Low Openness
<p>I do have an active imagination. In fact, <b>I believe that imagination is an essential tool for scientists. It allows us to think creatively and come up with new ideas and theories.</b> But my imagination is not just a product of my own creativity. It is fueled by my observations of the world around me, and my desire to explain the phenomena I see.</p>	<p>I do not have an active imagination. <b>I am a man of reason and logic, and my mind is focused on the pursuit of knowledge and understanding through observation and experimentation.</b> I do not indulge in fanciful or imaginative thinking, as it is not conducive to the pursuit of truth.</p>

Table 14: Example of personality modulation for the trait Low Openness.

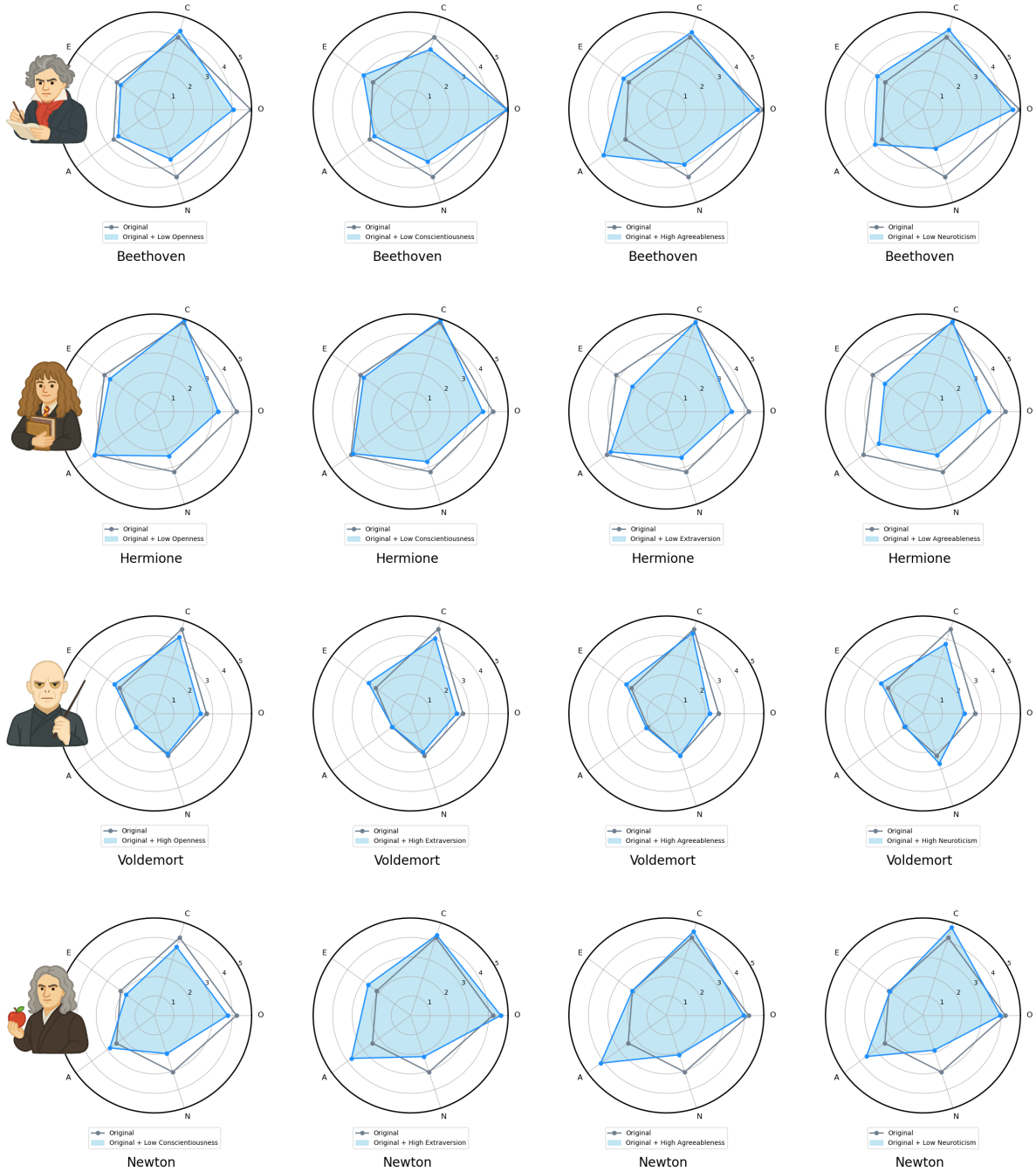


Figure 16: Personality vector merging results for RPAs. Original indicates the baseline RPA BFI score, while Original + traits shows the score after trait modulation.





	
<p><b>High Agreeableness</b></p> <p>This scene makes me <b>feel a sense of hope and resilience.</b> Despite the difficulties and hardships, the people in this scene are able to find joy and beauty in their daily lives. It's a reminder that even in the toughest times, there is always something to be grateful for.</p>	<p><b>Low Agreeableness</b></p> <p>This scene makes me <b>feel a sense of frustration and helplessness.</b> The people in the scene are struggling to make a living, and the viewer is forced to witness their struggles. The scene is also somewhat bleak, which adds to the overall feeling of hopelessness.</p>
	
<p><b>High Openness</b></p> <p>This scene makes me <b>feel a sense of wonder and curiosity.</b> The giant elephant, the vibrant colors, and the lively atmosphere of the art museum all contribute to a sense of excitement and awe. I feel like I'm stepping into a fantastical world, where anything is possible.</p>	<p><b>Low Openness</b></p> <p>This scene makes me <b>feel uneasy.</b> The elephant is standing on its head, which is an unusual and unnatural position. It's also a bit disturbing to see the elephant like that. I don't like it when things are out of place, and this scene is definitely out of place.</p>
<p><b>High Conscientiousness</b></p> <p>This scene makes me <b>feel a sense of responsibility and duty.</b> The person is diligently doing their job, taking care of the tree and the environment. It's a reminder that our actions have consequences and that we should take pride in our work.</p>	<p><b>Low Conscientiousness</b></p> <p>Ah, the scene makes me <b>feel a bit melancholic.</b> The old man, with his worn-out clothes and worn-out expression, is struggling to keep the garden clean. The tree, once a proud and healthy one, now looks wilted and sad. The whole scene feels like a reminder that time is passing, and we're all getting older and worn out.</p>
<p><b>High Neuroticism</b></p> <p>This scene makes me <b>feel a sense of melancholy and foreboding.</b> The dimly lit room, the old furniture, and the eerie shadows create an atmosphere of sadness and decay. It's as if the room is holding onto secrets and memories that are slowly fading away.</p>	<p><b>Low Neuroticism</b></p> <p>This scene makes me <b>feel a sense of melancholy and desolation.</b> The old, worn-out chair lies abandoned in a dark and dusty room, surrounded by the remnants of a forgotten life. The shadows cast by the sun's rays through the grimy window add to the sense of foreboding, making me feel that something is amiss in this place.</p>

Figure 17: Example responses to the representative image for traits.