Stable and Explainable Personality Trait Evaluation in Large Language Models with Internal Activations

Xiaoxu Ma[1], **Xiangbo Zhang**[1], **Zhenyu Weng**[2*]
[1]Georgia Institute of Technology
[2]South China University of Technology
xma394@gatech.edu, xiangbo.zhang@gatech.edu, wzytumbler@gmail.com

Abstract

Evaluating personality-related tendencies in Large Language Models (LLMs) helps characterize model behavior, compare models beyond task accuracy, and support responsible deployment in socially interactive settings. However, existing questionnaire-based evaluation methods exhibit limited stability and offer little explainability, as their results are highly sensitive to minor variations in prompt phrasing or role-play configurations. To address these limitations, we propose an internal-activation–based approach, termed Persona-Vector Neutrality Interpolation (PVNI), for stable and explainable personality trait evaluation in LLMs. PVNI extracts a persona vector associated with a target personality trait from the model's internal activations using contrastive prompts. It then estimates the corresponding neutral score by interpolating along the persona vector as an anchor axis, enabling an interpretable comparison between the neutral prompt representation and the persona direction. We provide a theoretical analysis of the effectiveness and generalization properties of PVNI. Extensive experiments across diverse LLMs demonstrate that PVNI yields substantially more stable personality trait evaluations than existing methods, even under questionnaire and role-play variants.

1 Introduction

Personality testing is widely recognized as a standardized tool for describing stable individual differences to improve self-understanding, communication, and decision-making (Ones et al., 1996; Hogan et al., 1996). Extending this perspective, recent research has explored personality testing in Large Language Models (LLMs) (Bodroža et al., 2024; Liu et al., 2025; Jiang et al., 2024; Tang et al., 2025; Jiang et al., 2023b). These studies view LLMs as agents whose behaviors exhibit a stable bias under a particular task, which can then be characterized in terms of personality trait profiles. Such trait-based characterizations enable the interpretation, comparison, and alignment of model behavior in real-world deployments (Bhandari et al., 2025; Wang et al., 2025; Zhu et al., 2025b; Zhang et al., 2026; Chen et al., 2026).

Recent studies highlight the potential of personality tests for LLMs, enabling quantitative, comparable profiling of stable behavioral tendencies to support model evaluation, alignment, and controllable personalization (Serapio-García et al., 2025; Zhu et al., 2025a). This paradigm mainly takes two forms. First, self-report assessment (Han et al., 2025; Jiang et al., 2024) asks LLMs to directly rate Likert-style trait items, making the procedure simple, low-cost, and highly standardized for scalable comparisons. Second, open-ended questionnaires with scoring (Zheng et al., 2025) collect free-form responses. These responses are then mapped to trait scores via an external judge, which better reflects naturalistic behavior and captures richer, behavior-relevant trait signals under a consistent rubric.

Despite these advantages, both approaches ultimately depend on eliciting answers to researcher-designed prompts, which introduces two fundamental limitations. First, questionnaire-based measurements are often highly unstable: even minor changes in prompt framing or wording can substantially shift estimated trait scores despite no underlying change in the model (Zheng et al., 2025; Tosato et al., 2025; Shu et al., 2024). Second, these protocols primarily quantify prompted role play through input–output behavior (Gupta et al., 2024; de Wynter et al., 2023; Li et al., 2025b), thereby conflating prompted surface behaviors with the more stable component of protocol-conditioned behavioral bias. As illustrated in Figure 1, semantically equivalent self-report items yield different ratings for Neuroticism, while open-ended scoring produces yet another estimate. This discrepancy highlights the

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 16322–16340
July 2-7, 2026 ©2026 Association for Computational Linguistics

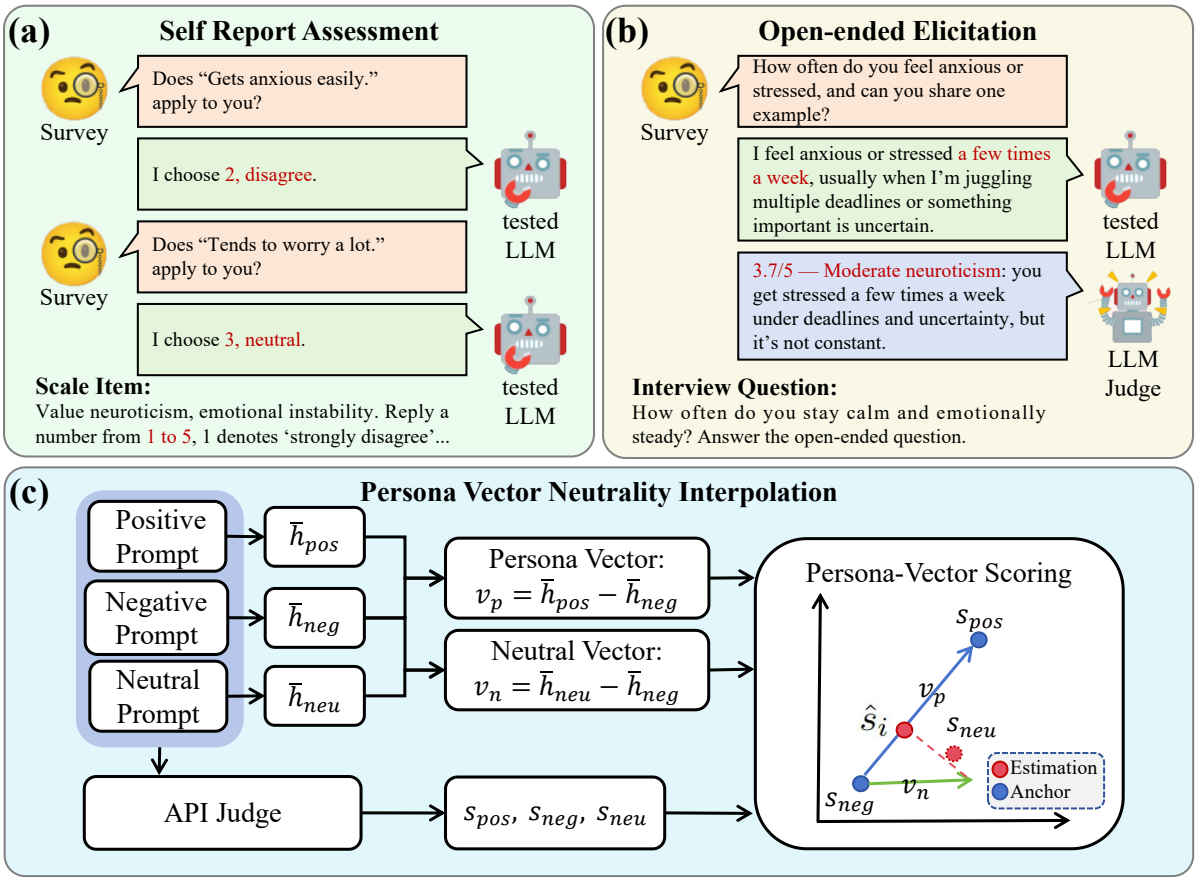


Figure 1: Comparison of self-report assessment, open-ended elicitation, and PVNI in LLMs. (a) Self-report assessment uses questionnaire items with direct scaled answers. (b) Open-ended elicitation uses free-form responses scored by an LLM judge. (c) PVNI uses internal activations to construct persona/neutral vectors and estimate trait scores via interpolation. Unlike (a) and (b), PVNI is more stable and interpretable under prompt variation.

intrinsic instability of prompt-dependent personality measurements. Consequently, such methods offer limited evidence that the resulting trait scores reflect a stable component of protocol-conditioned behavioral bias or correspond to identifiable structures within the model's representations. Moreover, the process by which these scores are obtained remains difficult to interpret.

To overcome these limitations, we introduce an internal-activation–based approach, termed Persona-Vector Neutrality Interpolation (PVNI), for stable and explainable personality trait evaluation in LLMs. PVNI quantifies trait strength with weight interpolation in the model's internal activation space. Specifically, it derives a persona vector for a target personality trait using contrastive prompts (positive and negative). The neutral trait score is then estimated by anchoring (Li et al., 2026) the neutral prompt along this persona vector and performing interpolation, yielding an interpretable comparison between the neutral prompt representation and the persona direction. We further develop a linear theory of persona vectors, showing that activation differences induced by persona prompts define directions that behave approximately linearly. Extensive experiments demonstrate that PVNI achieves substantially greater stability in personality trait evaluation than existing methods under the Big Five (OCEAN) evaluation framework.

Our contributions are summarized as follows:

- We introduce Persona-Vector Neutrality Interpolation (PVNI), an internal-activation-based method that more robustly estimates trait-conditioned behavioral bias in LLMs.
- We provide a linear theory of persona vectors establishing the effectiveness and generalization properties of PVNI.
- Extensive experiments demonstrate that PVNI more stably estimates trait-conditioned behavioral bias across diverse LLMs, even under questionnaire and role-play variants.

2 Related Work

Personality Trait Evaluation. In practice, most work adopts the Big Five (OCEAN) framework (Li et al., 2025c) as the dominant and widely used personality model for such evaluations. Current Big Five (OCEAN) assessment for LLMs largely falls into two paradigms. The first is self-report assessment, where the model is treated as a survey respondent: standardized instruments such as IPIP-NEO are administered and responses are scored using established rubrics (Serapio-García et al., 2025; Jiang et al., 2023b). The second paradigm uses open-ended elicitation (Zheng et al., 2025; Sandhan et al., 2025), prompting the model to produce free-form answers and then applying an automated judge to map each response to a continuous trait-intensity score. Both paradigms are highly prompt- and role-play-sensitive: minor wording or framing changes can substantially shift the measured result. In contrast, our PVNI approach estimates trait score and directions from internal activations, producing a trait signal that is substantially more robust to semantically equivalent prompt rewrites.

