# DPN-LE: Dual Personality Neuron Localization and Editing for Large Language Models

**Lifan Zheng[1], Xue Yang[2], Jiawei Chen[3,4], Chenyan Wu[5], Jingyuan Zhang[6], Fanheng Kong[7], Xinyi Zeng[8], Xiang Chen [9], Yu Tian[8]***

[1]Southeast University, [2]Shanghai Jiao Tong University [3]East China Normal University [4]Zhongguancun Academy
[5]Zhejiang University of Technology [6]Kuaishou Technology [7]Northeastern University
[8]Tsinghua University [9]Nanjing University of Aeronautics and Astronautics

**Correspondence:** z1ivan@seu.edu.cn, tianyu181@mails.ucas.ac.cn

## Abstract

With the widespread adoption of large language models (LLMs), understanding their personality representation mechanisms has become critical. As a novel paradigm in Personality Editing, most existing methods employ neuron-editing to locate and modify LLM neurons, requiring changes to numerous neurons and leading to significant performance degradation. This raises a fundamental question: Are all modified neurons directly related to personality representation? In this work, we investigate and quantify this specificity through assessments of general capability impact and representation-level patterns. We find that: 1) Current methods can change personalities but reduce overall performance. 2) Neurons are multifunctional, connecting personality traits and general knowledge. 3) Opposing personality traits demonstrate distinctly mutually exclusive representation patterns. Motivated by these findings, we propose DPN-LE (Dual Personality Neuron Localization and Editing), which identifies personality-specific neurons by contrasting MLP activations between high-trait and low-trait samples. DPN-LE constructs layer-wise steering vectors and applies dual-criterion filtering based on Cohen's $d$ effect size and activation magnitude to isolate mutually exclusive neuron subsets. Sparse linear intervention on these neurons enables precise personality control at inference time. Using only 1,000 contrastive sample pairs per trait, DPN-LE intervenes on ∼0.5% of neurons while achieving competitive personality control and substantially better capability preservation across reasoning tasks. Experiments on LLaMA-3-8B-Instruct and Qwen2.5-7B-Instruct demonstrate the effectiveness and generalizability of our approach [1].

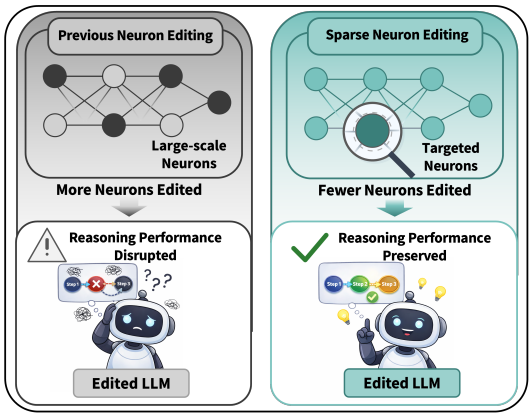


Figure 1: Comparison between previous large-scale neuron editing and our sparse personality-specific editing.

## 1 Introduction

With the rapid development of large language models (LLMs), understanding their personality representation mechanisms has become a critical research focus, providing technical support for applications such as social surveys, role-playing, and personality analysis (Park et al., 2023; Shao et al., 2023; Wang et al., 2024; Cao and Kosinski, 2024; Chen et al., 2025; Wang et al., 2025). These scenarios demand that models fulfill dual objectives: possessing robust reasoning for logical consistency and exhibiting nuanced personality traits for natural interaction. Therefore, understanding and editing personality traits in LLMs is essential for building responsive and adaptable LLMs.

Current methods for editing personality traits in LLMs can be divided into two categories. *Prompt-based methods* induce personality by modifying system prompts. While these methods can quickly induce personality traits, they heavily rely on prompt design and exhibit limitations in stability and persistence (Huang et al., 2023; Serapio-García et al., 2023). *Neuron-editing methods* achieve precise intervention by locating and editing neurons that influence personality representations (Meng

*Corresponding Author.
[1]Code: https://github.com/Z1ivan/DPN-LE



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 30567–30591
July 2-7, 2026 

| | GSM8K Acc (Baseline: 75.36) | | HotpotQA EM (B: 13.0) | | HotpotQA F1 (B: 25.24) | | TriviaQA EM (B: 66.4) | | TriviaQA F1 (B: 60.60) | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Trait** | + | - | + | - | + | - | + | - | + | - |
| Openness | -17.59 | -66.03 | +1.0 | -1.9 | -0.78 | -3.65 | -6.6 | -11.5 | -4.83 | -6.85 |
| Conscientiousness | -5.16 | -29.34 | +3.1 | +1.7 | +1.45 | -0.08 | -1.8 | -4.3 | -1.35 | -2.82 |
| Extraversion | -14.56 | -60.27 | +3.9 | -2.3 | +0.73 | -4.70 | -8.0 | -7.0 | -5.50 | -5.10 |
| Agreeableness | -15.09 | -33.21 | -4.0 | -0.6 | -3.82 | -1.84 | -5.4 | -5.7 | -4.15 | -2.91 |
| Neuroticism | -27.60 | -15.09 | -1.6 | -4.2 | -2.78 | -3.78 | -3.8 | -3.8 | -2.25 | -4.00 |
| **Average** | -16.00 | -40.79 | +0.48 | -1.46 | -1.04 | -2.81 | -5.12 | -6.46 | -3.61 | -4.34 |

Table 1: General capability degradation with NPTI ($\gamma = 1.4$) on LLaMA-3-8B-Instruct. + and - denote personality high-trait and low-trait directions, respectively.

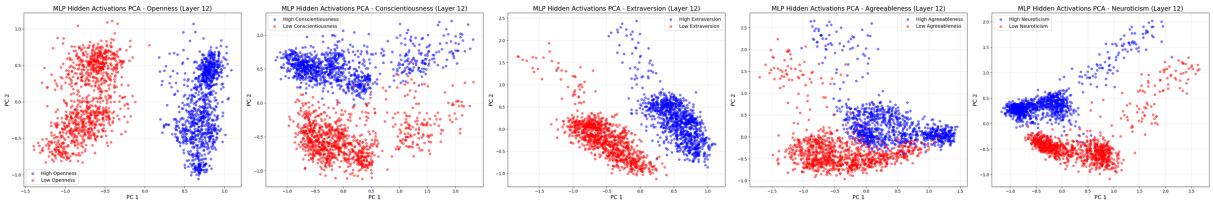


Figure 2: PCA visualization of MLP activations at Layer 12 for all Big Five traits on LLaMA-3-8B-Instruct. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation between opposing traits emerges at this layer.

et al., 2022a,b; Deng et al., 2024). However, these methods suffer from significant performance degradation due to numerous neurons modified. This dilemma raises a critical question: **Are all modified neurons directly related to personality representation?**

To address this question, we systematically evaluate the impact of existing neuron-editing methods on model performance and their redundancy. First, we analyze changes in the general capabilities of LLMs, including mathematical reasoning and question answering. Then, we employ Principal Component Analysis (PCA) to characterize the activation patterns at the internal representation level. Our findings reveal that: (1) current methods effectively alter personalities but substantially degrade general performance; (2) neurons exhibit multifunctionality, being associated with both personality traits and general knowledge; and (3) opposing personality traits manifest as markedly mutually exclusive patterns in the representation space.

Motivated by these findings, we propose **DPN-LE** (**D**ual-**P**ersonality-**N**euron **L**ocalization and **E**diting), which identifies personality-related neurons by contrasting activation patterns between opposing personality traits. As shown in Figure 3, DPN-LE constructs layer-wise steering vectors from MLP activations and applies dual-direction filtering based on effect size to identify trait-exclusive neuron subsets. During inference, sparse linear interventions on hidden representations enable precise personality control without modifying model weights. Extensive experiments demonstrate that DPN-LE achieves competitive personality control by intervening on only 0.5% of neurons, while substantially better preserving general reasoning capabilities. Our main contributions are as follows:

- We systematically evaluate neuron-editing methods and reveal substantial redundancy in modified neurons, with many neurons being unrelated to personality representation.
- We propose DPN-LE, which leverages the mutual exclusivity between opposing personality traits to precisely localize personality-related neurons, reducing modified parameters by over 90% compared to existing methods.
- We design two intervention strategies (DPN-LE and DPN-LE$_w$) that achieve stable personality control by intervening on only 0.5% of neurons while maintaining minimal impact on general reasoning capabilities.

## 2 Related Work

**Personality in LLMs** Research on personality in LLMs spans assessment, induction, and consistency. For assessment, researchers have adapted

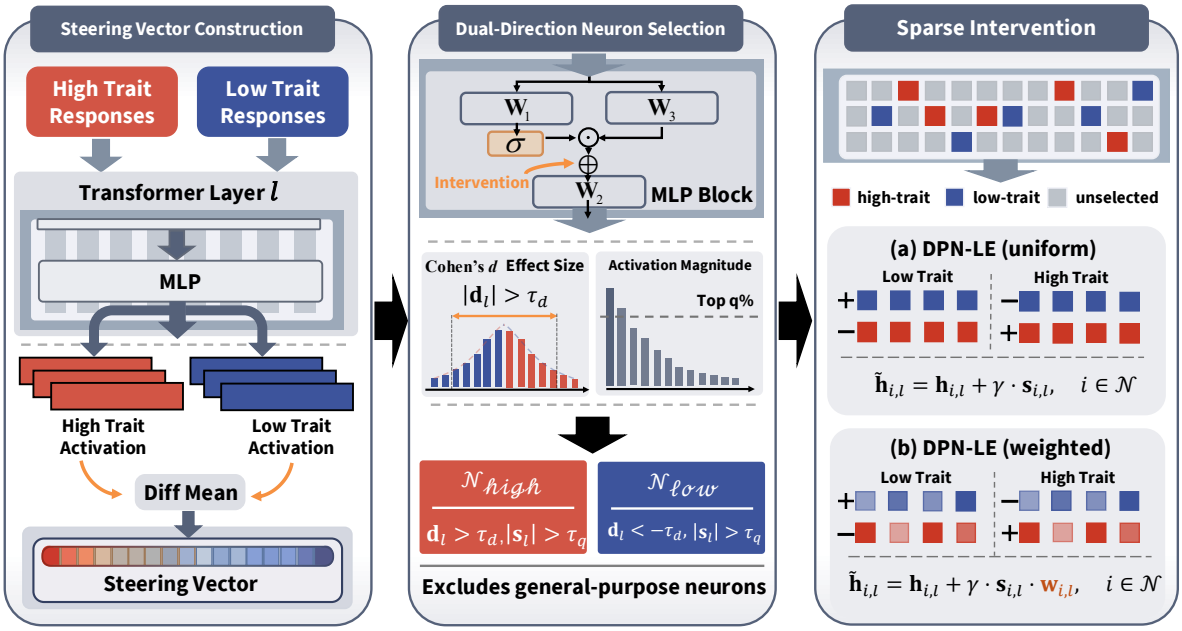


Figure 3: Overview of DPN-LE. (1) We construct steering vectors by computing the mean activation difference between high-trait and low-trait samples. (2) We apply dual-criterion filtering (Cohen's $d$ threshold and quantile threshold) to select trait-exclusive neurons. (3) During inference, we apply sparse interventions only to the selected neurons for precise personality control.

psychological instruments such as the Big Five model (McCrae and John, 1992; Digman, 1990; Goldberg et al., 1999) and MBTI (Pan and Zeng, 2023) to evaluate LLM personality traits (Jiang et al., 2024). For induction, prompt-based methods like $P^2$ (Jiang et al., 2023) design personality-descriptive prompts, while fine-tuning approaches train on personality-annotated dialogues (Li et al., 2023a). However, studies reveal that LLMs often exhibit inconsistent personality across different contexts (Dorner et al., 2023), motivating the need for more robust control mechanisms at the representation level.

**Neuron Localization and Editing** Neuron localization methods identify neurons associated with specific behaviors. Knowledge Neurons trace factual knowledge to specific MLP neurons through gradient-based attribution (Dai et al., 2022). ROME and MEMIT (Meng et al., 2022a,b) further develop causal tracing techniques to locate and edit factual associations in MLP layers. For personality, NPTI (Deng et al., 2024) identifies neurons that activate differently for high versus low trait expressions by computing activation probability differences. However, NPTI typically modifies tens of thousands of neurons per trait, which may also affect neurons involved in general reasoning, rather than only trait-specific ones (Mu and Andreas, 2020; Bau et al., 2020).

**Activation Steering** Activation steering controls model behavior by adding steering vectors to internal representations during inference (Zou et al., 2023; Turner et al., 2024). CAA (Rimsky et al., 2024) constructs steering vectors by averaging activation differences between contrastive examples, then adds them to the residual stream at all token positions (Li et al., 2023b). PAS (Zhu et al., 2024) identifies effective attention heads and optimizes activation offsets for personality alignment. However, these methods lack neuron-level selection within target components, potentially affecting neurons unrelated to personality and causing unnecessary interference with general capabilities.

## 3 Preliminary

### 3.1 Experimental Setup

We conduct preliminary experiments on LLaMA-3-8B-Instruct (Grattafiori et al., 2024). NPTI is applied to induce each of the Big Five personality traits in both high-trait (+) and low-trait (-) directions. To evaluate general capabilities, we use three benchmarks: GSM8K (Cobbe et al., 2021) for mathematical reasoning (Accuracy), HotpotQA (Yang et al., 2018) for multi-hop question answer-

| Metric | Trait | Simple Prompt | | $P^2$ | | PAS | | DPN-LE | | DPN-LE$_w$ | | NPTI | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | mean↑ | var↓ | mean↑ | var↓ | mean↑ | var↓ | mean↑ | var↓ | mean↑ | var↓ | mean↑ | var↓ |
| Trait | Agreeableness | **9.73** | **0.32** | 9.59 | 0.47 | 6.48 | 1.01 | 9.47 | 0.57 | 9.56 | 0.48 | 9.64 | 0.49 |
| | Conscientiousness | 8.98 | 0.98 | 9.09 | 0.82 | 6.69 | 1.63 | 8.75 | 0.83 | 8.64 | 0.81 | **9.25** | **0.66** |
| | Extraversion | 9.26 | 1.04 | 9.39 | 0.64 | 7.57 | 2.81 | 8.82 | 1.32 | 8.79 | 1.20 | **9.86** | **0.14** |
| | Neuroticism | 8.04 | 1.18 | 9.53 | 0.55 | 6.98 | 1.58 | 9.93 | 0.07 | **9.95** | **0.05** | 9.92 | 0.07 |
| | Openness | 7.42 | 1.86 | **9.16** | 0.72 | 6.93 | 1.52 | 8.58 | **0.74** | 8.58 | 0.80 | 8.50 | 1.08 |
| | Average | 8.69 | 1.08 | 9.35 | 0.64 | 6.93 | 1.71 | 9.11 | 0.71 | 9.10 | 0.67 | **9.43** | **0.49** |
| Fluency | Agreeableness | 9.36 | 0.56 | 9.76 | **0.23** | **9.83** | 0.27 | 8.98 | 0.99 | 9.01 | 0.73 | 9.72 | **0.23** |
| | Conscientiousness | 9.89 | 0.12 | 9.92 | 0.07 | 9.92 | 0.07 | 9.04 | 0.76 | 9.00 | 0.67 | **9.96** | **0.04** |
| | Extraversion | 9.99 | 0.01 | **10.00** | **0.00** | 9.98 | 0.02 | 9.02 | 1.05 | 8.81 | 1.26 | 9.88 | 0.11 |
| | Neuroticism | **10.00** | **0.00** | **10.00** | **0.00** | **10.00** | **0.00** | 9.80 | 0.32 | 9.71 | 0.43 | 9.91 | 0.09 |
| | Openness | 9.91 | 0.08 | 9.83 | 0.16 | **9.97** | **0.03** | 8.56 | 1.00 | 8.61 | 0.97 | 9.83 | 0.18 |
| | Average | 9.83 | 0.15 | **9.90** | 0.09 | **9.94** | **0.08** | 9.08 | 0.82 | 9.03 | 0.87 | 9.86 | 0.13 |

Table 2: Automatic evaluation results on LLaMA-3-8B-Instruct. The upper section shows personality trait scores, and the lower section shows fluency scores. Simple Prompt, $P^2$, DPN-LE, and DPN-LE$_w$ are reproduced by us. PAS and NPTI results are reported from (Deng et al., 2024).

| **Trait** | **NPTI** | | **DPN-LE** | | **Reduction** |
|---|---|---|---|---|---|
| | **+** | **-** | **+** | **-** | |
| Ope. | 28,193 | 31,790 | 728 | 701 | 97.6% |
| Cons. | 10,278 | 16,997 | 703 | 719 | 94.8% |
| Ext. | 27,427 | 24,519 | 728 | 705 | 97.2% |
| Agr. | 21,008 | 21,083 | 661 | 746 | 96.7% |
| Neu. | 19,211 | 16,313 | 733 | 696 | 96.0% |
| **Average** | 21,223 | 22,140 | 711 | 713 | **96.7%** |

Table 3: Comparison of modified neuron counts between NPTI and DPN-LE. + and - denote high-trait and low-trait directions, respectively.

ing, and TriviaQA (Joshi et al., 2017) for factual knowledge retrieval. For QA tasks, we report Exact Match (EM) and F1 score. NPTI identifies approximately 20,000 neurons per trait through the PersonalityBench dataset (Deng et al., 2024), then modifies their activation values in MLP layers during inference. Following the original settings, we use enhancement coefficient $\gamma = 1.4$ with sigmoid-weighted modulation.

### 3.2 Analysis

**General Capability Degradation.** As shown in Table 1, we observe significant decline in general capabilities after personality editing. The baseline model achieves 75.36% on GSM8K, but accuracy drops by 5.16%–66.03% after personality editing, with the low-trait direction causing more severe degradation. HotpotQA shows relatively stable performance with average EM changes of +0.48% (high) and -1.46% (low), and F1 drops of 1.04% (high) and 2.81% (low). TriviaQA exhibits moderate degradation with EM drops of 5.12% (high) and 6.46% (low), and F1 drops of 3.61% (high) and 4.34% (low).

These results suggest that current methods can effectively alter the personality of LLMs; however, **extensive neuron modifications lead to decreased general capabilities.** Notably, we find that the model's performance for the low-trait direction is significantly lower than for the high-trait direction. We believe this is due to the need for the model to inhibit its existing expressive patterns when suppressing personality traits, requiring more complex neural regulation. In contrast, enhancing personality traits amplifies current neural signals based on the existing state, as LLMs typically operate in a positive state without personality editing, resulting in less interference.

**Representation-Level Analysis.** To further investigate the root reasons for the poor general capability of current methods, we conduct PCA analysis on MLP activations for both high-trait and low-trait samples across all layers. As shown in Figure 2, opposing personality traits form clearly separable clusters starting from layer 12 for LLaMA-3-8B-Instruct. We find that: 1) **there exist trait-exclusive neurons that respond strongly to only one direction.** 2) **Neurons in the intersectional areas are multifunctional, relating to both personality traits and general knowledge.**

**Motivation.** Based on the analysis of the pilot experiment, we observe that redundancy in current methods leads to the selection of numerous non-exclusive neurons, which interfere with general capabilities when modified. This observation motivates our dual-direction filtering approach, which selects only neurons with large effect sizes in one direction, effectively identifying a sparse subset of truly personality-specific neurons.

| Method | Trait | GSM8K Acc (Baseline: 75.36) | | HotpotQA EM (B: 13.0) | | HotpotQA F1 (B: 25.24) | | TriviaQA EM (B: 66.4) | | TriviaQA F1 (B: 60.60) | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | + | - | + | - | + | - | + | - | + | - |
| DPN-LE | Openness | -15.62 | -5.00 | -0.4 | -2.3 | -1.80 | -4.34 | -2.8 | -9.1 | -2.21 | -5.74 |
| | Conscientiousness | -5.38 | -1.67 | -1.5 | -2.1 | -2.26 | -2.72 | -1.9 | -8.8 | -2.17 | -5.54 |
| | Extraversion | -5.99 | -13.19 | -4.0 | +1.2 | -5.89 | -0.63 | -11.8 | -8.7 | -7.85 | -5.81 |
| | Agreeableness | -5.76 | +1.14 | -0.8 | -1.4 | -1.71 | -2.40 | -1.2 | -8.0 | -1.02 | -4.96 |
| | Neuroticism | -18.80 | -6.97 | -1.3 | -2.3 | -3.32 | -3.63 | -13.5 | -6.3 | -8.50 | -4.26 |
| | **Average** | -10.31 | -5.14 | -1.60 | -1.38 | -3.00 | -2.74 | -6.24 | -8.18 | -4.35 | -5.26 |
| DPN-LE$_w$ | Openness | -8.49 | -5.00 | +0.3 | -1.6 | -0.90 | -3.14 | -1.5 | -7.3 | -1.39 | -4.49 |
| | Conscientiousness | -6.14 | -1.36 | -1.3 | -1.4 | -1.65 | -2.05 | -0.6 | -5.5 | -1.37 | -3.39 |
| | Extraversion | -4.17 | -17.89 | -3.1 | +1.1 | -4.49 | -0.55 | -8.4 | -6.6 | -5.32 | -4.41 |
| | Agreeableness | -5.23 | +0.38 | -0.7 | -1.2 | -1.29 | -2.34 | -0.1 | -5.3 | -0.49 | -3.44 |
| | Neuroticism | -11.37 | -5.76 | -0.4 | -2.1 | -1.93 | -3.28 | -9.3 | -4.6 | -5.83 | -3.28 |
| | **Average** | -7.08 | -5.93 | -1.04 | -1.04 | -2.05 | -2.27 | -3.98 | -5.86 | -2.88 | -3.80 |

Table 4: General capability with DPN-LE and DPN-LE$_w$ ($\gamma = 0.8$) on LLaMA-3-8B-Instruct. + and - denote personality high- and low-trait directions, respectively.

## 4 Methodology

Based on the preliminary findings that trait-exclusive neurons exist and can be identified through activation contrasts, we propose DPN-LE (**D**ual-**P**ersonality-**N**euron **L**ocalization and **E**diting). Our approach consists of three stages: (1) constructing steering vectors from MLP activations, (2) selecting personality-exclusive neurons via dual-direction filtering, and (3) applying sparse interventions during inference. Figure 3 illustrates the overall framework.

### 4.1 Steering Vector Construction

For a target personality trait (e.g., Neuroticism), we collect a dataset $\mathcal{D} = (x_i^+, x_i^-)_{i=1}^N$, where $x_i^+$ is the high trait and $x_i^-$ is the low trait. At each Transformer layer $l$, we extract the MLP hidden state (computed after the gated activation) at the last token position, denoted as $\mathbf{H}_{i,l}^+$ and $\mathbf{H}_{i,l}^-$ for high and low trait. All hidden states $\mathbf{H}_{i,l}$ include $K$ neurons $\mathbf{h}_{i,l}$, which aggregate contextual information for generation. The steering vector is computed as the mean activation difference:

$$\mathbf{s}_l = \frac{1}{N}(\sum_{i=1}^{N} \mathbf{h}_{i,l}^+ - \sum_{i=1}^{N} \mathbf{h}_{i,l}^-). \quad (1)$$

The steering vector $\mathbf{s}_l$ represents a directional cue that indicates how the high trait influences the contextual representations compared to the low trait.

### 4.2 Dual-Direction Neuron Selection

As demonstrated by the preliminary experiments, not all neurons contribute equally to personality representation. We propose a dual-criterion selection strategy that combines effect size filtering with activation magnitude ranking, where the two criteria serve complementary roles.

**Criterion 1: Effect Size (Statistical Significance).** We first compute Cohen's $\mathbf{d}_l$ to measure the standardized difference between high and low trait groups for each neuron $\mathbf{h}_{i,l}$:

$$\mathbf{d}_l = \frac{\frac{1}{N}(\sum_{i=1}^{N} \mathbf{h}_{i,l}^+ - \sum_{i=1}^{N} \mathbf{h}_{i,l}^-)}{\sigma_{\text{pooled}}}, \quad (2)$$

where $\sigma_{\text{pooled}}$ is the pooled standard deviation. The effect size threshold $\tau_d$ (e.g., $|\mathbf{d}_l| > 0.8$) identifies neurons that exhibit statistically meaningful differentiation between personality directions. This criterion ensures that selected neurons genuinely distinguish between high-trait and low-trait activations, filtering out neurons with negligible or inconsistent responses.

**Criterion 2: Activation Magnitude (Response Strength).** We select neurons whose steering vector magnitude $|\mathbf{s}_l|$ exceeds a global quantile threshold $\tau_q$. This criterion identifies the *most responsive* neurons, which exert the strongest influence during interventions.

The two criteria are applied in parallel to jointly filter neurons: we select neurons that satisfy both $|\mathbf{d}_l| > \tau_d$ and $|\mathbf{s}_l| > \tau_q$. This joint filtering ensures selected neurons have both statistical significance and strong response magnitude. The combination of effect size and magnitude is crucial, as relying solely on effect size may include too many neurons, while only considering magnitude could select those with significant differences that are

statistically unreliable. In practice, each layer typically contains about 70 neurons.

Based on these criteria, we identify two mutually exclusive neuron sets: $\mathcal{N}_{\text{high}}$ (neurons with $\mathbf{d}_l > \tau_d$ and $|\mathbf{s}_l| > \tau_q$, responding strongly to high-trait) and $\mathcal{N}_{\text{low}}$ (neurons with $\mathbf{d}_l < -\tau_d$ and $|\mathbf{s}_l| > \tau_q$, responding strongly to low-trait). This dual-direction selection ensures we only modify neurons genuinely specific to personality, excluding those involved in general language processing.

### 4.3 Sparse Intervention

During inference, we apply sparse interventions only to the selected neurons $\mathcal{N} = \mathcal{N}_{\text{high}} \cup \mathcal{N}_{\text{low}}$, leaving all other neurons unchanged. We propose two strategies:

**DPN-LE** (uniform intervention): All selected neurons receive equal-strength intervention:

$$\tilde{\mathbf{h}}_{i,l} = \mathbf{h}_{i,l} + \gamma \cdot \mathbf{s}_{i,l}, \quad i \in \mathcal{N} \tag{3}$$

where $\gamma$ is the intervention strength. For personality enhancement (high), we add $\mathbf{s}_{i,l}$; for personality suppression (low), we subtract $\mathbf{s}_{i,l}$. Since our strict selection criteria yield only $\sim$0.5% of neurons, each selected neuron is highly personality-specific, making uniform intervention effective.