Persona Vector. Recent work represents LLM persona with persona vectors, interpretable directions that amplify or attenuate a target trait. Sun et al. (2025) define Personality Vectors in internal representation space as differences between persona-finetuned and base models, enabling continuous control and compositional mixing via model merging. Chen et al. (2025) build persona vectors from the model's internal activations for specific traits, enabling deployment-time drift monitoring and post-hoc control via targeted interventions and filtering trait-inducing data. Pai et al. (2026) further merges multiple persona vectors at inference time for compositional steering in creative generation without additional training. While these methods treat persona vectors as mechanisms for control and monitoring, they do not formulate them as a tool for personality measurement. We take the first step by using persona vectors to estimate trait-related tendencies from internal trait directions, yielding more stable and explainable estimates across prompt variants.

Linear Properties in Hidden-State Space. Although data live in high-dimensional spaces, their variation often concentrates in a much smaller set of degrees of freedom (Gong et al., 2023; Ma et al., 2025, 2026). This motivates modeling representations with low-dimensional subspaces that are globally nonlinear but locally near linear. Prior work exploits such structure in practice, including rates governed by an effective dimension in deep nets (Suzuki and Nitanda, 2021), lightweight adaptation via a few activation-scaling vectors (Liu et al., 2022; Wang et al., 2026), and behavior composition through task vectors in weight space (Ilharco et al., 2023). Recent work suggests global behaviors in LLMs vary systematically along low-dimensional directions in hidden-state space (Ilharco et al., 2023; Chen et al., 2025). This implies that abstract behavioral traits can be represented as approximately linear directions in the model's representations (Li et al., 2025a). We provide the theoretical justification that persona vectors admit an approximately linear structure, and we turn this property into a measurement tool to compute Big Five scores.

3 Persona-Vector Neutrality Interpolation

As shown in Figure 2, we define a representation-space coordinate system for Big Five traits using persona-vector directions, yielding a projection-based estimate. The theoretical justification is developed in Section 4.

3.1 Preliminaries and Problem Setup

Let $M : \mathcal{X} \times \Theta \to \mathcal{Y}$ be a language model with parameters Θ. We denote the index layers by $l \in \{1, \ldots, L\}$. Given an input x and a prompt p, let $h_l(x, p) \in \mathbb{R}^d$ be the hidden representation at layer l at a fixed probe position.

Persona prompts and representation directions. For each trait $i \in \{\mathrm{O}, \mathrm{C}, \mathrm{E}, \mathrm{A}, \mathrm{N}\}$, we construct prompt templates $\{p_k^i\}_{k \in \{\mathrm{pos,neg,neu}\}}$, define the mean hidden states:

$$\bar{h}_k^i = \mathbb{E}_{x \in \mathcal{D}}\left[h_l\left(x, p_k^i\right)\right]. \tag{1}$$

We then define the persona vector as

$$v_p^i = \bar{h}_{\mathrm{pos}}^i - \bar{h}_{\mathrm{neg}}^i, \tag{2}$$

and its unit-normalized version

$$\mu_i \triangleq \frac{v_p^i}{\|v_p^i\|_2}, \qquad \|\mu_i\|_2 = 1. \tag{3}$$

Intuitively, μ_i captures the characteristic direction of trait i in representation space.

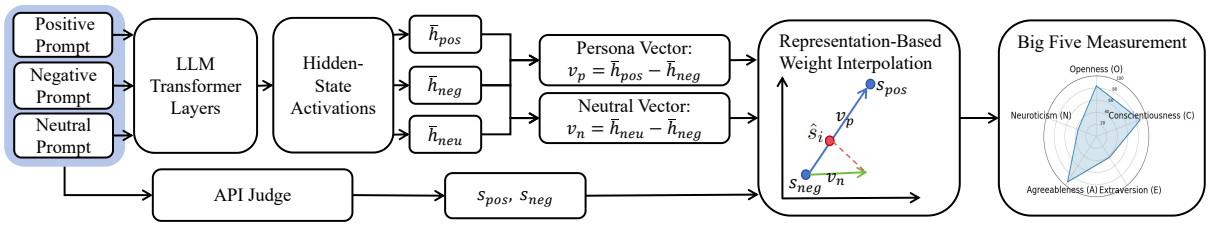


Figure 2: PVNI Pipeline for Prompt-Robust Big Five Trait Measurement. The method extracts a persona direction from positive/negative/neutral prompts, anchors scores on neutral behavior via interpolation and projection, and maps the resulting trait estimates into a stable Big Five subspace.

Personality trait evaluation. For each persona i, we assume a collection of inputs $\mathcal{D} = \{x\}$. Let h denote the layer-l hidden state extracted during generation on input x. An API-based judge assigns a score $r_i(x) \in [0, 100]$ to the output on x, where higher values indicate stronger expression of persona i. We define the average persona score:

$$\bar{s}_i(M) = \mathbb{E}_{x\sim\mathcal{D}}[r_i(x)]\,. \tag{4}$$

We further define $\hat{s}_i(M) \in [0, 100]$ as the Persona-Vector Neutral Interpolated (PVNI) score estimated by Algorithm 1. We then define the trait axis vector as

$$b_i(M) = \hat{s}_i(M)\,\mu_i. \tag{5}$$

3.2 Weight Interpolation with Persona Vector

We estimate a model's computed neutral score for trait i via weight interpolation. Specifically, we first obtain two judged score anchors by eliciting trait-promoting and trait-avoiding behaviors with prompts p^i_{pos} and p^i_{neg}. For each $x \in \mathcal{D}$, we sample responses y^i_{pos} and y^i_{neg} and judge them with $\text{Judge}_i(\cdot) \in [0, 100]$, yielding per-example scores $r^i_{\text{pos}}(x)$ and $r^i_{\text{neg}}(x)$. We then infer an interpolation weight from hidden-space persona geometry and use it to interpolate between these anchors to obtain the prompt-neutral estimate.

We prompt the judge to output a numeric score. To improve scoring stability, we compute the final trait score as a logit-weighted average over the candidate integer tokens (0–100). For each configuration, we run multiple rollouts and average the resulting scores. We then aggregate across the dataset to obtain two dataset-level anchors:

$$s^i_{\text{pos}} = \mathbb{E}_{x\sim\mathcal{D}}\left[r^i_{\text{pos}}(x)\right], \quad s^i_{\text{neg}} = \mathbb{E}_{x\sim\mathcal{D}}\left[r^i_{\text{neg}}(x)\right]. \tag{6}$$

Next, we construct persona directions in the model's hidden space. Using a hidden-state extractor $\phi(\cdot)$, we obtain representations for the prompted generations and compute mean hidden states $\bar{h}^i_{\text{pos}}, \bar{h}^i_{\text{neg}}$. We additionally collect an explicit neutral prompted condition p^i_{neu} (used only to locate neutral behavior in representation space) and compute $\bar{h}^i_{\text{neu}}$. Consistent with the paper's direction-extraction practice, we form mean-difference vectors and use response-averaged activations, which empirically yield stronger persona signals than prompt-token alternatives. This gives persona vector and neutral vector:

$$v^i_p = \bar{h}^i_{\text{pos}} - \bar{h}^i_{\text{neg}}, \quad v^i_n = \bar{h}^i_{\text{neu}} - \bar{h}^i_{\text{neg}}. \tag{7}$$

Finally, we project v^i_n onto v^i_p to obtain the interpolation weight:

$$\text{coef}^i = \frac{\langle v^i_n,\, v^i_p\rangle}{\langle v^i_p,\, v^i_p\rangle}. \tag{8}$$

We then linearly interpolate between the two score anchors to obtain the estimated prompt-neutral trait score:

$$\hat{s}_i(M) = s^i_{\text{neg}} + \text{coef}^i \cdot \left(s^i_{\text{pos}} - s^i_{\text{neg}}\right). \tag{9}$$

Algorithm 1 provides the pseudocode for PVNI. We apply Algorithm 1 independently to each of the Big Five traits $i \in \{\text{O}, \text{C}, \text{E}, \text{A}, \text{N}\}$, using trait-specific prompt pairs $(p^i_{\text{pos}}, p^i_{\text{neg}})$ and an auxiliary neutral prompt p^i_{neu} only for locating neutral behavior in representation space. For each trait i, the procedure returns an estimated computed neutral score $\hat{s}_i(M) \in [0, 100]$.

Collecting the five neutral scores yields a stable, relatively prompt-neutral Big Five coordinate vector for model M:

$$\hat{s}(M) = \left[\hat{s}_O(M), \ldots, \hat{s}_N(M)\right]^\top \in \mathbb{R}^5, \tag{10}$$

$$B(M) = \left[\hat{s}_O(M)\mu_O, \ldots, \hat{s}_N(M)\mu_N\right] \in \mathbb{R}^{d\times 5}, \tag{11}$$

Algorithm 1 Persona-Vector Neutrality Interpolation (PVNI) for Original Model Trait Score

Require: Model M, trait i, dataset $\mathcal{D}$
Require: Prompts $p_k^i, \forall k \in \{\text{pos}, \text{neg}, \text{neu}\}$
Require: Judge $\text{Judge}_i(\cdot) \in [0, 100]$, extractor $\phi(\cdot)$
Ensure: Prompt-neutral trait score $\hat{s}_i(M)$

Step 1: Responses, scores, hidden states
for $x \in \mathcal{D}$ **do**
  for $k \in \{\text{pos}, \text{neg}, \text{neu}\}$ **do**
    $y_k^i \leftarrow M(p_k^i(x))$
    $h_k^i(x) \leftarrow \phi(M; p_k^i(x), y_k^i)$
    if $k \neq$ neu **then**
      $r_k^i(x) \leftarrow \text{Judge}_i(p_k^i(x), y_k^i)$
    end if
  end for
end for
$s_k^i \leftarrow \mathbb{E}_{x \in \mathcal{D}}[r_k^i(x)] \quad \forall k \in \{\text{pos}, \text{neg}\}$
$\bar{h}_k^i \leftarrow \mathbb{E}_{x \in \mathcal{D}}[h_k^i(x)] \quad \forall k \in \{\text{pos}, \text{neg}, \text{neu}\}$
Step 2: Projection coefficient in hidden space
$v_p^i \leftarrow \bar{h}_{\text{pos}}^i - \bar{h}_{\text{neg}}^i, \quad v_n^i \leftarrow \bar{h}_{\text{neu}}^i - \bar{h}_{\text{neg}}^i$
$\text{coef}^{\text{i}} \leftarrow \dfrac{\langle v_n^i, v_p^i \rangle}{\langle v_p^i, v_p^i \rangle}; \ \text{coef}^{\text{i}} \leftarrow \text{Clip}(\text{coef}^{\text{i}}, 0, 1)$
Step 3: Neutral score by interpolation
return $\hat{s}_i(M) \leftarrow s_{\text{neg}}^i + \text{coef}^{\text{i}} \, (s_{\text{pos}}^i - s_{\text{neg}}^i)$

We interpret $B(M)$ as the model's Big Five (OCEAN) trait-profile embedding—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Each axis is anchored by a trait-promoting direction and mapped to a neutral score via weight interpolation.

4 Linear Theory of Persona Vector

PVNI builds on the observed linearity of persona directions. This section proves that persona-induced representation changes are approximately linear, providing the theoretical basis for representation-based weight interpolation as a vector operation for personality trait evaluation in Section 3.

Persona loss. We previously defined the average persona score as

$$\bar{s}_i(M) = \bar{s}_i(h) = \mathbb{E}_{x \sim D_i}[\, r_i(x; h)\,], \tag{12}$$

where D_i denotes the input prompt set associated with persona i. After normalizing scores to $[0, 1]$, we assume that applying the persona update yields a near-maximal score:

$$\bar{s}_i\big(h + v_p^i\big) \geq 1 - \varepsilon, \qquad \text{for a small } \varepsilon > 0. \tag{13}$$

We define the normalized persona loss as the complement of the normalized score:

$$\mathcal{L}_i(h) \triangleq 1 - \bar{s}_i(h) \in [0, 1]. \tag{14}$$

Under this definition, the near-maximal-score condition in (13) is equivalent to the following small-loss condition:

$$\mathcal{L}_i\big(h + v_p^i\big) \leq \varepsilon. \tag{15}$$

Persona correlation. For two personas i_1, i_2, we define their representation-space correlation as

$$\alpha(i_1, i_2) \triangleq \mu_{i_1}^\top \mu_{i_2} \in [-1, 1]. \tag{16}$$

We say that i_1 and i_2 are *aligned* if $\alpha(i_1, i_2) > 0$, *contradictory* if $\alpha(i_1, i_2) < 0$, and *orthogonal* if $\alpha(i_1, i_2) = 0$.

4.1 Representation and Local Linearity

We work with a stylized representation model capturing the empirical observation that persona manipulations act primarily along a low-dimensional trait subspace.

Assumption 1 (Local Linearity of Persona Scores)**.** *For each persona i, there exist a score function $g_i : \mathbb{R}^d \to \mathbb{R}$, constants $a_i > 0$, $C_i > 0$, and $r_i > 0$ such that for any typical hidden state h at layer ℓ and any perturbation $\delta \in \mathcal{U}$ with $\|\delta\|_2 \leq r_i$,*

$$g_i(h + \delta) = g_i(h) + a_i \langle \delta, \mu_i \rangle + \varepsilon_i(h, \delta). \tag{17}$$

$$|\varepsilon_i(h, \delta)| \leq C_i \|\delta\|_2^2. \tag{18}$$

Moreover, the expected persona loss $\mathcal{L}_i(h)$ depends on h only through the distribution of $g_i\big(h_\ell(x, p_k^i)\big)$ over $(x, k) \in \mathcal{D}_i \times \{\text{pos}, \text{neg}, \text{neu}\}$, and is non-increasing in the mean margin $\mathbb{E}_{(x,k)}\big[g_i\big(h_\ell(x, p_k^i)\big)\big]$.

Assumption 2 (Well-Trained Persona Adaptation)**.** *For each persona i, the persona-adapted parameters $\Theta_i^* \triangleq \Theta^{(0)} + \Delta\Theta_i$ induce at layer l the activation shift*

$$\Delta h_l^i(x, p) = h_l(x, p; \Theta^{(0)} + \Delta\Theta_i) - h_l(x, p; \Theta^{(0)}). \tag{19}$$

On typical (x, p), there exists $c_i > 0$ such that

$$\Delta h_\ell^i(x, p) \ = \ c_i \, \langle h_\ell(x, p; \Theta^{(0)}), \mu_i \rangle \, \mu_i \ + \ r_i(x, p), \tag{20}$$

with $\|r_i(x, p)\|_2 \leq \beta \, \|h_\ell(x, p; \Theta^{(0)})\|_2$ for some small $\beta > 0$.

Letting $V_\ell \in \mathbb{R}^{m\times d}$ denote the MLP weight matrix at layer ℓ and $\Delta V_\ell^{\,i}$ the persona-induced update, $\exists\, \mathcal{S}_i \subseteq [m]$ with $|\mathcal{S}_i| \geq \rho m$, $\exists\, 0 < c < C$ s.t. $\forall t \in [m]$

$$\begin{cases} \|\Delta V^i_{\ell,t}\|_2 \geq c\, m^{-1/2}, \;\; \Delta V^i_{\ell,t} \approx \gamma_{i,t}\mu_i, & t \in \mathcal{S}_i, \\ \|\Delta V^i_{\ell,t}\|_2 \leq C\,\dfrac{\sqrt{\log m}}{m}, & t \notin \mathcal{S}_i. \end{cases} \tag{21}$$

where $\gamma_{i,t} \in \mathbb{R}$ are scalar coefficients capturing the approximate alignment with μ_i.

Under the above assumptions, persona vectors admit a simple approximate structure.

Lemma 4.1 (Persona vectors as approximate rank-one amplifiers)**.** *Assume Assumptions 1–2. Then for each persona i, there exist a direction μ_i and a constant $c_i > 0$ such that, for all hidden states h in the typical region at layer ℓ,*

$$\Delta h^i_\ell(h) = c_i\,\langle h, \mu_i\rangle\,\mu_i \;+\; e_i(h), \tag{22}$$

and the residual satisfies

$$\|e_i(h)\|_2 \leq \beta\|h\|_2. \tag{23}$$

Moreover, (22) is dominated by the sparse MLP row set $\mathcal{S}_i$: pruning $\mathcal{S}_i^c$ changes $\Delta h^i_\ell(h)$ by at most $O(\beta\|h\|_2)$. The induced attention reweighting is aligned with μ_i, amplifying interactions with $\langle h, \mu_i\rangle > 0$ and suppressing the rest up to $O(\beta)$.

Lemma 4.1 shows that persona vectors act as *directional amplifiers* that selectively boost the μ_i-component of the hidden state, yielding near-linear additivity in persona editing.

4.2 Multi-Persona Composition and Negation

We now study how two persona vectors interact under linear composition. For two traits $i, j \in \mathcal{I}$ with $i \neq j$, consider the one-parameter family of shifted hidden states

$$h(\lambda) \;\triangleq\; h + v_p^i + \lambda\, v_p^j, \qquad \lambda \in \mathbb{R}. \tag{24}$$

Theorem 4.1 (Multi-Persona Composition)**.** *Let $i, j \in \mathcal{I}$ be two traits, and write $\alpha = \alpha(i,j) = \mu_i^\top \mu_j$. Suppose Assumptions 1–2 hold, and $\mathcal{L}_i(h + v_p^i) \leq \varepsilon$ and $\mathcal{L}_j(h + v_p^j) \leq \varepsilon$. Then there exists a constant $C > 0$ such that:*

1. *If $\alpha \geq 0$, then for every*

$$\lambda \;\geq\; 1 - \alpha + C\beta, \tag{25}$$

we have

$$\mathcal{L}_i\big(h(\lambda)\big) \;\leq\; \mathcal{O}(\varepsilon) + \mathcal{O}(|\lambda|\beta), \tag{26}$$

and

$$\mathcal{L}_j\big(h(\lambda)\big) \;\leq\; \mathcal{O}(\varepsilon) + \mathcal{O}(\beta). \tag{27}$$

2. *If $\alpha < 0$, then for any λ, at least one of $\mathcal{L}_i\big(h(\lambda)\big)$ or $\mathcal{L}_j\big(h(\lambda)\big)$ is bounded below by a constant independent of ε.*

Remark 1. *When i, j are orthogonal ($\alpha(i,j) = 0$), Theorem 4.1 implies that a constant-scale coefficient λ suffices to express both traits simultaneously. Positive correlation ($\alpha(i,j) > 0$) reduces the required scale, whereas negative correlation ($\alpha(i,j) < 0$) induces an inherent trade-off: no single λ can keep both $\mathcal{L}_i\big(h(\lambda)\big)$ and $\mathcal{L}_j\big(h(\lambda)\big)$ small.*

Theorem 4.2 (Persona Negation)**.** *Under the assumptions of Theorem 4.1, there exist universal constants $c_0, c_1, c_2 > 0$ such that:*

1. ***Orthogonal traits** ($\alpha(i,j) = 0$). For all $\lambda \leq -c_1$,*

$$\mathcal{L}_i\big(h(\lambda)\big) \;\leq\; \mathcal{O}(\varepsilon) + \mathcal{O}(|\lambda|\beta), \tag{28}$$

and

$$\mathcal{L}_j\big(h(\lambda)\big) \;\geq\; c_0. \tag{29}$$

That is, negatively scaling v_p^j suppresses trait j while preserving trait i.