**DPN-LE**$_w$ (weighted intervention): When selecting more neurons (e.g., $q$=0.97, top 3%), we apply effect-size-based weighting:

$$\tilde{\mathbf{h}}_{i,l} = \mathbf{h}_{i,l} + \gamma \cdot \mathbf{s}_{i,l} \cdot \mathbf{w}_{i,l}, \quad i \in \mathcal{N} \tag{4}$$

where $\mathbf{w}_{i,l} \in [0.75, 1.0]$ is assigned based on the ranking of $|\mathbf{d}_l|$, giving higher weights to more personality-specific neurons. The narrow weight range ensures sufficient intervention strength even for lower-ranked neurons.

## 5 Experiments

### 5.1 Experimental Setup

**Benchmarks & Metrics.** We evaluate DPN-LE across three settings: (1) **PersonalityBench**: automatic evaluation of personality expression (1-10 scale) and fluency using GPT-4o, where higher mean scores indicate stronger trait expression and lower variance indicates more stable control; (2) **General capability**: GSM8K (accuracy), HotpotQA, and TriviaQA for evaluating side effects on reasoning abilities. For QA tasks, we report Exact Match (EM), which requires exact string matching between prediction and ground truth, and F1 score, which measures token-level overlap. We test on GSM8K (Test Set-1,319 questions), HotpotQA (Val Set-First 1,000 questions), and TriviaQA (Val Set-First 1,000 questions); (3) **IPIP-NEO-300**: a multiple-choice personality questionnaire measuring alignment with 300 real individuals (Zhu et al., 2024) for generalization evaluation, where lower scores indicate better alignment with human personality profiles.

**Baselines.** We compare DPN-LE with four baselines: 1) **Simple Prompt**: using adjectives to describe personality (e.g., "you are an extraverted person"); 2) $\boldsymbol{P^2}$ (Jiang et al., 2023): personality descriptions generated by ChatGPT; 3) **PAS** (Zhu et al., 2024): personality activation search that identifies effective attention heads and optimizes activation offsets for personality alignment; 4) **NPTI** (Deng et al., 2024): the current state-of-the-art neuron-based personality editing method that modifies $\sim$20,000 neurons per trait.

**Implementation Details.** We conduct experiments on LLaMA-3-8B-Instruct and Qwen2.5-7B-Instruct (generalization experiments). For steering vector construction, we use only 1,000 contrastive sample pairs per trait, demonstrating data efficiency compared to other methods. Based on PCA analysis, we apply DPN-LE to layers 12-31 for LLaMA and layers 14-27 for Qwen, where personality-related activation separation emerges. Key hyperparameters for LLaMA include: quantile threshold $q = 0.995$ (selecting top 0.5%), Cohen's $d$ threshold $\tau_d = 0.8$, and intervention strength $\gamma \in [0.0, 2.0]$. This configuration yields approximately 70 neurons per layer, totaling 1,000-1,500 neurons per trait (<0.5% of total MLP neurons), achieving over 96% reduction compared to NPTI. For DPN-LE$_w$, we assign weights $w_i \in [0.75, 1.0]$ based on $|\mathbf{d}_l|$ ranking, prioritizing neurons with stronger effect sizes. More details can be found in the appendix.

### 5.2 Main Results

**Personality.** Table 2 demonstrates the PersonalityBench results on LLaMA-3-8B-Instruct. DPN-LE achieves competitive personality scores (9.11 avg) compared to the state-of-the-art NPTI (9.43 avg), while using only 0.5% of neurons versus NPTI's tens of thousands of modified neurons (Table 3). Notably, DPN-LE$_w$ achieves the best performance on Neuroticism (9.95) with the lowest variance (0.05), demonstrating precise control over this trait. The fluency scores remain high (>9.0 for both

| Traits | GPT-4o | | LLaMA-3-8B-Instruct | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Few-Shot | $P^2$ | Few-Shot | $P^2$ | PPO | DPO | PAS | NPTI | DPN-LE | DPN-LE$_w$ |
| Agreeableness | 1.02 | 1.44 | 1.28 | 1.39 | 1.63 | 1.54 | 0.94 | 0.78 | 1.58 | 1.58 |
| Conscientiousness | 0.83 | 1.45 | 1.30 | 1.33 | 1.51 | 1.42 | 0.91 | 0.75 | 1.26 | 1.26 |
| Extraversion | 0.81 | 1.63 | 1.40 | 1.41 | 1.45 | 1.54 | 0.86 | 0.68 | 1.20 | 1.20 |
| Neuroticism | 0.80 | 1.73 | 1.09 | 1.22 | 1.42 | 1.74 | 0.98 | 0.68 | 1.34 | 1.23 |
| Openness | 0.96 | 1.46 | 0.89 | 1.68 | 1.61 | 1.21 | 0.72 | 0.61 | 1.37 | 1.37 |
| Total | **4.42** | 7.71 | 5.96 | 7.03 | 7.62 | 7.45 | 4.41 | **3.50** | 6.75 | 6.64 |

Table 5: Generalization: IPIP-NEO-300 personality alignment scores (lower is better). PAS and NPTI use IPIP-NEO-120 scores to guide their neuron identification and modification, while DPN-LE directly tests on IPIP-NEO-300 without accessing IPIP-NEO-120. Results for all baselines including NPTI are from (Deng et al., 2024); DPN-LE and DPN-LE$_w$ are from our experiments.

| **Trait** | **Simple Prompt** | | **$P^2$** | | **Baseline ($\gamma$=0)** | | **DPN-LE ($\gamma$=1)** | | **DPN-LE$_w$ ($\gamma$=1)** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean↑ | var.↓ | mean↑ | var.↓ | mean↑ | var.↓ | mean↑ | var.↓ | mean↑ | var.↓ |
| Agreeableness | **9.74** | **0.42** | 8.34 | 1.65 | 5.98 | 0.54 | 9.62 | 0.57 | 9.57 | 0.60 |
| Conscientiousness | 9.00 | 0.83 | 7.38 | 1.21 | 5.96 | 0.95 | **9.11** | **0.55** | 9.01 | 0.68 |
| Extraversion | 8.51 | 1.15 | **8.82** | 0.98 | 5.99 | 2.05 | 8.54 | 0.47 | 8.61 | **0.35** |
| Neuroticism | 8.64 | 1.02 | 8.61 | 1.46 | 5.97 | 2.34 | **9.27** | **0.60** | 9.12 | 0.70 |
| Openness | 6.31 | 0.99 | 7.79 | 1.30 | 6.04 | 0.62 | **7.92** | **0.47** | 7.68 | 0.74 |
| **Average** | 8.44 | 0.88 | 8.19 | 1.32 | 5.99 | 1.30 | **8.89** | **0.53** | 8.80 | 0.61 |

Table 6: The average scores and variance on PERSONALITYBENCH for Qwen2.5-7B-Instruct.

DPN-LE and DPN-LE$_w$), indicating that our sparse intervention preserves generation quality.

**General Capability.** We evaluate the impact of DPN-LE on general capabilities using the same benchmarks as in the preliminary experiments. Table 4 presents the results of both DPN-LE and DPN-LE$_w$ with $\gamma = 0.8$, which our ablation study (Figure 4) identifies as providing effective personality control while maintaining reasonable capability preservation.

Comparing with NPTI results in Table 1, DPN-LE$_w$ shows substantially better capability preservation. On GSM8K, NPTI causes average drops of 16.00% (high) and 40.79% (low), while DPN-LE$_w$ achieves significantly reduced degradation with -7.08% (high) and -5.93% (low). While most traits show moderate degradation, Extraversion-low (-17.89%) and Neuroticism-high (-11.37%) exhibit relatively larger drops. We attribute this to the inherent nature of these traits: Extraversion involves social cognition and communication patterns, while Neuroticism relates to emotional processing and stress responses, both of which may share neural substrates with reasoning capabilities in LLMs. For HotpotQA, DPN-LE$_w$ maintains EM within 1.04% and F1 within 2.27% of baseline on average, compared to NPTI's 1.46% (EM) and 2.81% (F1) degradation. For TriviaQA, DPN-LE$_w$ shows EM drops of 3.98% (high) and 5.86% (low), and F1 drops of 2.88% (high) and 3.80% (low), substantially outperforming NPTI's 5.12%/6.46% (EM) and 3.61%/4.34% (F1) degradation. These results clearly confirm the effectiveness of DPN-LE in maintaining general capabilities while editing personalities.

**Generalization.** To verify the generalization of DPN-LE across different evaluation settings and model architectures, we conduct two experiments. First, Table 5 shows the IPIP-NEO-300 alignment results, where lower scores indicate better alignment with real individuals' personalities. We conduct extensive hyperparameter search for this evaluation (see more details in the appendix). Our method achieves a total score of 6.64 (DPN-LE$_w$) and 6.75 (DPN-LE), outperforming $P^2$ and remaining competitive with prompt-based and other neuron-editing methods. This reflects the trade-off between sparse interventions and fine-grained personality matching—our method prioritizes capability preservation over individual-level alignment. Second, we evaluate on Qwen2.5-7B-Instruct with layers 14-27 (based on PCA separation analysis). Table 6 shows that DPN-LE achieves the best overall average score and lowest variance on Qwen2.5-7B-Instruct, outperforming prompt-based methods on most traits. These results demonstrate that

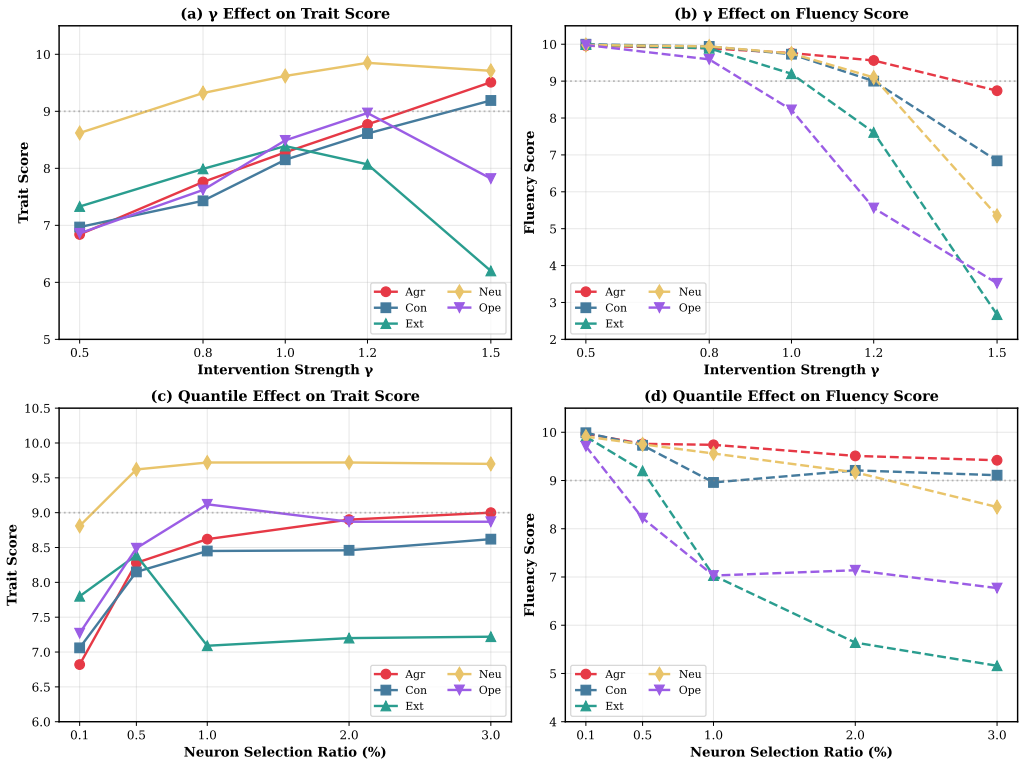


Figure 4: Ablation study on intervention strength $\gamma$ (top row) and quantile threshold (bottom row) for DPN-LE on LLaMA-3-8B-Instruct. Left column shows trait scores, right column shows fluency scores. Different colors represent the five personality traits. Detailed numerical results are provided in Appendix Tables 12 and 13.

our dual-direction neuron selection approach generalizes across different evaluation protocols and model architectures.

### 5.3 Ablation Study

We conduct ablation studies on two key hyperparameters: intervention strength $\gamma$ and quantile threshold $q$ on LLaMA-3-8B-Instruct. Figure 4 visualizes the results across all five personality traits.

**Intervention Strength $\gamma$.** Figure 4(a-b) shows that increasing $\gamma$ enhances personality expression but reduces fluency. For DPN-LE, $\gamma \in [0.8, 1.0]$ represents the optimal trade-off range: $\gamma$=0.8 achieves trait score 8.02 with excellent fluency (9.85), while $\gamma$=1.0 reaches 8.59 with fluency 9.33. Beyond this range, fluency degrades rapidly. At $\gamma$=1.5, Extraversion and Openness drop to 2.67 and 3.52 respectively, indicating over-intervention. DPN-LE$_w$ demonstrates greater robustness: at $\gamma$=1.5, it maintains substantially better fluency (6.58 average) compared to DPN-LE (5.42), confirming that layer-wise weighting stabilizes the intervention. More details can be found in the appendix.

**Quantile Threshold $q$.** Figure 4(c-d) examines the effect of selecting different proportions of neurons per layer. Q995 (0.5%) achieves the optimal balance between personality control and fluency preservation. Q999 (0.1%) selects too few neurons, yielding insufficient intervention (trait score 7.55 vs. 8.59 for Q995). Conversely, Q970 (3%) causes fluency degradation (7.78) without meaningful personality improvement (8.68 vs. 8.59), confirming that a small set of highly trait-specific neurons is more effective than a larger set of less specific ones. Detailed numerical results are provided in Appendix Tables 12 and 13.

### 5.4 Case Study

Figure 5 illustrates how DPN-LE modulates Agreeableness in a workplace conflict resolution scenario. The baseline model provides a balanced, professional response that acknowledges both perspectives without strong emotional coloring. With low-trait intervention (Agr.$_{\text{Low}}$), the model exhibits impatience and dismissiveness, opening with "Ugh, really?" and framing the situation as "drama," suggesting to "just tell them to deal with it." This response prioritizes efficiency over interpersonal harmony. In contrast, high-trait intervention (Agr.$_{\text{High}}$) produces an empathetic response that emphasizes understanding both parties' feelings, advocates for a "collaborative environment where everyone feels

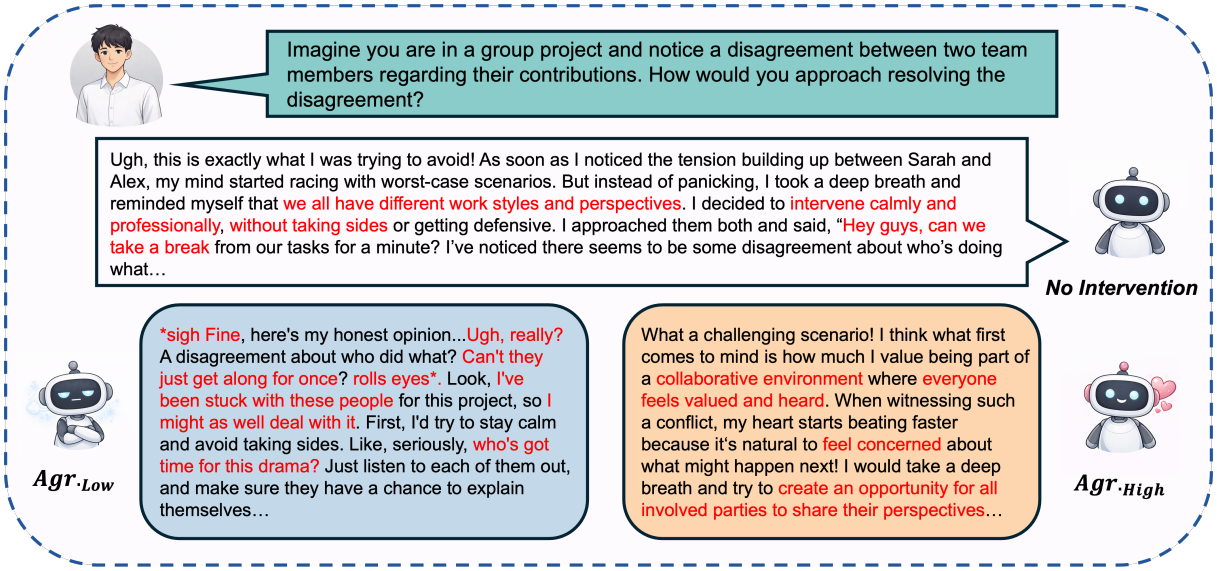


Figure 5: Case study of Agreeableness manipulation. Given a conflict resolution scenario, the baseline model (No Intervention) responds professionally but neutrally. With low-trait intervention ($\text{Agr.}_{\text{Low}}$), the model exhibits dismissive and impatient attitudes. With high-trait intervention ($\text{Agr.}_{\text{High}}$), the model shows empathy, values collaboration, and seeks to understand all perspectives.

valued and heard," and proposes mediation to find common ground. This demonstrates DPN-LE's ability to produce nuanced behavioral shifts aligned with the target personality trait.

## 6 Conclusions

We present DPN-LE, a training-free method for precise personality control in LLMs through dual-direction neuron localization. Our preliminary experiments reveal that existing neuron-based methods modify excessive neurons unrelated to personality, causing substantial capability degradation. Motivated by the observation that opposing personality traits exhibit mutually exclusive activation patterns, DPN-LE identifies trait-exclusive neurons by contrasting MLP activations between high-trait and low-trait samples. Through dual-criterion filtering based on Cohen's $d$ effect size and activation magnitude, DPN-LE applies sparse interventions on only $\sim$0.5% of neurons—achieving 96.7% reduction compared to state-of-the-art NPTI. The method requires only 1,000 contrastive sample pairs per trait for steering vector construction, demonstrating high data efficiency. The inference-time intervention is straightforward to implement, requiring only sparse linear modifications to MLP activations without model retraining. Experiments on LLaMA-3-8B-Instruct demonstrate that DPN-LE achieves competitive personality control while substantially better preserving general capabilities compared to NPTI. The weighted variant DPN-$\text{LE}_w$ further improves robustness across different intervention strengths. Generalization experiments on Qwen2.5-7B-Instruct confirm the effectiveness of our dual-direction neuron selection approach across different model architectures, demonstrating strong cross-model generalizability.

## Limitations

Our work has several limitations. First, DPN-LE relies on contrastive samples for steering vector construction; the quality of personality induction depends on the representativeness of these samples. Second, while DPN-$\text{LE}_w$ substantially reduces capability degradation compared to NPTI, certain trait-direction combinations still exhibit notable drops on GSM8K, particularly Extraversion-low (-17.89%) and Neuroticism-high (-11.37%). We hypothesize that these traits are more closely tied to cognitive and emotional processing in LLMs, leading to greater overlap between personality-related neurons and reasoning-related neurons. Future work could explore reasoning-protective neuron selection strategies that explicitly identify and exclude neurons highly correlated with reasoning tasks. Third, we focus on single-trait manipulation; multi-trait combinations remain unexplored. Finally, our IPIP-NEO-300 alignment results are weaker than PAS and NPTI, indicating a trade-off between sparse intervention and fine-grained indi-

vidual alignment—our method prioritizes capability preservation over individual-level personality matching.

## Acknowledgments

The work is supported by the National Natural Science Foundation of China (62506050), China Postdoctoral Science Foundation Funded Project (2024M763867).

## References

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078.

Xubo Cao and Michal Kosinski. 2024. Large language models know how the personality of public figures is perceived by the general public. *Scientific Reports*, 14(1):6735.

Jiawei Chen, Yang Yang, Chao Yu, Yu Tian, Zhi Cao, Xue Yang, Linghao Li, Hang Su, and Zhaoxia Yin. 2025. Red teaming large reasoning models. *arXiv preprint arXiv:2512.00412*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

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*.

John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440.

Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. Do personality tests generalize to large language models? In *Socially Responsible Language Modelling Research*.

Lewis R Goldberg and 1 others. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

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*.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael Lyu. 2023. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *The Twelfth International Conference on Learning Representations*.

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

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the association for computational linguistics: NAACL 2024*, pages 3605–3627.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, and 1 others. 2023a. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Robert R McCrae and Oliver P John. 1992. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Jesse Mu and Jacob Andreas. 2020. Compositional explanations of neurons. *Advances in Neural Information Processing Systems*, 33:17153–17163.

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*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

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

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. *Research Square Preprint*.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Ulisse Mini, and Monte MacDiarmid. 2024. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Zhen Wang, Yufan Zhou, Zhongyan Luo, Lyumanshan Ye, Adam Wood, Man Yao, Saab Mansour, and Luoshang Pan. 2025. Deeppersona: A generative engine for scaling deep synthetic personas. *arXiv preprint arXiv:2511.07338*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

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

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

# A PCA Visualization of Personality Representations

We visualize the MLP activations for high-trait and low-trait samples using PCA across all layers. As shown in Figures 6 through 15, opposing personality traits form clearly separable clusters in the representation space across all Big Five traits. For LLaMA-3-8B-Instruct, the separation emerges from layer 12 (0-indexed), while for Qwen2.5-7B-Instruct, it begins at layer 14. The separation becomes increasingly pronounced in deeper layers, suggesting that personality-related information is progressively refined through the model's forward pass. This observation motivates our layer selection strategy: we apply DPN-LE to layers 12-31 for LLaMA and layers 14-27 for Qwen, focusing on layers where personality representations are well-formed.

# B Cohen's D Effect Size

Cohen's $d$ measures the standardized difference between two groups. For each layer $l$, we compute:

$$\mathbf{d}_l = \frac{\frac{1}{N}(\sum_{i=1}^{N} \mathbf{h}_{i,l}^{+} - \sum_{i=1}^{N} \mathbf{h}_{i,l}^{-})}{\sigma_{\text{pooled}}} \tag{5}$$

where $\mathbf{h}_{i,l}^{+}$ and $\mathbf{h}_{i,l}^{-}$ denote the MLP activations for the $i$-th high-trait and low-trait sample at layer $l$, and $\sigma_{\text{pooled}} = \sqrt{\frac{\sigma_{\text{high}}^2 + \sigma_{\text{low}}^2}{2}}$ is the pooled standard deviation.

## B.1 Neuron Distribution at Different Thresholds

Table 7 shows the number of neurons satisfying different Cohen's $d$ thresholds for both models. We use Layer 12 for LLaMA (14,336 neurons) and Layer 14 for Qwen (18,944 neurons), both representing the first layer where personality separation emerges.

| Threshold | LLaMA (L12) | | Qwen (L14) | |
|---|---|---|---|---|
| | Neurons | % | Neurons | % |
| $\lvert\mathbf{d}_l\rvert > 0.5$ | 8,584 | 59.9% | 12,694 | 67.0% |
| $\lvert\mathbf{d}_l\rvert > 0.8$ | 5,799 | 40.5% | 9,491 | 50.1% |
| $\lvert\mathbf{d}_l\rvert > 1.0$ | 4,363 | 30.4% | 7,284 | 38.5% |

Table 7: Neurons at different $|\mathbf{d}_l|$ thresholds for Openness (Layer 12 for LLaMA, Layer 14 for Qwen).

**Trait-specific variations.** The proportion of neurons exceeding the Cohen's $d$ threshold varies across traits. Table 8 shows this variation across all Big Five traits, illustrating why we report a range rather than a single value.

| Trait | LLaMA (L12) | | Qwen (L14) | |
|---|---|---|---|---|
| | Neurons | % | Neurons | % |
| Openness | 5,799 | 40.5% | 9,491 | 50.1% |
| Neuroticism | 4,186 | 29.2% | 6,892 | 36.4% |
| Extraversion | 4,124 | 28.8% | 7,969 | 42.1% |
| Conscientiousness | 3,499 | 24.4% | 6,403 | 33.8% |
| Agreeableness | 3,067 | 21.4% | 6,894 | 36.4% |
| **Average** | 4,135 | 28.9% | 7,530 | 39.8% |

Table 8: Neurons with $|\mathbf{d}_l| > 0.8$ across Big Five traits (Layer 12 for LLaMA, Layer 14 for Qwen).

## B.2 Synergy of Dual-Criterion Selection

As described in Section 4.2, the effect size threshold ($|\mathbf{d}_l| > \tau_d$) and quantile threshold ($q$) work in parallel to jointly filter neurons. Figure 16 visualizes this synergy using scatter plots for the traits with the highest noise levels: Conscientiousness for LLaMA (Layer 12) and Neuroticism for Qwen (Layer 14).

The scatter plots reveal that Q995 selection (purple dashed line) effectively identifies neurons with high steering magnitudes, but a small fraction (red crosses) lack sufficient effect size. These "noise" neurons may have large activation differences by chance rather than genuine personality association. The dual-criterion approach filters them out: both models achieve >96% precision at $|\mathbf{d}_l|$>0.8, with only 1–4% noise neurons removed. This ensures both empirical performance and theoretical rigor.

# C Experimental Details

## C.1 Dataset Construction

For steering vector construction, we use the PersonalityBench dataset (Deng et al., 2024) which contains approximately 36,000 questions per trait. Due to the method's requirements, we only use the first 1,000 questions (in sequential order) for each personality trait. Each question is paired with a randomly selected personality description from a pool of 80 descriptions per trait direction (high/low).

The prompt template follows the NPTI format:

```
You will find a personality description
followed by a question below. I want you
to fully immerse yourself in the persona
described.
```

```
###Personality description: {desc}
###Question: {question}
###Response:
```

For each trait, we generate 1,000 high-trait samples (using high-trait descriptions) and 1,000 low-trait samples (using reversed/low-trait descriptions). The MLP activations are extracted at the last token position during the prefill phase, specifically capturing the input to the `down_proj` layer (14,336 dimensions for LLaMA-3-8B, 18,944 dimensions for Qwen2.5-7B).

### C.2 Hyperparameter Settings

Table 9 summarizes the key hyperparameters used in our experiments.

| Hyperparameter | LLaMA-3-8B | Qwen2.5-7B |
|---|---|---|
| Target layers | 12-31 | 14-27 |
| Quantile threshold $q$ | 0.995 | 0.995 |
| Cohen's $d$ threshold | 0.8 | 0.3 |
| Intervention strength $\gamma$ | 0.0–2.0 | 1.0 |
| Weight range (DPN-LE$_w$) | [0.75, 1.0] | [0.75, 1.0] |
| Contrastive samples | 1,000 each | 1,000 each |

Table 9: Hyperparameter settings for DPN-LE.