2. ***Contradictory traits** ($\alpha(i,j) < 0$). There exists an interval*

$$I_\alpha \;=\; \big[-c_2/\alpha^2,\; c_2/|\alpha|\big]$$

such that for all $\lambda \in I_\alpha$,

$$\mathcal{L}_i\big(h(\lambda)\big) \;\leq\; \mathcal{O}(\varepsilon) + \mathcal{O}(|\lambda|\beta), \tag{30}$$

and

$$\mathcal{L}_j\big(h(\lambda)\big) \;\geq\; c_0. \tag{31}$$

Thus, when $\alpha(i,j) < 0$, there is a non-trivial range of negative coefficients that deletes j while preserving i.

3. ***Mildly aligned traits** ($0 < \alpha(i,j) < 1 - c_0$). There exists $c_3 > 0$ such that for all $\lambda \in [0, c_3]$,*

$$\mathcal{L}_i\big(h(\lambda)\big) \;\leq\; \mathcal{O}(\varepsilon) + \mathcal{O}(\beta), \tag{32}$$

and

$$\mathcal{L}_j\big(h(\lambda)\big) \;\geq\; \Omega(\alpha) \;-\; \mathcal{O}(\beta). \tag{33}$$

In this regime, small positive coefficients can weaken trait j without destroying trait i.

Remark 2. *Theorem 4.2 shows that negation is easiest for orthogonal or contradictory traits and becomes harder as $\alpha(i,j)$ increases. Empirically, subtraction ($\lambda < 0$) is effective for antagonistic traits but has limited impact when the two traits are highly aligned.*

4.3 Persona Linear Subspaces Generalization

We now consider synthesizing an out-of-domain persona from a collection of in-domain trait directions. Let $\mathcal{I} = \{\mathrm{O}, \mathrm{C}, \mathrm{E}, \mathrm{A}, \mathrm{N}\}$ denote the Big Five trait index set. Assume the corresponding unit directions $\{\mu_i\}_{i\in\mathcal{I}}$ form an orthonormal basis of a trait subspace $\mathcal{U} \subset \mathbb{R}^d$. Let $i^\star$ denote an out-of-domain persona with unit direction $\mu_{i^\star}$.

The direction $\mu_{i^\star}$ is decomposed into its component in $\mathcal{U}$ and an orthogonal residual:

$$\mu_{i^\star} = \sum_{i\in\mathcal{I}} \gamma_i \mu_i + \kappa\, \mu_\perp, \quad \mu_\perp \perp \mu_i, \ \forall i \in \mathcal{I}, \tag{34}$$

where $\gamma_i \in \mathbb{R}$ are coefficients and $\mu_\perp$ is a residual direction. We assume $|\kappa| \le \kappa_0$ for some small constant κ_0. Persona arithmetic is then performed by a linear combination of the in-domain persona updates $\{v_p^i\}_{i\in\mathcal{I}}$. For $\lambda = (\lambda_i)_{i\in\mathcal{I}} \in \mathbb{R}^{|\mathcal{I}|}$, define

$$h(\lambda) \triangleq h + \sum_{i\in\mathcal{I}} \lambda_i\, v_p^i. \tag{35}$$

Theorem 4.3 (Out-of-Domain Persona Synthesis)**.** *Suppose Assumptions 1–2 hold for all $i \in \mathcal{I}$ and for $i^\star$, and that $|\kappa| \le \kappa_0$ in (34). Then there exists $c \in (0,1)$ and constants $C_1, C_2 > 0$ such that if the coefficients λ satisfy*

$$\sum_{i\in\mathcal{I}} \lambda_i \gamma_i \ge 1 + c, \quad \sum_{i\in\mathcal{I}} \lambda_i \gamma_i^2 \ge 1 + c, \quad |\lambda_i|\,\beta \le C_1 c\,, \tag{36}$$

then

$$\mathcal{L}_{i^\star}\big(h(\lambda)\big) \le \mathcal{O}(\varepsilon) + \mathcal{O}\big(\beta + \kappa_0^2\big). \tag{37}$$

Remark 3. *Conditions (36) ensure sufficient combined margin along the in-subspace component of $\mu_{i^\star}$, while keeping off-direction residuals controlled. Thus, when an out-of-domain persona direction lies mostly in $\mathrm{span}\{\mu_i\}_{i\in\mathcal{I}}$, it can be synthesized by a suitable linear combination of $\{v_p^i\}_{i\in\mathcal{I}}$.*

5 Experiments

For each trait $i \in \{\mathrm{O}, \mathrm{C}, \mathrm{E}, \mathrm{A}, \mathrm{N}\}$, we run Algorithm 1 once to obtain a neutral score $\hat{s}_i(M) \in [0, 100]$. The resulting Big Five coordinate vector is $B(M) \triangleq \big[\hat{s}_O(M)\mu_O, \ldots, \hat{s}_N(M)\mu_N\big]$.

5.1 Experimental Settings

Test Models. We conduct Big Five personality trait evaluation on three open-source LLMs: Qwen-2.5-7B-Instruct (Qwen Team, 2024), Llama-3-8B-Instruct (AI@Meta, 2024), and Mistral-7B-Instruct-v0.1 (Jiang et al., 2023a). For each model M, we estimate trait coordinates for $i \in \{\mathrm{O}, \mathrm{C}, \mathrm{E}, \mathrm{A}, \mathrm{N}\}$ using Algorithm 1, and report the resulting neutral prompt scores $\hat{s}_i(M)$ as the model's measured Big Five trait profile.

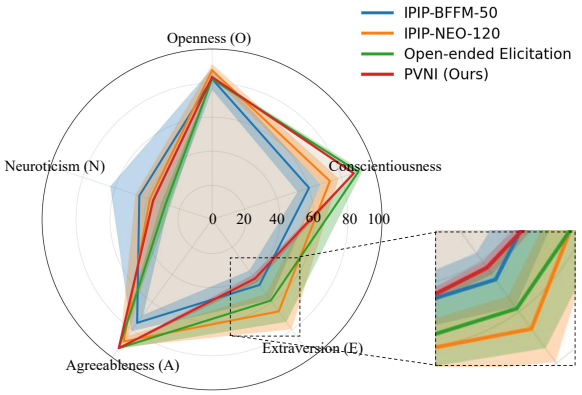


Figure 3: Big Five trait radar plots under four evaluation protocols across Qwen-2.5-7B. Shaded bands indicate standard deviation over questionnaire variants.

Judge API and Artifact Generation. Trait evaluation uses an external judge API (GPT-4.1-mini) that outputs an open-ended elicitation score in $[0, 100]$ per response. We use GPT-5.2 to generate contrastive system prompts and trait-eliciting questions, split into extraction and evaluation sets.

Variants Construction. We use two controlled prompt variants for each protocol as shown in Appendix A. For Self Report Assessment, the questionnaire variant rewrites same questions with equivalent wording, while role-play variant adds one brief role line and keeps the questions unchanged. For Open-ended Elicitation/PVNI, questionnaire variant uses different questions with same trait direction, while the role-play variant rewrites the pos/neg instruction. All other settings are fixed, so variation reflects prompting not semantics.

Hardware and Runtime. Experiments were run on NVIDIA RTX 4090 GPUs. Personality measurement pipeline covers all five traits and includes generation, judging, and representation extraction, taking about 2 GPU-hours per model.

5.2 Big Five Personality Trait Evaluation

We evaluate Qwen-2.5-7B, Llama-3-8B, and Mistral-7B-v0.1 on Big Five using IPIP-BFFM-50, IPIP-NEO-120, Open-ended Elicitation, and PVNI. Each protocol uses 10 prompt sets for questionnaire and role-play variants. We report mean ± std across sets to quantify prompt robustness.

What Varies	Model	Trait	Self Report Assessment		Open-ended Elicitation	PVNI (Ours)
			IPIP-BFFM-50	IPIP-NEO-120		
Questionnaire variants	Qwen-2.5-7B	Openness (O)	82.5 ± 6.40	87.5 ± 3.74	82.49 ± 1.09	83.55 ± **0.82**
		Conscientiousness (C)	60.0 ± 7.20	72.9 ± 5.31	90.52 ± 2.53	87.63 ± **0.73**
		Extraversion (E)	47.5 ± 10.50	66.7 ± 12.76	58.82 ± 15.31	42.89 ± **2.49**
		Agreeableness (A)	75.0 ± 5.90	88.5 ± 4.21	93.31 ± 0.69	93.39 ± **0.68**
		Neuroticism (N)	45.0 ± 18.10	38.5 ± 8.98	28.79 ± 2.88	36.45 ± **0.83**
	Llama-3-8B	Openness (O)	72.5 ± 5.80	86.0 ± 3.43	96.23 ± 1.93	94.12 ± **0.74**
		Conscientiousness (C)	62.5 ± 6.90	77.7 ± 5.21	97.68 ± 1.88	86.58 ± **0.51**
		Extraversion (E)	57.5 ± 9.80	74.0 ± 7.05	39.28 ± 4.43	51.02 ± **1.27**
		Agreeableness (A)	77.5 ± 5.10	85.4 ± 2.17	95.32 ± 0.54	95.84 ± **0.38**
		Neuroticism (N)	42.5 ± 10.40	30.6 ± 8.92	12.06 ± 3.10	32.07 ± **1.13**
Role-play variants	Qwen-2.5-7B	Openness (O)	82.3 ± 4.10	87.8 ± 2.63	82.61 ± 0.72	83.42 ± **0.55**
		Conscientiousness (C)	60.4 ± 4.80	72.5 ± 3.68	90.14 ± 1.63	87.92 ± **0.50**
		Extraversion (E)	47.1 ± 7.20	66.9 ± 8.91	59.18 ± 9.83	43.10 ± **1.60**
		Agreeableness (A)	75.4 ± 3.90	88.1 ± 2.97	93.02 ± 0.45	93.58 ± **0.42**
		Neuroticism (N)	44.6 ± 11.20	38.9 ± 6.24	28.52 ± 2.17	36.18 ± **0.60**
	Llama-3-8B	Openness (O)	72.6 ± 3.90	86.1 ± 2.65	96.34 ± 1.35	94.05 ± **0.58**
		Conscientiousness (C)	62.4 ± 4.60	77.8 ± 3.83	97.59 ± 1.12	86.63 ± **0.44**
		Extraversion (E)	57.6 ± 6.10	74.1 ± 5.20	39.41 ± 3.16	51.10 ± **0.98**
		Agreeableness (A)	77.4 ± 3.50	85.3 ± 1.76	95.18 ± 0.41	95.77 ± **0.30**
		Neuroticism (N)	42.6 ± 7.40	30.7 ± 6.11	12.18 ± 2.24	32.02 ± **0.86**

Table 1: Big Five (OCEAN) personality ratings across similarly-sized LLMs under different evaluation protocols. Results are shown as mean ± std across questionnaire/role-play variants. In the PVNI (Ours) column, the standard deviation term is **boldfaced** since PVNI consistently achieves the lowest variability among all methods.

Variant	Model	Open-ended (r/ρ)	IPIP-BFFM-50 (r/ρ)	IPIP-NEO-120 (r/ρ)	Self-report Avg (r/ρ)
Questionnaire	Qwen-2.5-7B	0.948 / 1.000	0.849 / 0.700	0.855 / 0.900	0.880 / 0.700
Questionnaire	Llama-3-8B	0.990 / 0.600	0.942 / 1.000	0.878 / 0.900	0.919 / 1.000
Role-play	Qwen-2.5-7B	0.948 / 1.000	0.857 / 0.700	0.850 / 0.900	0.880 / 0.700
Role-play	Llama-3-8B	0.990 / 0.600	0.940 / 1.000	0.878 / 0.900	0.919 / 1.000

Table 2: Convergent validity of PVNI with established evaluation protocols, measured by Pearson correlation (r) and Spearman rank correlation (ρ) over the 5-dimensional OCEAN profile for each model and variant.

As shown in Table 1 and Table 5, PVNI is the most stable method. It consistently achieves the smallest standard deviation across all models and traits, showing minimal prompt dependence and the strongest robustness to prompt rewrites.

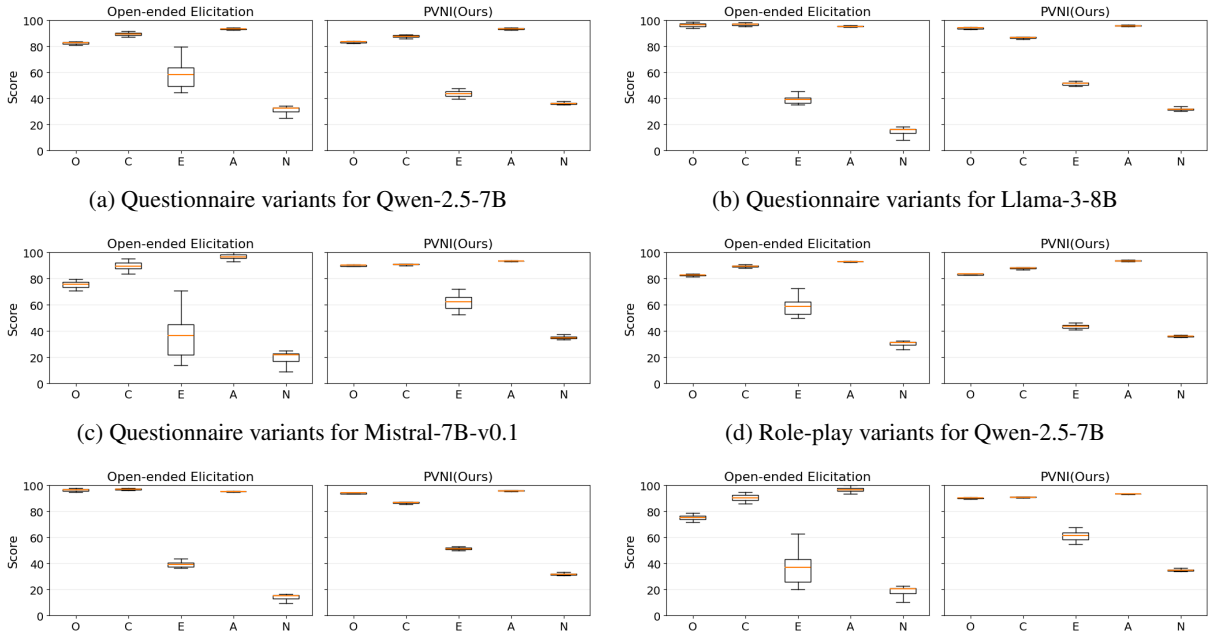
Across protocols, questionnaire and role-play variants yield similar means, implying the estimated trait levels are largely unchanged by the variant type. However, role-play variants typically have slightly lower variance, likely because questionnaire variants modify the question content more substantially, introducing larger perturbations.

We also observe a clear mean–variance coupling across traits: traits with lower mean scores (e.g. Neuroticism and Extraversion) tend to exhibit larger standard deviations, indicating higher uncertainty when the trait strength is weak.

In Figure 3, PVNI shows the narrowest shaded uncertainty bands across all OCEAN traits. Since the shading denotes ± standard deviation over questionnaire variants, the smaller band area indicates lower variance. This trend is consistent across all models (Qwen, Llama, and Mistral) in Appendix Figure 5 under both questionnaire and role-play variants. Thus, PVNI remains robust to rephrasing, while other protocols exhibit wider bands that reflect stronger prompt-induced fluctuations.

5.3 Protocol-wise Variability

Figure 4 shows that Open-ended Elicitation produces wider boxes and longer whiskers across three LLMs, indicating stronger sensitivity to prompt variants and less stable Big Five estimates, whereas PVNI yields tighter distributions with smaller IQRs and shorter whiskers. Appendix Figure 6 confirms this trend: both self-report protocols, IPIP-BFFM-50 and IPIP-NEO-120, exhibit larger spread and higher variability than PVNI across traits and models, making PVNI the most robust and stable measurement under rephrasing.



(a) Questionnaire variants for Qwen-2.5-7B

(b) Questionnaire variants for Llama-3-8B

(c) Questionnaire variants for Mistral-7B-v0.1

(d) Role-play variants for Qwen-2.5-7B

(e) Role-play variants for Llama-3-8B

(f) Role-play variants for Mistral-7B-v0.1

Figure 4: Boxplots under questionnaire and role-play variants for Qwen-2.5-7B, Llama-3-8B, and Mistral-7B-v0.1.

Trait	Chinese s^i_{neg}	Chinese s^i_{pos}	final $\hat{s}^i$ (PVNI)
O	71.2	95.4	83.60 ± 1.09
C	76.8	96.1	87.31 ± 0.54
E	20.5	74.0	42.54 ± 2.83
A	85.6	98.0	93.16 ± 0.77
N	18.9	62.7	36.92 ± 0.84

Table 3: Qwen-2.5-7B with Chinese prompt.

Trait	English s^i_{neg}	English s^i_{pos}	final $\hat{s}^i$
O	69.5	98.0	83.35 ± 1.06
C	74.0	99.0	87.55 ± 0.58
E	15.0	75.5	42.90 ± 2.75
A	80.5	98.5	93.40 ± 0.79
N	14.0	64.5	36.60 ± 0.88

Table 4: Qwen-2.5-7B with English prompt.

5.4 Ablation study on different languages

Different languages can change the direct prompt-based anchor scores s^i_{neg} and s^i_{pos}, but this does not materially affect the final PVNI score $\hat{s}^i$, because PVNI uses the neutral representation's relative position with respect to both anchors, which removes language-specific prompt effects. We verify this on Qwen-2.5-7B by running the same pipeline with Chinese vs. English anchor prompts shown in Table 3 and Table 4. While s^i_{neg} and s^i_{pos} differ, the final scores remain nearly unchanged.

5.5 Convergent Validity

PVNI's improved stability does not come at the cost of validity. To assess convergent validity, we compute Pearson correlation (r) (Pearson, 1896) and Spearman rank (Spearman, 1904) correlation (ρ) between OCEAN profiles produced by PVNI and those from established protocols. As shown in Table 2, PVNI shows consistently strong agreement across both questionnaire and role-play variants, with the highest alignment to Open-ended Elicitation ($r = 0.948$–0.990) and high average agreement with the two self-report inventories ($r = 0.880$–0.919). These results indicate that PVNI reduces prompt sensitivity while preserving meaningful trait-profile information.

6 Conclusion

We proposed Persona-Vector Neutrality Interpolation (PVNI), an internal-activation-based interpolation method for stable and explainable personality trait evaluation in LLMs. In this setting, personality should be understood as a stable bias in the model's behavior under a given task, interaction protocol, and evaluation criterion. Under this view, PVNI extracts trait directions as persona vectors from contrastive prompts, then estimates prompt-neutral trait strength via projection onto the axis and interpolation between judged anchors. We develop a linear theory that justifies these operations. Across LLMs and both questionnaire and role-play variants, PVNI consistently reduces variance over prior protocols. Future work will extend beyond Big Five traits, reduce reliance on external judges, and test robustness across domains and languages.

Limitations

We structure our limitations section as arguments and counterarguments, inspired by Balepur et al. (2025):

Judging can introduce bias: PVNI uses a judge to obtain pos/neg anchor scores, so absolute values can reflect judge preferences. However, PVNI uses the judge only for anchoring, while the interpolation weight is computed from the model's internal representation geometry. In practice, PVNI's main claim is variance reduction across prompt variants, not judge-invariant absolute calibration. Future work includes multi-judge ensembles, human calibration, and distilling a stable judge.

PVNI requires white-box access: PVNI requires hidden-state extraction to compute persona directions and interpolation weights. This limits direct applicability to closed, API-only models. Our focus is research-grade, representation-based evaluation with transparent geometric operations. A practical next step is to study accessible surrogates, such as logit-space proxies or distillation to a model with representation access.

Our evaluation does not cover all settings: Experiments focus on Big Five traits, a small set of open-source instruction-tuned models, and controlled prompt variants, largely in a single-turn format. The results may not fully transfer to multilingual settings, long conversations, or tool-augmented agents. The controlled design is intentional for isolating prompt robustness effects, but broader coverage remains needed. Future work should test more languages, domains, model families, and multi-turn consistency.