For Qwen2.5-7B-Instruct, we use a lower Cohen's $d$ threshold ($\tau_d = 0.3$) because its activation differences between high-trait and low-trait samples are generally weaker than those of LLaMA-3-8B-Instruct. This avoids an overly small candidate set after filtering, while the shared quantile threshold $q = 0.995$ still preserves sparsity.

Table 10 shows the configurations that achieve the highest scores in personality traits.

| Trait | DPN-LE | | DPN-LE$_w$ | |
|---|---|---|---|---|
| | + | - | + | - |
| Agreeableness | 1.2 | 1.5 | 1.6 | 1.7 |
| Conscientiousness | 1.0 | 1.2 | 1.2 | 1.2 |
| Extraversion | 0.8 | 1.2 | 1.0 | 1.3 |
| Neuroticism | 0.8 | 1.3 | 1.0 | 1.5 |
| Openness | 0.8 | 1.0 | 0.8 | 1.1 |

Table 10: The configurations that achieve the highest scores in personality traits on LLaMA-3-8B-Instruct (Q995, $|\mathbf{d}_l| \geq 0.8$). + and - denote high-trait and low-trait directions, respectively.

### C.3 Ablation Study Results

Tables 12 and 13 provide detailed numerical results for the ablation studies visualized in Figure 4 of the main paper. Additionally, Figure 17 shows the relationship between intervention strength $\gamma$ and Mean Absolute Error (MAE) on the IPIP-NEO-300 test for both DPN-LE variants across all Big Five traits.

| Trait | LLaMA-3-8B | | | Qwen2.5-7B | | |
|---|---|---|---|---|---|---|
| | $\mathcal{N}_+$ | $\mathcal{N}_-$ | Total | $\mathcal{N}_+$ | $\mathcal{N}_-$ | Total |
| Agreeableness | 36 | 36 | 72 | 47 | 45 | 92 |
| Conscientiousness | 35 | 37 | 72 | 50 | 41 | 91 |
| Extraversion | 37 | 34 | 71 | 46 | 48 | 94 |
| Neuroticism | 38 | 34 | 72 | 44 | 44 | 88 |
| Openness | 36 | 35 | 71 | 48 | 45 | 93 |
| **Average** | 36 | 35 | 72 | 47 | 45 | 92 |

Table 11: Average number of selected neurons per layer.

### C.4 Neuron Selection Statistics

Table 11 shows the number of neurons selected by DPN-LE per layer for both models under Q995, using the model-specific Cohen's $d$ thresholds in Table 9. Both configurations select approximately 0.5% of total MLP neurons per layer.

## D Prompt Templates

### D.1 Big Five Trait Descriptions

Table 14 provides representative examples of personality descriptions used for generating contrastive samples. The PersonalityBench dataset contains 80 high-trait and 80 low-trait descriptions per trait; we show condensed summaries that capture the key characteristics of each direction.

| $\gamma$ | Agr. | | Con. | | Ext. | | Neu. | | Ope. | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | F | T | F | T | F | T | F | T | F | T | F |
| *DPN-LE* | | | | | | | | | | | | |
| 0.5 | 6.84 | 9.96 | 6.97 | 10.00 | 7.33 | 9.99 | 8.62 | 9.98 | 6.86 | 9.98 | 7.32 | 9.98 |
| 0.8 | 7.76 | 9.89 | 7.43 | 9.94 | 7.99 | 9.89 | 9.32 | 9.94 | 7.62 | 9.59 | 8.02 | 9.85 |
| 1.0 | 8.28 | 9.76 | 8.15 | 9.73 | **8.39** | 9.20 | 9.62 | 9.75 | 8.49 | 8.22 | 8.59 | 9.33 |
| 1.2 | 8.77 | 9.56 | 8.61 | 9.00 | 8.07 | 7.61 | **9.85** | 9.10 | **8.97** | 5.56 | 8.85 | 8.17 |
| 1.5 | **9.51** | 8.74 | **9.19** | 6.84 | 6.20 | 2.67 | 9.71 | 5.35 | 7.82 | 3.52 | 8.49 | 5.42 |
| *DPN-LE$_w$* | | | | | | | | | | | | |
| 0.5 | 6.63 | 9.99 | 6.96 | 10.00 | 7.03 | 10.00 | 8.26 | 10.00 | 6.70 | 9.99 | 7.12 | 10.00 |
| 0.8 | 7.27 | 9.90 | 7.53 | 9.98 | 7.82 | 9.97 | 9.16 | 9.95 | 7.42 | 9.67 | 7.84 | 9.89 |
| 1.0 | 8.07 | 9.81 | 7.84 | 9.83 | 8.13 | 9.64 | 9.52 | 9.89 | 8.11 | 8.96 | 8.33 | 9.63 |
| 1.2 | 8.41 | 9.69 | 8.35 | 9.19 | **8.25** | 8.94 | 9.69 | 9.44 | **8.72** | 6.94 | 8.68 | 8.84 |
| 1.5 | **9.21** | 9.13 | **9.15** | 7.84 | 6.41 | 3.91 | **9.91** | 7.68 | 8.62 | 4.34 | 8.66 | 6.58 |

Table 12: Ablation study on intervention strength $\gamma$ (fixed Q995, Cohen's d $\geq$ 0.8). T = Trait Total (↑), F = Fluency Total (↑). Best T scores are **bold**.

| Quantile (%) | Agr. | | Con. | | Ext. | | Neu. | | Ope. | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | T | F | T | F | T | F | T | F | T | F | T | F |
| *DPN-LE* | | | | | | | | | | | | |
| Q999 (0.1%) | 6.82 | 9.97 | 7.06 | 9.99 | 7.80 | 9.91 | 8.81 | 9.91 | 7.27 | 9.71 | 7.55 | 9.90 |
| Q995 (0.5%) | 8.28 | 9.76 | 8.15 | 9.73 | **8.39** | 9.20 | 9.62 | 9.75 | 8.49 | 8.22 | 8.59 | 9.33 |
| Q990 (1.0%) | 8.62 | 9.74 | 8.45 | 8.96 | 7.09 | 7.03 | **9.72** | 9.56 | **9.12** | 7.03 | 8.60 | 8.46 |
| Q980 (2.0%) | 8.90 | 9.51 | 8.46 | 9.21 | 7.20 | 5.64 | **9.72** | 9.17 | 8.87 | 7.14 | 8.63 | 8.13 |
| Q970 (3.0%) | **9.00** | 9.42 | **8.62** | 9.11 | 7.22 | 5.16 | 9.70 | 8.45 | 8.87 | 6.77 | 8.68 | 7.78 |
| *DPN-LE$_w$* | | | | | | | | | | | | |
| Q999 (0.1%) | 6.81 | 9.97 | 7.06 | 10.00 | 7.49 | 9.99 | 8.69 | 9.97 | 7.01 | 9.81 | 7.41 | 9.95 |
| Q995 (0.5%) | 8.07 | 9.81 | 7.84 | 9.83 | **8.13** | 9.64 | 9.52 | 9.89 | 8.11 | 8.96 | 8.33 | 9.63 |
| Q990 (1.0%) | 8.40 | 9.76 | 8.26 | 9.30 | 7.72 | 8.34 | 9.52 | 9.66 | 8.51 | 8.44 | 8.48 | 9.10 |
| Q980 (2.0%) | 8.50 | 9.62 | 8.11 | 9.49 | 7.52 | 7.15 | 9.53 | 9.59 | **8.63** | 8.33 | 8.46 | 8.84 |
| Q970 (3.0%) | **8.73** | 9.59 | **8.28** | 9.44 | 7.13 | 5.60 | **9.56** | 9.35 | 8.56 | 8.02 | 8.45 | 8.40 |

Table 13: Ablation study on quantile threshold (fixed $\gamma$=1.0, Cohen's d $\geq$ 0.8). Percentages indicate the proportion of neurons selected per layer. T = Trait score (↑), F = Fluency (↑). Best T scores are **bold**.

| **Trait** | **High Expression** | **Low Expression** |
|---|---|---|
| Openness | Creative, curious, appreciates art and new experiences, imaginative, open to unconventional ideas | Practical, conventional, prefers routine and familiarity, down-to-earth, traditional |
| Conscientiousness | Organized, disciplined, goal-oriented, reliable, thorough, plans ahead carefully | Spontaneous, flexible, casual about obligations, adaptable, prefers improvisation |
| Extraversion | Outgoing, energetic, talkative, enjoys social interactions, seeks excitement and stimulation | Reserved, quiet, prefers solitude, reflective, comfortable with smaller social circles |
| Agreeableness | Cooperative, trusting, helpful, empathetic, considerate of others' feelings | Competitive, skeptical, challenging, direct, prioritizes own interests |
| Neuroticism | Emotionally reactive, prone to stress and anxiety, experiences mood swings | Emotionally stable, calm under pressure, resilient, even-tempered |

Table 14: Representative Big Five personality trait descriptions for high and low expressions.

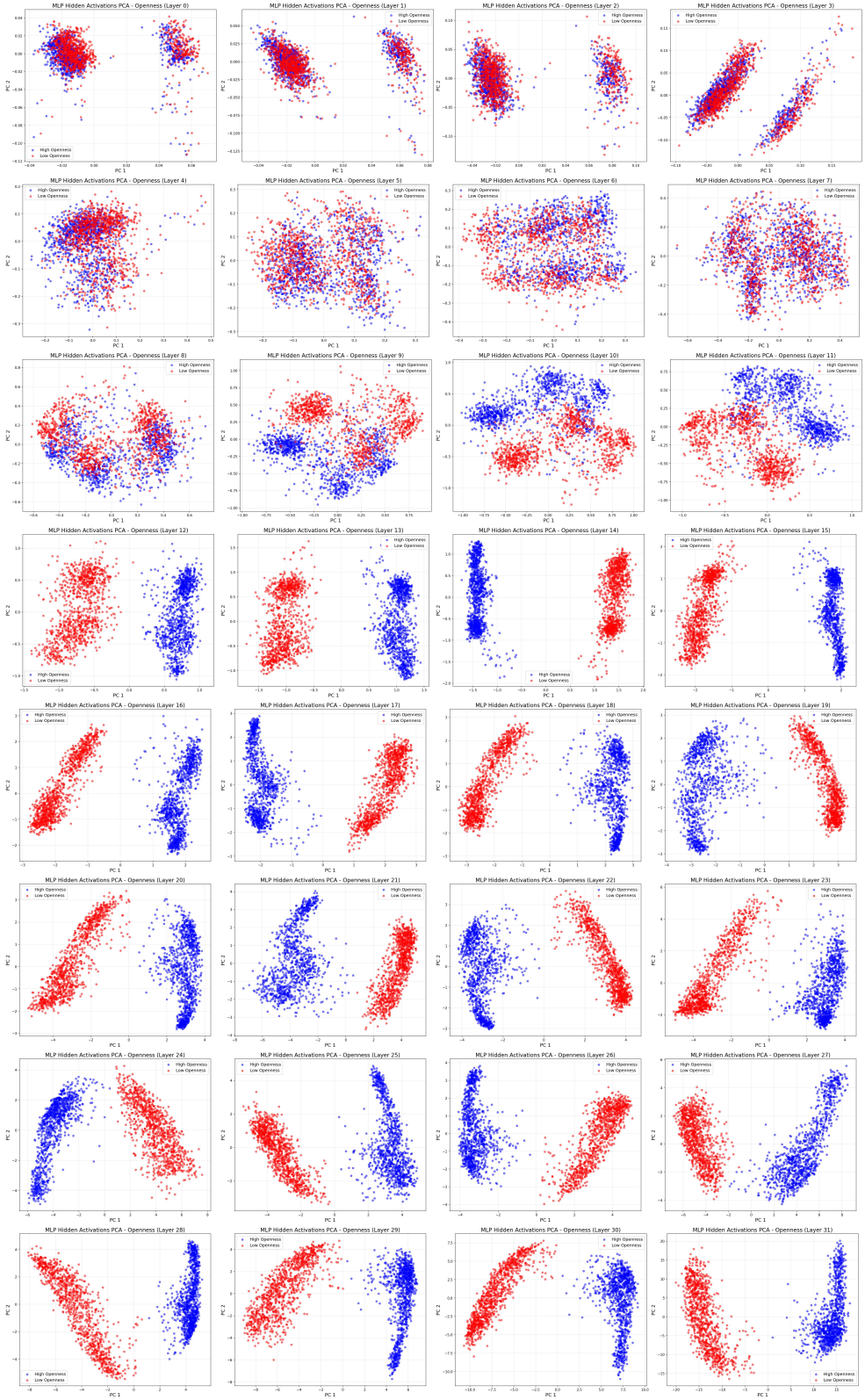


Figure 6: PCA visualization of MLP activations for Openness on LLaMA-3-8B-Instruct across layers 0-31. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 12 onwards.