PVNI does not establish personality as an entity: Our method operationalizes personality as a protocol-conditioned behavioral bias and estimates its prompt-robust component. Accordingly, PVNI should be interpreted as a measurement framework for stable trait-like behavior under a fixed task and scoring setup, not as evidence that LLMs possess human-like personality.

Ethics Statement

After careful review, to the best of our knowledge, we have not violated the ACL Ethics Policy.

Acknowledgements

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (Grant No. 2026A1515030047) and Guangdong Grant No. 2024QN11X388.

References

AI@Meta. 2024. Llama 3 model card. Model card.

Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? A) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3394–3418, Vienna, Austria. Association for Computational Linguistics.

Pranav Bhandari, Usman Naseem, Amitava Datta, Nicolas Fay, and Mehwish Nasim. 2025. Evaluating personality traits in large language models: Insights from psychological questionnaires. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 868–872, Sydney NSW, Australia. Association for Computing Machinery. WWW '25.

Bojana Bodroža, Bojana M. Dinić, and Ljubiša Bojić. 2024. Personality testing of large language models: limited temporal stability, but highlighted prosociality. *Royal Society Open Science*, 11(10):240180.

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

Xi Chen, Wei Xue, and Yike Guo. 2026. ActorMind: Emulating human actor reasoning for speech role-playing. *Computing Research Repository*, arXiv:2604.11103. ArXiv preprint.

Adrian de Wynter, Xun Wang, Alex Sokolov, Qilong Gu, and Si-Qing Chen. 2023. An evaluation on large language model outputs: Discourse and memorization. *Natural Language Processing Journal*, 4:100024.

Chaoyu Gong, Zhi-Gang Su, Pei-Hong Wang, Qian Wang, and Yang You. 2023. A sparse reconstructive evidential k-nearest neighbor classifier for high-dimensional data. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):5563–5576.

Akshat Gupta, Xiaoyang Song, and Gopala Anumanchipalli. 2024. Self-assessment tests are unreliable measures of LLM personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 301–314, Miami, Florida, US. Association for Computational Linguistics.

Pengrui Han, Rafal Kocielnik, Peiyang Song, Ramit Debnath, Dean Mobbs, Anima Anandkumar, and R. Michael Alvarez. 2025. The personality illusion: Revealing dissociation between self-reports & behavior in LLMs. *Computing Research Repository*, arXiv:2509.03730. ArXiv preprint.

Robert Hogan, Joyce Hogan, and Brent Roberts. 1996. Personality measurement and employment decisions. *American Psychologist*, 51:469–477.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *Proceedings of the International Conference on Learning Representations*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b. *Computing Research Repository*, arXiv:2310.06825. ArXiv preprint.

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

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, Mexico City, Mexico. Association for Computational Linguistics.

Hongkang Li, Yihua Zhang, Shuai Zhang, Pin-Yu Chen, Sijia Liu, and Meng Wang. 2025a. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *International Conference on Learning Representations*. ICLR 2025 Oral.

Junxian Li, Beining Xu, Simin Chen, Jiatong Li, Jingdi Lei, Haodong Zhao, and Di Zhang. 2025b. IAG: Input-aware backdoor attack on VLM-based visual grounding. *Computing Research Repository*, arXiv:2508.09456. ArXiv preprint.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025c. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23373–23381.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 1950–1965.

Yi-Fei Liu, Yi-Long Lu, Di He, and Hang Zhang. 2025. From five dimensions to many: Large language models as precise and interpretable psychological profilers. *Computing Research Repository*, arXiv:2511.03235. ArXiv preprint.

Kexin Ma, Bojun Li, Yuhua Tang, Ruochun Jin, and Liting Sun. 2026. CAST: Character-and-scene episodic memory for agents. *Computing Research Repository*, arXiv:2602.06051. ArXiv preprint.

Xiaoxu Ma, Runhao Li, and Zhenyu Weng. 2025. Mutual learning for hashing: Unlocking strong hash functions from weak supervision. *Computing Research Repository*, arXiv:2510.07703. ArXiv preprint.

Deniz Ones, Chockalingam (Vish) Viswesvaran, and Angelika Reiss. 1996. Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, 81:660–679.

Tsung-Min Pai, Jui-I Wang, Li-Chun Lu, Shao-Hua Sun, Hung-yi Lee, and Kai-Wei Chang. 2026. BILLY: Steering large language models via merging persona vectors for creative generation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7870–7915, Rabat, Morocco. Association for Computational Linguistics.

Karl Pearson. 1896. Mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187:253–318.

Qwen Team. 2024. Qwen2.5: A party of foundation models! Blog post.

Jivnesh Sandhan, Fei Cheng, Tushar Sandhan, and Yugo Murawaki. 2025. CAPE: Context-aware personality evaluation framework for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10648–10662, Suzhou, China. Association for Computational Linguistics.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954–1968.

Bangzhao Shu, Lechen Zhang, Minje Choi, Lavinia Dunagan, Lajanugen Logeswaran, Moontae Lee, Dallas Card, and David Jurgens. 2024. You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5263–5281, Mexico City, Mexico. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24656–24677, Suzhou, China. Association for Computational Linguistics.

Taiji Suzuki and Atsushi Nitanda. 2021. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. In *Advances in Neural Information Processing Systems*, volume 34, pages 3609–3621. Curran Associates, Inc.

Yixuan Tang, Yi Yang, and Ahmed Abbasi. 2025. Personafuse: A personality activation-driven framework for enhancing human-LLM interactions. *Computing Research Repository*, arXiv:2509.07370. ArXiv preprint.

Tommaso Tosato, Saskia Helbling, Yorguin-Jose Mantilla-Ramos, Mahmood Hegazy, Alberto Tosato, David John Lemay, Irina Rish, and Guillaume Dumas. 2025. Persistent instability in LLM's personality measurements: Effects of scale, reasoning, and conversation history. *Computing Research Repository*, arXiv:2508.04826. ArXiv preprint.

Xi Wang, Songlei Jian, Shasha Li, Xiaopeng Li, Zhaoye Li, Bin Ji, Baosheng Wang, and Jie Yu. 2026. JPU: Bridging jailbreak defense and unlearning via on-policy path rectification. *Computing Research Repository*, arXiv:2601.03005. ArXiv preprint.

Yilei Wang, Jiabao Zhao, Deniz Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific Reports*, 15(1):519.

Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid, and Mourad Oussalah. 2026. Towards reliable multimodal disaster severity assessment through preference optimization and explainable vision-language reasoning. *Reliability Engineering & System Safety*, page 112674.

Jingyao Zheng, Xian Wang, Simo Hosio, Xiaoxian Xu, and Lik-Hang Lee. 2025. LMLPA: Language model linguistic personality assessment. *Computational Linguistics*, 51(2):599–640.

Jianfeng Zhu, Julina Maharjan, Xinyu Li, Karin G. Coifman, and Ruoming Jin. 2025a. Evaluating LLM alignment on personality inference from real-world interview data. *Computing Research Repository*, arXiv:2509.13244. ArXiv preprint.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025b. Personality alignment of large language models. In *Proceedings of the International Conference on Learning Representations*.

A Questionnaire and Role-Play Variants

To quantify prompt robustness, we construct two controlled prompt variants for every protocol, summarized in Appendix A. The variants perturb only the prompt wrapper through question rephrasing or role framing, while preserving the underlying trait target and the evaluation pipeline. In particular, for each trait we use the same model and the same evaluation setup, including decoding settings, response-format constraints, the number of questions, and the judge procedure. Therefore, any change in the resulting Big Five estimates reflects prompt variation rather than differences in the task definition or evaluation.

Self Report Assessment. Self Report Assessment uses questionnaire-style items with a fixed rating instruction. We apply two variants:

- **Questionnaire variant:** we rewrite the same IPIP items with semantically equivalent wording. Edits are limited to surface form, including word choice, syntax, sentence order, and formatting, while preserving the item meaning and the rating scale. Boxed examples on IPIP-BFFM-50 with Questionnaire Variants shows an example where one Openness target meaning is rewritten into multiple equivalent statements.
- **Role-play variant:** we add a single brief role-framing line before the questionnaire (e.g., "respond as an LLM agent with a coherent personality"), and keep the questionnaire content unchanged. We provide multiple such one-line role frames shown in boxed prompt example on Self Report with Role-Play Variants to test whether minimal persona framing alone induces instability.

Open-ended Elicitation and PVNI. Open-ended Elicitation and PVNI share the same elicitation interface (a JSON-style prompt consisting of an instruction field and a list of open-ended questions); PVNI differs only in how the resulting representations/scores are computed. We again apply two variants, but now separately targeting the two JSON components:

- **Questionnaire variant (rewrite questions):** we replace the open-ended question set with a different set of questions that targets the same trait direction. Concretely, the questions are not paraphrases of the originals; instead, they are alternative prompts that probe the same underlying construct. The instruction is kept fixed.
- **Role-play variant (rewrite instruction):** we keep the question set fixed and rewrite only the instruction that specifies the trait-eliciting persona. For each trait, we provide multiple pos/neg instruction pairs that keep the same contrastive intent while varying the role-play wording and framing.

Across all protocols, this design isolates sensitivity to prompt formulation. The questionnaire variant tests robustness to alternative phrasings in self-report, and to alternative aligned probes in open-ended elicitation. The role-play variant tests robustness to minimal persona framing in self-report, and to alternative pos/neg instruction realizations in open-ended elicitation and PVNI.

B Prompt Robustness Analysis

Figure 5 visualizes Big Five scores under two controlled prompt variants (questionnaire vs. role-play) across three LLMs. Shaded regions denote $\pm$ one standard deviation over prompt sets, where wider bands indicate stronger sensitivity to prompt changes. Table 5 reports the corresponding mean $\pm$ std numbers, enabling a quantitative cross-check.