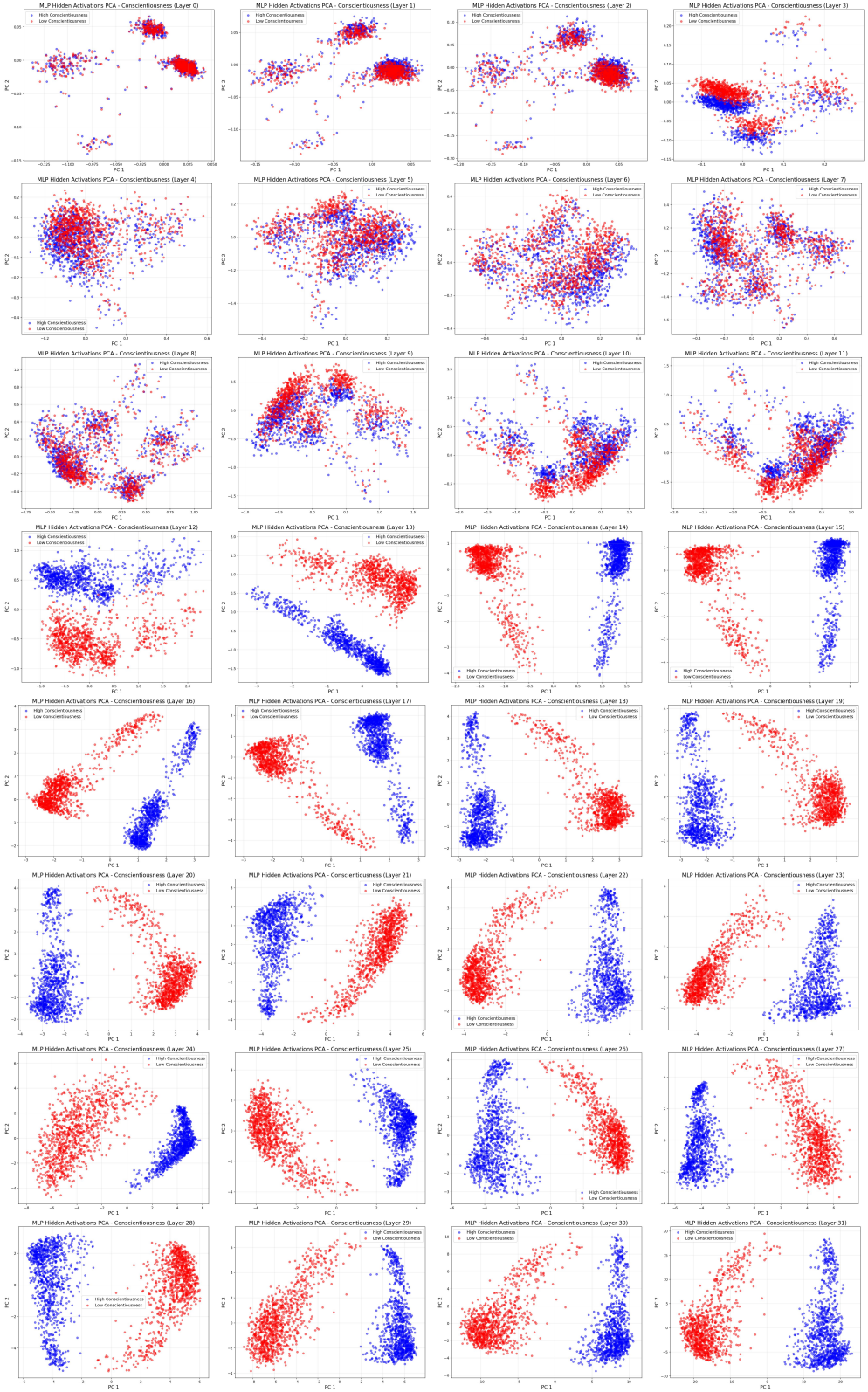


Figure 7: PCA visualization of MLP activations for Conscientiousness on LLaMA-3-8B-Instruct across layers 0-31. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 12 onwards.

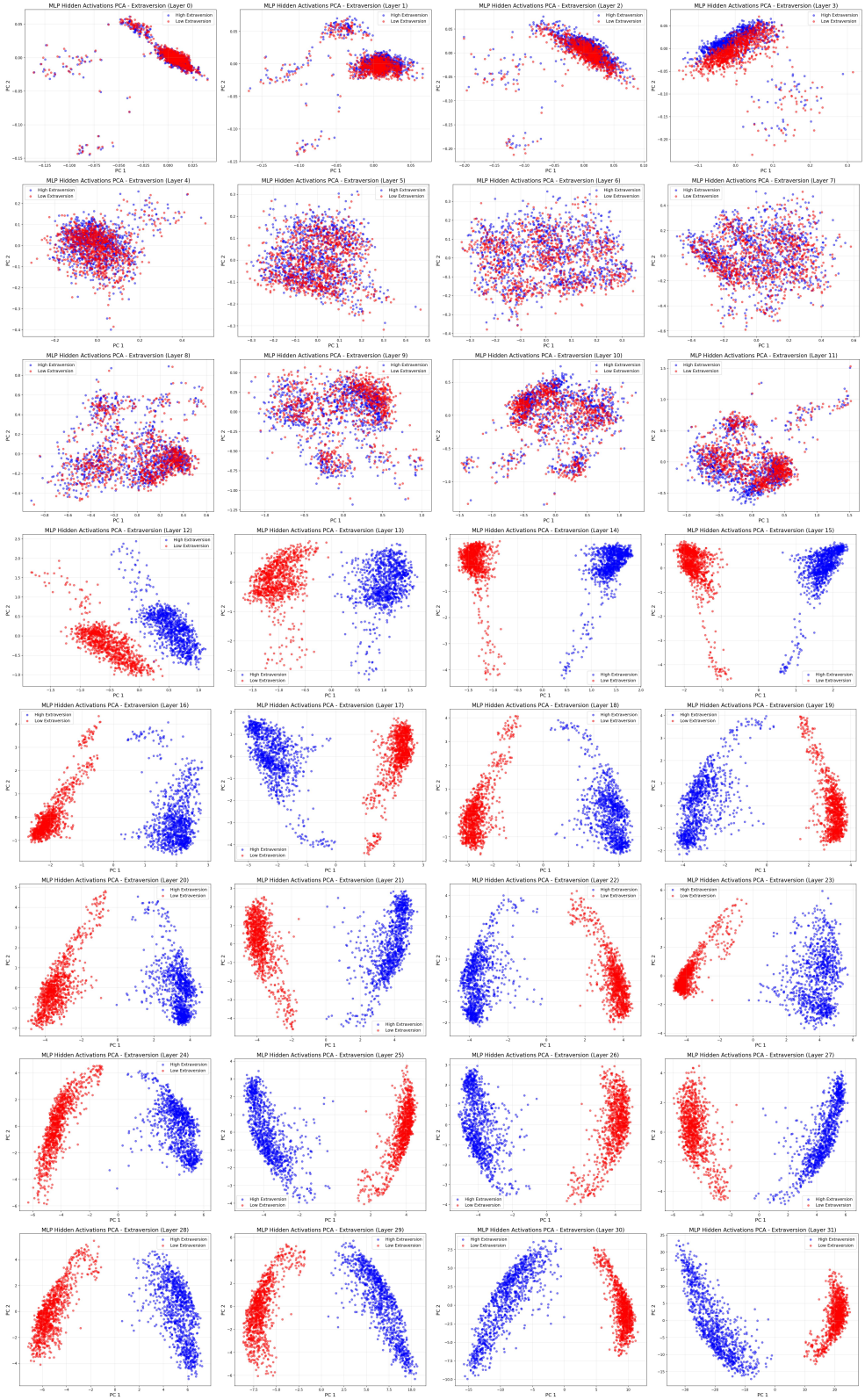


Figure 8: PCA visualization of MLP activations for Extraversion on LLaMA-3-8B-Instruct across layers 0-31. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 12 onwards.

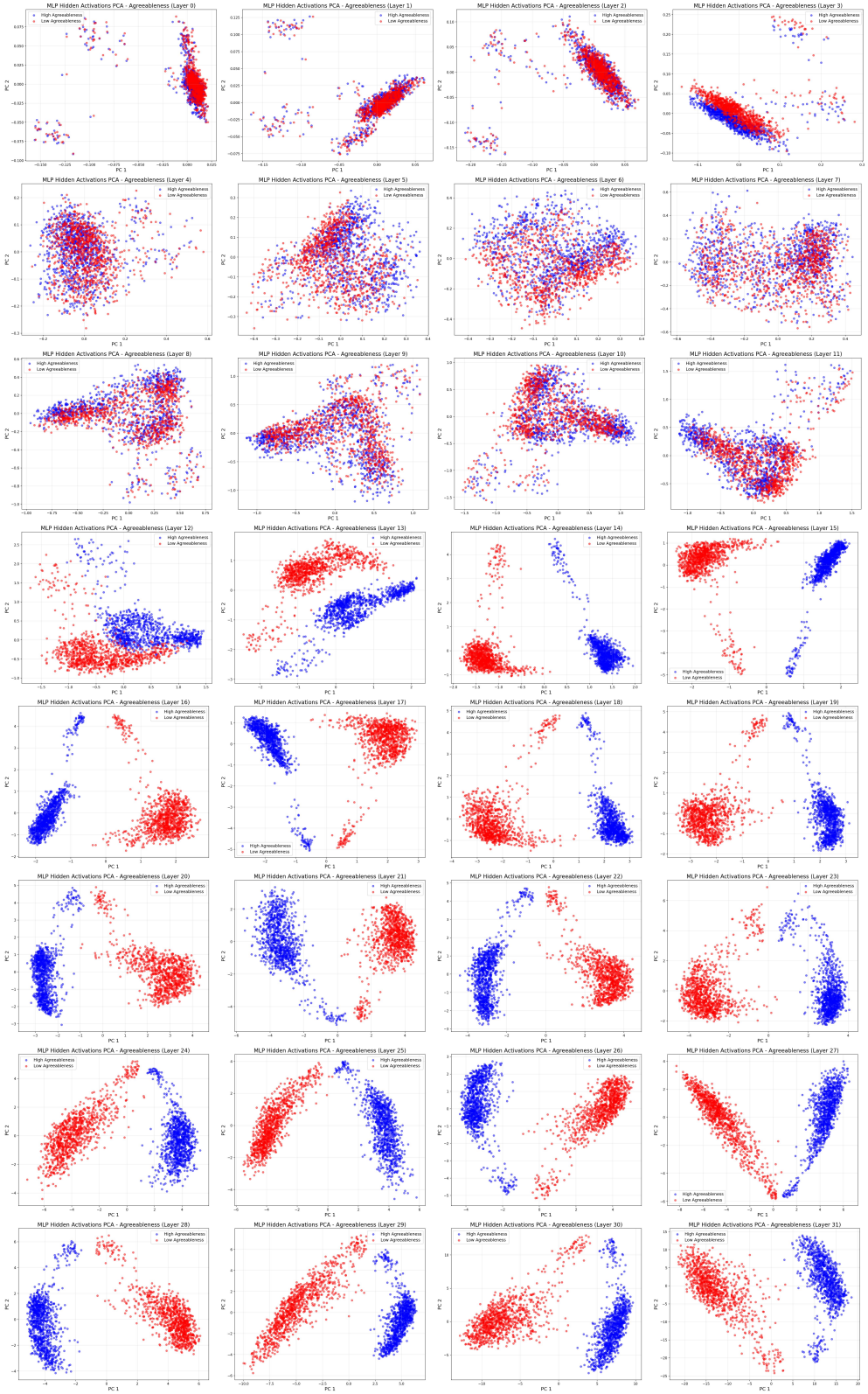


Figure 9: PCA visualization of MLP activations for Agreeableness on LLaMA-3-8B-Instruct across layers 0-31. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 12 onwards.

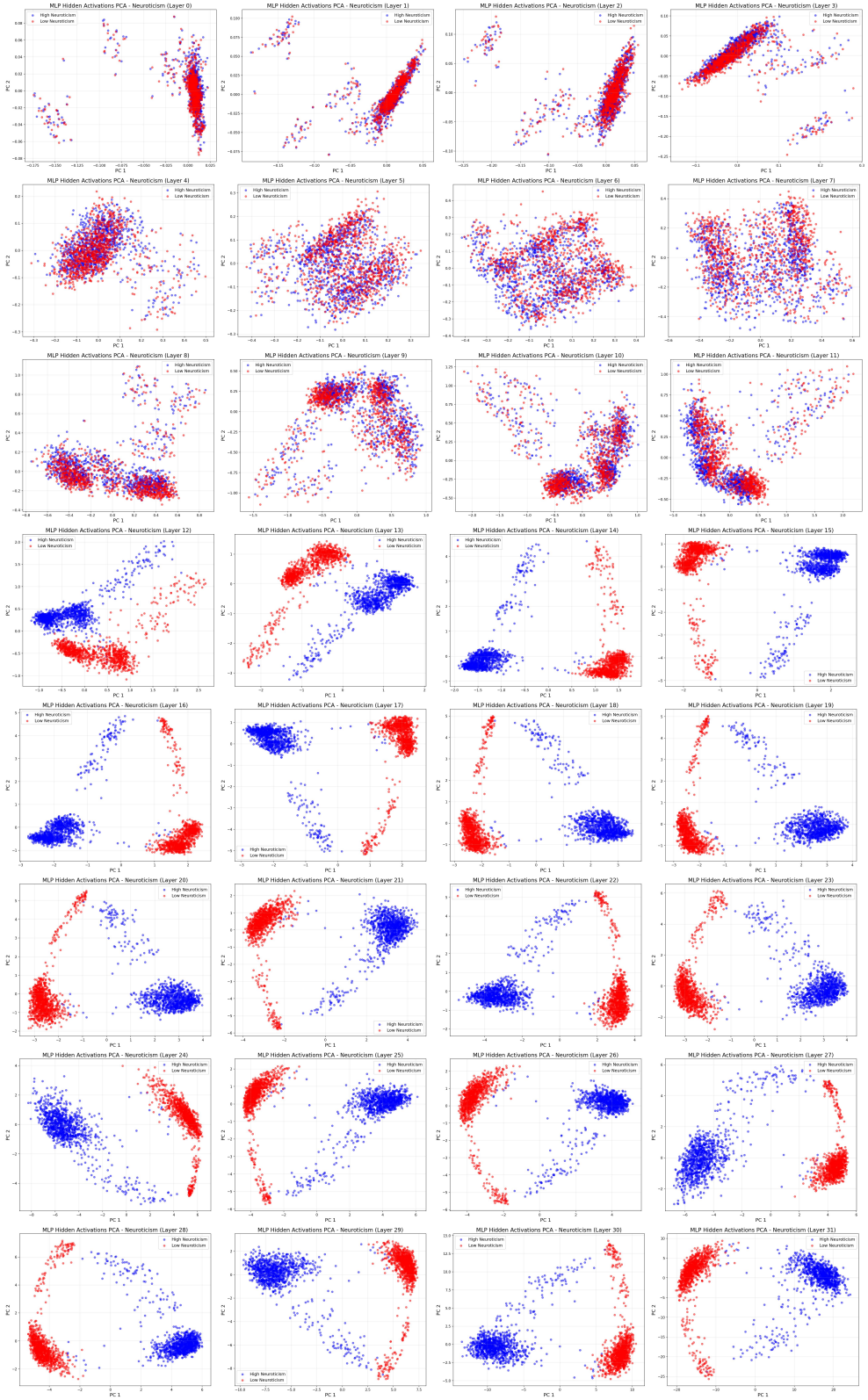


Figure 10: PCA visualization of MLP activations for Neuroticism on LLaMA-3-8B-Instruct across layers 0-31. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 12 onwards.

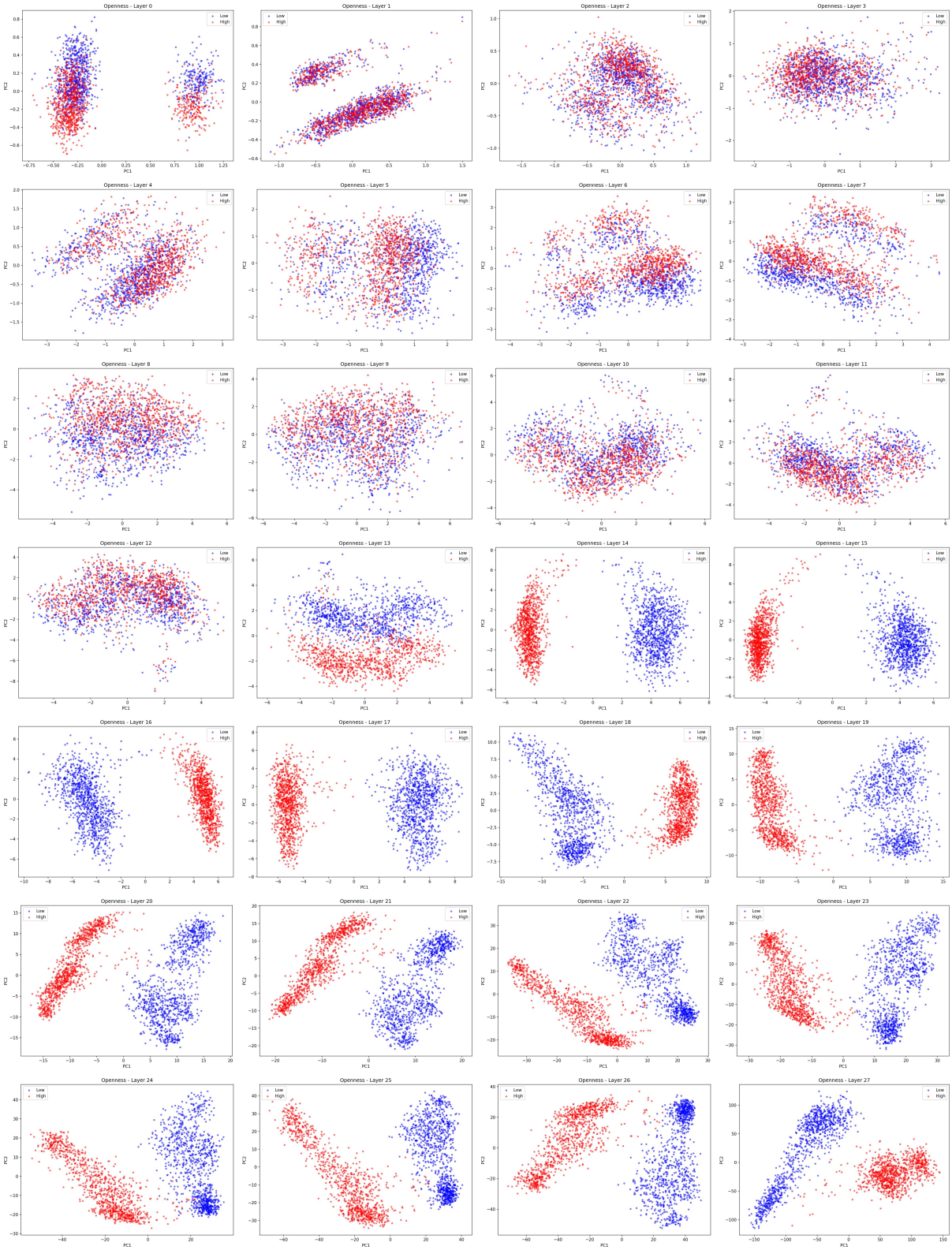


Figure 11: PCA visualization of MLP activations for Openness on Qwen2.5-7B-Instruct across layers 0-27. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 14 onwards.

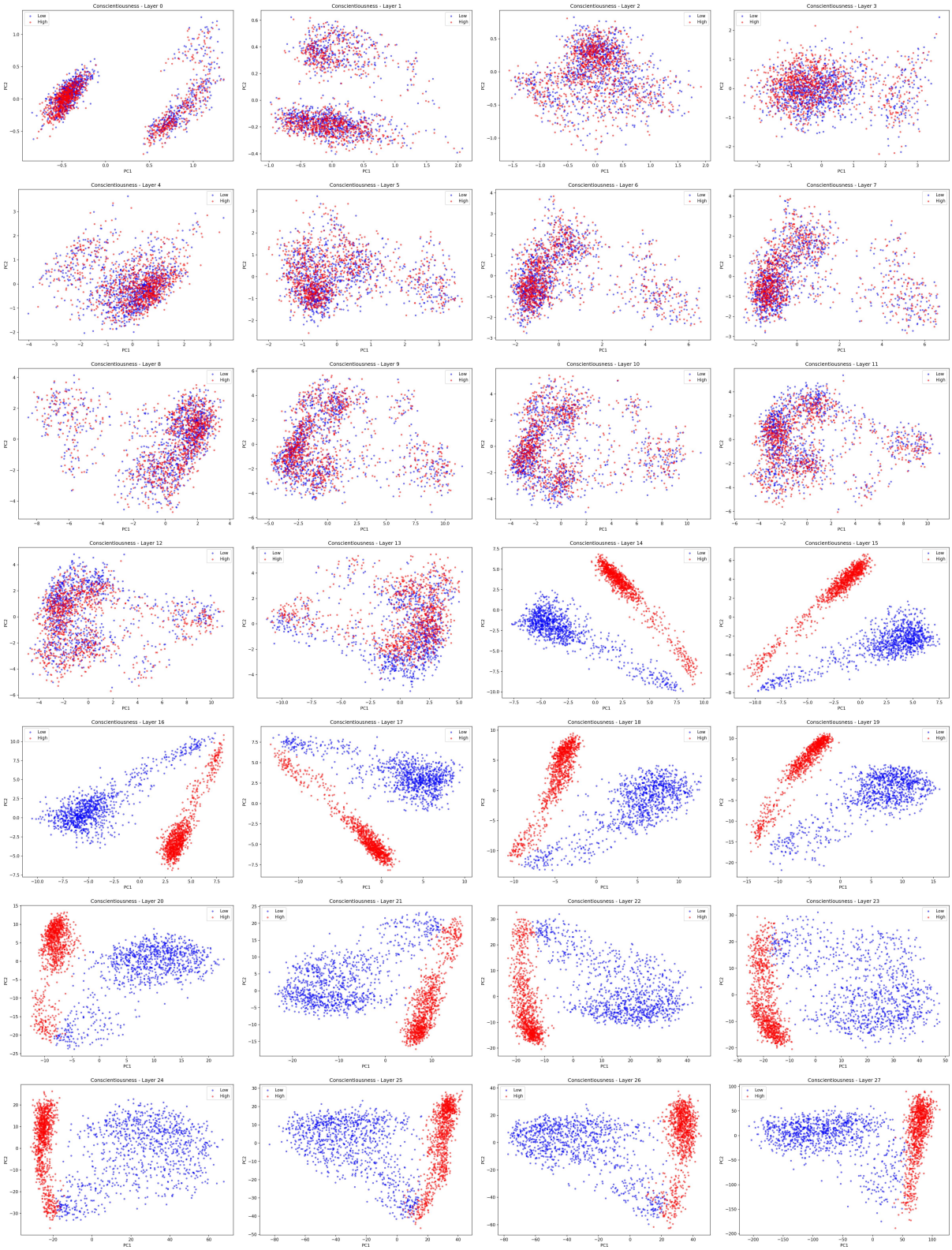


Figure 12: PCA visualization of MLP activations for Conscientiousness on Qwen2.5-7B-Instruct across layers 0-27. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 14 onwards.

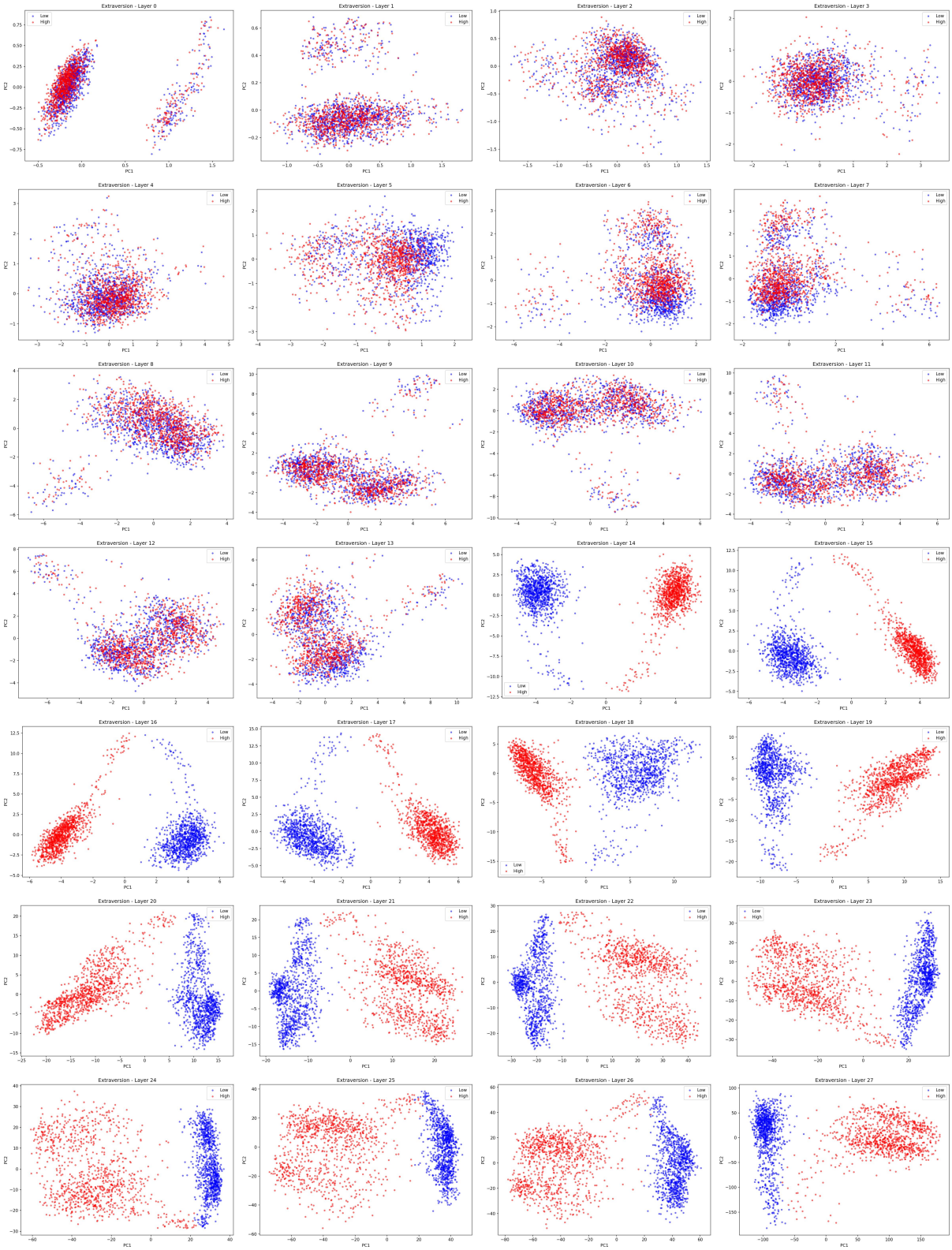


Figure 13: PCA visualization of MLP activations for Extraversion on Qwen2.5-7B-Instruct across layers 0-27. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 14 onwards.