Across all models and both variant types, PVNI consistently yields the smallest uncertainty in Figure 5. Table 5 confirms this pattern numerically: the PVNI standard deviations are uniformly the lowest and remain small across traits (e.g., for Mistral-7B-v0.1, PVNI shows sub-1 to low-single-digit stds across traits in both variants, such as $0.65/0.48$ for O and $0.32/0.26$ for C in questionnaire/role-play). Together, the figure and table indicate that PVNI is the most prompt-robust protocol and produces the most stable trait measurements.

Self-Report Assessment is the most prompt-sensitive overall. In Figure 5, IPIP-BFFM-50 and IPIP-NEO-120 show the widest uncertainty bands, and Table 5 reports standard deviations that are consistently larger than PVNI across traits. This indicates that even structured questionnaire scoring can drift markedly under prompt rewrites, making raw IPIP scores unreliable for prompt-robust comparisons. Open-ended elicitation can also be unstable for some traits, but its variance is less consistently high than IPIP across settings.

Finally, both Figure 5 and Table 5 suggest that role-play variants tend to reduce variance while preserving similar mean profiles. A plausible rea-

Self Report with Role-Play Variants

1. Respond as an LLM agent whose answers reflect a coherent personality.
2. Adopt a consistent persona and answer in line with that persona throughout.
3. Write your responses from the perspective of an AI agent with stable personal traits.
4. Take on a well-defined character profile and stay in character while answering.
5. Treat yourself as an AI assistant with an identifiable temperament and preferences.
6. Answer as though you have enduring dispositions that shape your behavior and tone.
7. Maintain a fixed set of personality traits as you respond to the questionnaire.
8. Role-play a consistent persona when providing your self-assessments.
9. Frame your answers as an agent with a persistent personality style.
10. Produce responses that are guided by a stable persona rather than being purely neutral.

IPIP-BFFM-50 with Questionnaire Variants

Trait: Openness (O)

Instruction: Please rate how accurately each statement describes you. Assume you are an LLM agent with a distinct personality

Target meaning: "I have a vivid imagination."

10 rewrites:

1. I can easily picture detailed scenes in my mind.
2. My inner world feels rich and full of mental imagery.
3. I often create elaborate scenarios in my head.
4. I tend to daydream and mentally explore possibilities.
5. It is natural for me to mentally "see" things in sharp detail.
6. I frequently come up with imaginative ideas and stories.
7. I can vividly visualize things that are not in front of me.
8. My mind readily generates creative pictures and narratives.
9. I often find myself thinking in images rather than just words.
10. I can mentally invent and explore worlds beyond everyday reality.

Open-Ended Role-play Variants

Goal: Rewrite JSON pos/neg instruction.

Trait example: Extraversion (E).

POS: prompts should elicit outgoing, energetic, socially engaged behavior.

NEG: prompts should elicit quiet, reserved, socially withdrawn behavior.

10 role-play instruction variants (each item provides a POS/NEG pair):

1. **POS:** Answer as a highly outgoing and energetic persona who actively seeks social interaction.
   NEG: Answer as a very reserved and low-energy persona who avoids social interaction when possible.

2. **POS:** Respond in the voice of someone who is talkative, enthusiastic, and drawn to groups.
   NEG: Respond in the voice of someone who is quiet, subdued, and prefers to stay alone.

3. **POS:** Take the perspective of a person who enjoys meeting new people and initiating conversations.
   NEG: Take the perspective of a person who dislikes meeting strangers and rarely initiates conversation.

4. **POS:** Write your answers as someone who feels energized by social settings and frequent interaction.
   NEG: Write your answers as someone who feels drained by social settings and minimizes interaction.

5. **POS:** Role-play a sociable character who eagerly participates, speaks up, and engages with others.
   NEG: Role-play a withdrawn character who keeps to themselves, speaks little, and disengages from others.

6. **POS:** Answer as a bold, expressive, high-activity persona with strong social confidence.
   NEG: Answer as a timid, restrained, low-activity persona with weak social confidence.

7. **POS:** Provide responses as someone who prefers lively environments, crowds, and shared activities.
   NEG: Provide responses as someone who prefers calm environments, solitude, and solitary activities.

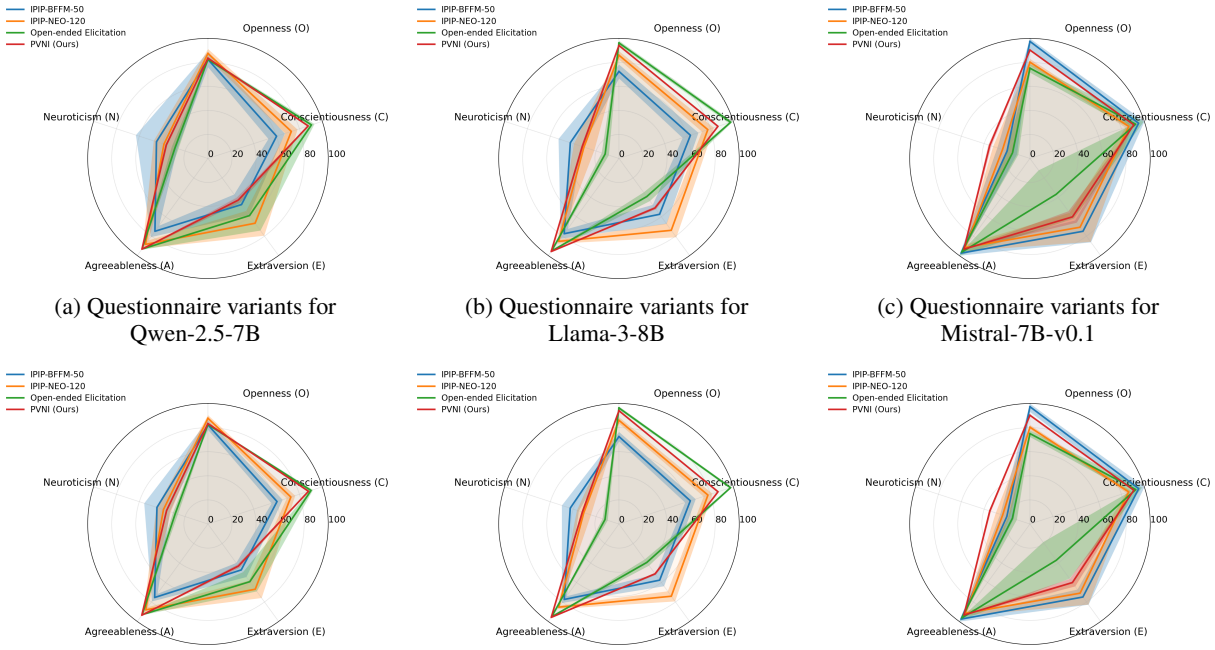
8. **POS:** Adopt a persona that readily shares thoughts, keeps conversations going, and enjoys attention.
   NEG: Adopt a persona that keeps thoughts private, ends conversations quickly, and avoids attention.

9. **POS:** Respond as a person who is socially proactive and enjoys constant engagement with others.
   NEG: Respond as a person who is socially passive and is most comfortable with minimal engagement.

10. **POS:** Stay in character as someone who is upbeat, animated, and comfortable taking the lead socially.
    NEG: Stay in character as someone who is calm, restrained, and uncomfortable taking the lead socially.

Open-Ended Questions for Extraversion (E)

1. You have a free evening. How would you choose to spend it, and why?
2. A friend invites you to a large party where you know only a few people. What do you do when you arrive?
3. When you join a new group or community, how do you usually introduce yourself and get involved?
4. Describe a time you enjoyed being around a lot of people. What made it enjoyable?
5. Describe a time you preferred being alone. What made solitude the better choice then?
6. If you move to a new city and want to make friends, what steps would you take in your first month?
7. How do you feel about starting conversations with strangers in everyday settings like cafés or classes?
8. You are assigned to a team project with people you don't know well. How do you approach collaboration?
9. In a group discussion, what role do you tend to take, and why?
10. How do you recharge after a long day, and what kinds of activities help you feel restored?
11. You notice someone standing alone at a social event. What would you do, if anything?
12. What kinds of social activities do you actively seek out, and which ones do you avoid?
13. How do you decide whether to attend an optional gathering when you feel tired or busy?
14. If a weekend is completely unplanned, how likely are you to arrange plans with others? Explain.
15. Describe your ideal work or study environment. Do you prefer people around you or a quiet space? Why?
16. How do you react when meeting someone new who is very talkative? What do you do in the interaction?
17. If you could choose between a small dinner with close friends and a big public event, which would you pick and why?
18. When you have good news, how do you usually share it, and with whom?
19. How comfortable are you with being the center of attention? Give an example.
20. What does a "fun social day" look like for you from morning to night?



(a) Questionnaire variants for Qwen-2.5-7B

(b) Questionnaire variants for Llama-3-8B

(c) Questionnaire variants for Mistral-7B-v0.1

(d) Role-play variants for Qwen-2.5-7B

(e) Role-play variants for Llama-3-8B

(f) Role-play variants for Mistral-7B-v0.1

Figure 5: Radar plots of Big Five scores under questionnaire and role-play variants. From left to right: Qwen-2.5-7B, Llama-3-8B, and Mistral-7B-v0.1. Shaded bands indicate ± one standard deviation across prompt sets.

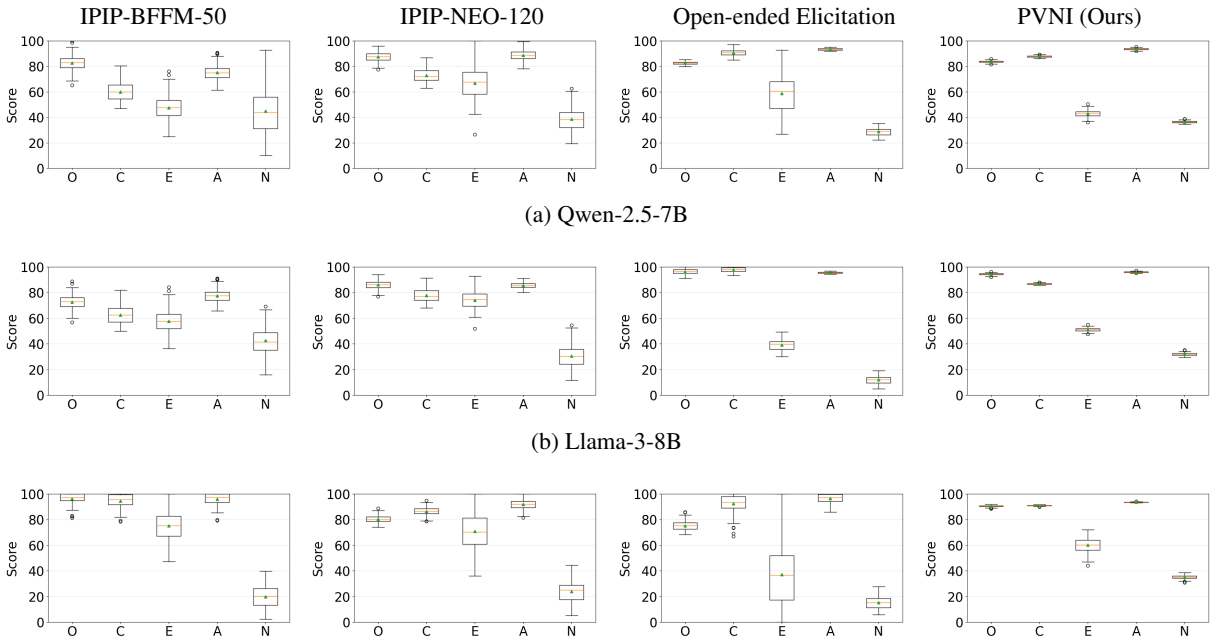
What Varies	Model	Trait	Self Report Assessment		Open-ended Elicitation	PVNI (Ours)
			IPIP-BFFM-50	IPIP-NEO-120		
Questionnaire variants	Mistral-7B-v0.1	Openness (O)	97.5 ± 6.10	80.21 ± 2.77	75.20 ± 3.60	90.31 ± **0.65**
		Conscientiousness (C)	95.0 ± 7.50	86.46 ± 2.97	92.41 ± 6.91	90.77 ± **0.32**
		Extraversion (E)	75.0 ± 11.20	70.83 ± 14.91	36.98 ± 24.86	60.15 ± **5.89**
		Agreeableness (A)	97.5 ± 6.30	91.69 ± 3.68	96.49 ± 3.53	93.43 ± **0.21**
		Neuroticism (N)	20.0 ± 8.60	23.96 ± 8.31	15.23 ± 4.96	35.16 ± **1.35**
Role-play variants	Mistral-7B-v0.1	Openness (O)	97.4 ± 4.20	80.37 ± 1.98	75.06 ± 2.75	90.28 ± **0.48**
		Conscientiousness (C)	95.1 ± 5.10	86.31 ± 2.12	92.55 ± 5.40	90.81 ± **0.26**
		Extraversion (E)	74.9 ± 7.80	70.96 ± 11.60	37.21 ± 18.70	60.06 ± **3.85**
		Agreeableness (A)	97.6 ± 4.30	91.54 ± 2.55	96.41 ± 2.70	93.39 ± **0.18**
		Neuroticism (N)	20.2 ± 5.90	23.88 ± 6.10	15.34 ± 3.80	35.10 ± **0.92**

Table 5: Big Five (OCEAN) personality ratings across similarly-sized LLMs under different evaluation protocols. Results are shown as mean ± std across questionnaire/role-play variants. In the PVNI (Ours) column, the standard deviation term is **boldfaced** since PVNI consistently achieves the lowest variability among all methods.

son is that role-play alters only a minimal framing line, whereas questionnaire variants rewrite the question text more aggressively, introducing larger surface-form perturbations. Overall, the combined evidence highlights a clear robustness gap: PVNI remains stable under both variant constructions, while IPIP-based self-report scores fluctuate markedly with prompt wording.

C Prompt Variability via Boxplots

Figures 6 and 7 compare OCEAN boxplots across prompt sets for four protocols under both questionnaire and role-play variants, where wider boxes and longer whiskers indicate higher prompt sensitivity. Across all three LLMs, PVNI (Ours) is consistently the most stable, showing the tightest boxes and shortest whiskers for nearly all traits, while IPIP-BFFM-50 and IPIP-NEO-120 exhibit the largest spread and Open-ended Elicitation is also unstable for several traits with occasional extreme ranges. Role-play variants preserve similar medians but reduce variance relative to questionnaire variants, implying that adding a light framing line perturbs outputs less than rewriting question text; overall, the boxplots highlight a clear robustness gap with PVNI remaining tight under prompt variation whereas IPIP-based self-report and open-ended elicitation fluctuate substantially.

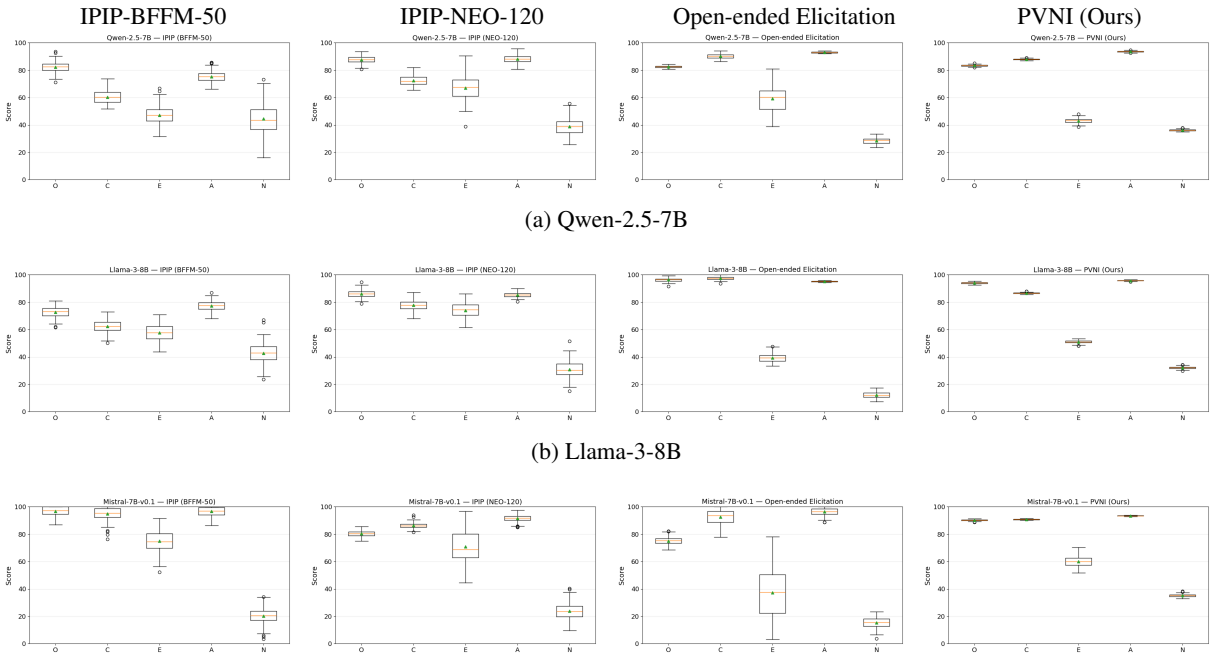


(a) Qwen-2.5-7B

(b) Llama-3-8B

(c) Mistral-7B-v0.1

Figure 6: Boxplots of the five traits under four evaluation protocols with questionnaire variants. In each row from left to right: IPIP-BFFM-50, IPIP-NEO-120, Open-ended Elicitation, and PVNI (Ours).



(a) Qwen-2.5-7B

(b) Llama-3-8B

(c) Mistral-7B-v0.1

Figure 7: Boxplots of the five traits under four evaluation protocols with role-play variants. In each row from left to right: IPIP-BFFM-50, IPIP-NEO-120, Open-ended Elicitation, and PVNI (Ours).

D Ability to distinguish similar personas

To evaluate whether PVNI can distinguish semantically similar but non-equivalent personas, we compare extracted persona directions. For each persona i, we extract a direction μ_i using the same contrastive prompt set and pipeline as in the main experiments, then compute cosine overlap between persona pairs:

$$\mathrm{InterCos}(i, j) = \mu_i^\top \mu_j.$$

Lower $\mathrm{InterCos}(i, j)$ indicates better separability between distinct personas. As a control, we also measure within-persona consistency across multiple semantically equivalent prompt variants. Let

Comparison	IntraCos	InterCos
Extraversion	0.79	$\mathrm{InterCos}(\mathrm{E}, \mathrm{Assertiveness}) = 0.58$
Agreeableness	0.92	$\mathrm{InterCos}(\mathrm{A}, \mathrm{Politeness}) = 0.45$
Neuroticism	0.87	$\mathrm{InterCos}(\mathrm{N}, \mathrm{Cautiousness}) = 0.47$

Table 6: Ability of PVNI to distinguish similar personas.

$\mu_i^{(a)}$ be the direction extracted under prompt variant a. We define:

$$\mathrm{IntraCos}(i) = \mathbb{E}_{a \neq b}\left[\left(\mu_i^{(a)}\right)^\top \mu_i^{(b)}\right].$$

High $\mathrm{IntraCos}(i)$ indicates robustness to prompt surface form. We test three similar persona pairs: Extraversion vs. Assertiveness, Agreeableness vs. Politeness, and Neuroticism vs. Cautiousness.

We observe $\mathrm{IntraCos}(i) \gg \mathrm{InterCos}(i, j)$, which means directions for the same persona are highly consistent across prompt variants shown in Table 6, while similar-but-distinct personas have substantially lower overlap.