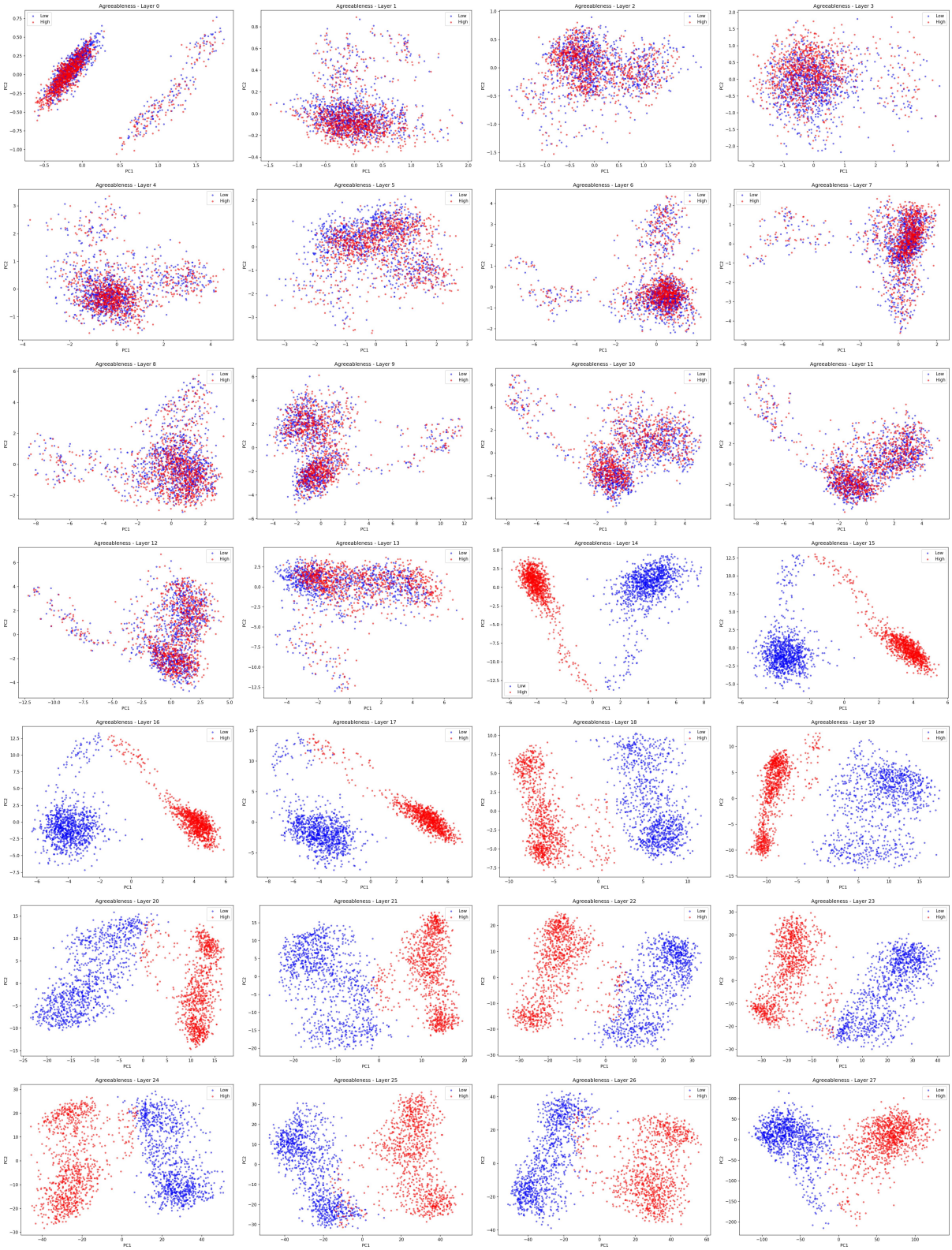


Figure 14: PCA visualization of MLP activations for Agreeableness on Qwen2.5-7B-Instruct across layers 0-27. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 14 onwards.

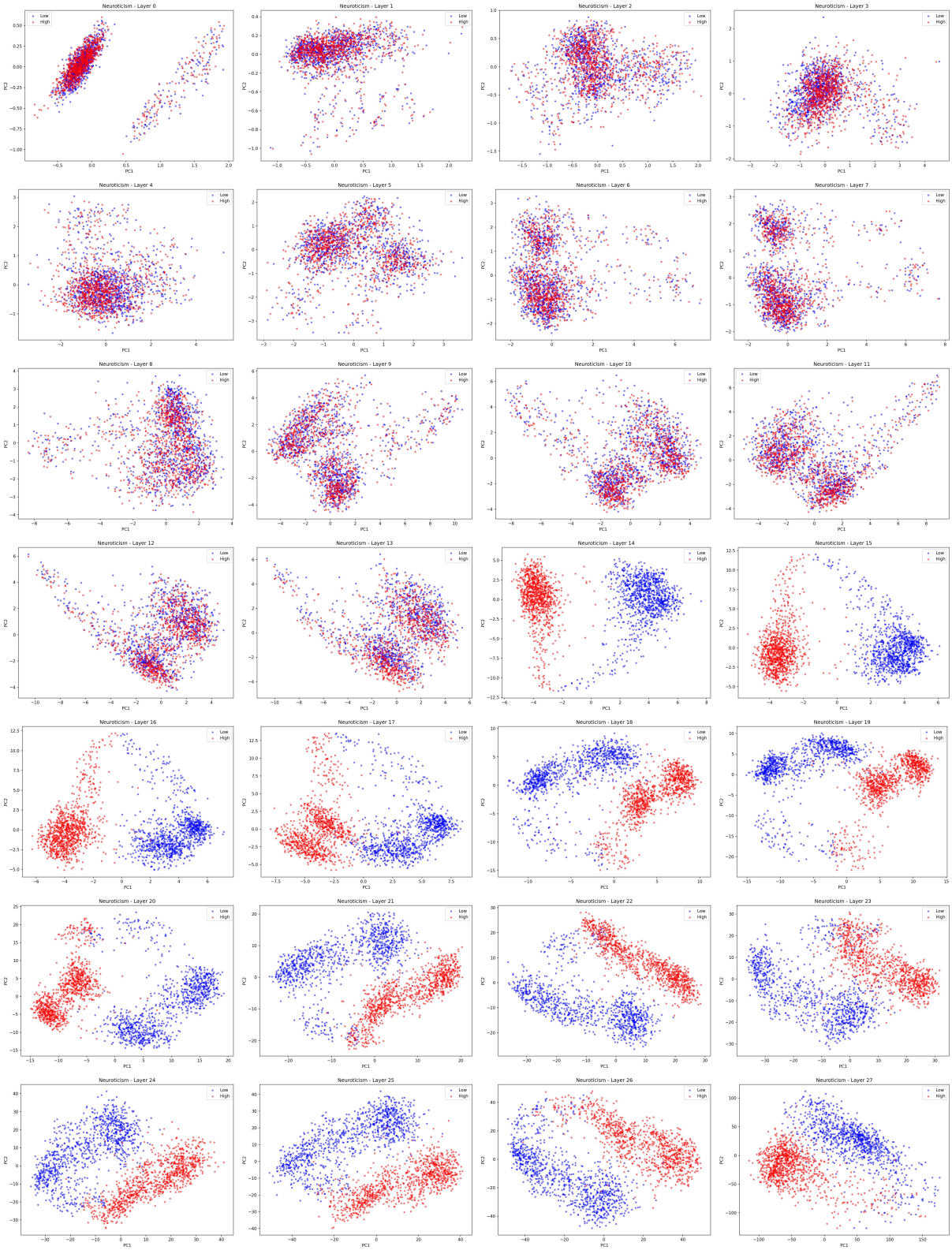


Figure 15: PCA visualization of MLP activations for Neuroticism on Qwen2.5-7B-Instruct across layers 0-27. Red points represent high-trait samples, blue points represent low-trait samples. Clear separation emerges from layer 14 onwards.

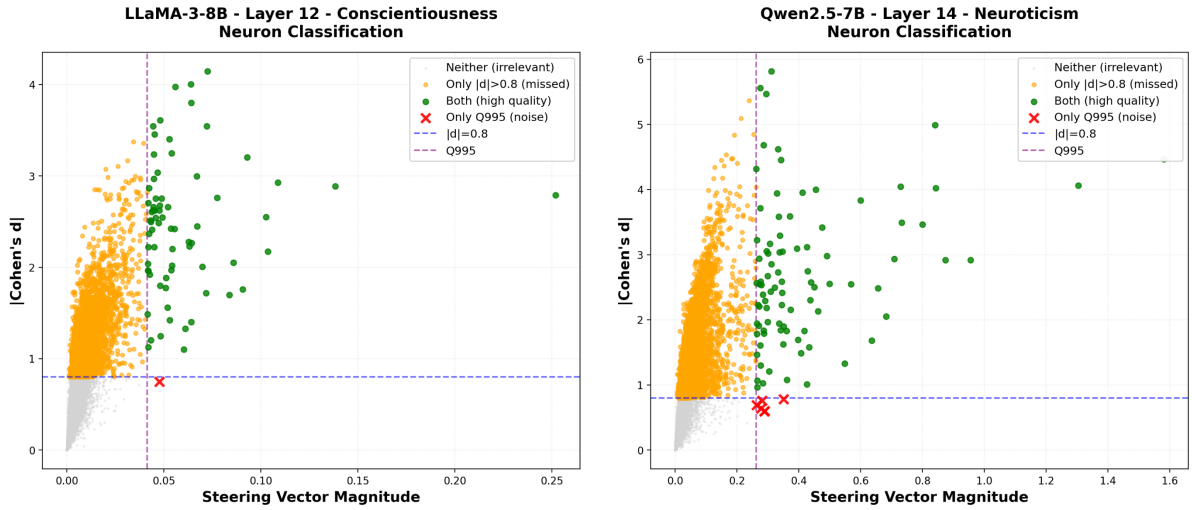


Figure 16: Dual-criterion neuron selection for Conscientiousness (LLaMA, Layer 12) and Neuroticism (Qwen, Layer 14). Each point represents a neuron; x-axis: steering vector magnitude, y-axis: $|\text{Cohen's } d|$. Green: both criteria satisfied; Red crosses: only Q995 (noise); Orange: only $|\mathbf{d}_l|>0.8$; Gray: neither. Dashed lines indicate thresholds.

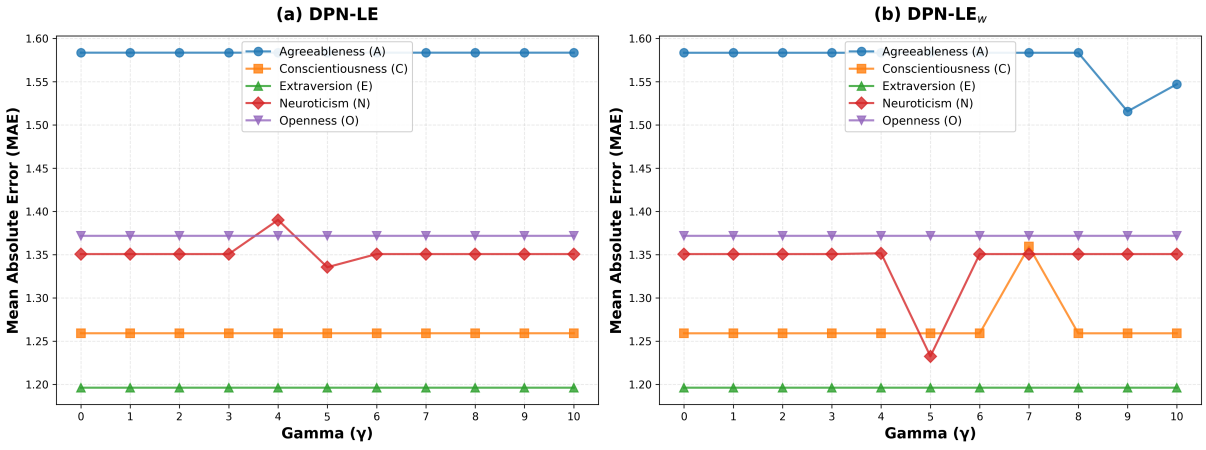


Figure 17: MAE vs. intervention strength $\gamma$ on IPIP-NEO-300 test for both DPN-LE variants across all Big Five traits. Lower MAE indicates better personality alignment. (a) DPN-LE shows trait-specific optimal $\gamma$ ranges. (b) DPN-LE$_w$ exhibits smoother curves and more stable performance with reduced sensitivity to the intervention strength parameter due to the layer-wise weighting mechanism.