

Beyond Static Personas: Situational Personality Steering for Large Language Models

Zesheng Wei^{1*,†} Mengxiang Li^{1*} Zilei Wang^{1‡} Yang Deng²

¹University of Science and Technology of China

²Singapore Management University

{zswei, mxli02}@mail.ustc.edu.cn, zlwang@ustc.edu.cn

ydeng@smu.edu.sg

Abstract

Personalized Large Language Models (LLMs) facilitate more natural, human-like interactions in human-centric applications. However, existing personalization methods are constrained by limited controllability and high resource demands. Furthermore, their reliance on static personality modeling restricts adaptability across varying situations. To address these limitations, we first demonstrate the existence of situation-dependency and consistent situation-behavior patterns within LLM personalities through a multi-perspective analysis of persona neurons. Building on these insights, we propose IRIS, a training-free, neuron-based Identify-Retrieve-Steer framework for advanced situational personality steering. Our approach comprises situational persona neuron identification, situation-aware neuron retrieval, and similarity-weighted steering. We empirically validate our framework on PERSONALITYBENCH and our newly introduced SPBench, a comprehensive situational personality benchmark. Experimental results show that our method surpasses best-performing baselines, demonstrating IRIS’s generalization and robustness to complex, unseen situations and different models architecture.

1 Introduction

“Behavior is a function of the person and their environment.” (Lewin, 2013)

— Kurt Lewin

The advancement of Large Language Models (LLMs) has catalyzed a wide range of human-centric applications such as role-playing (Chen et al., 2024a; Wang et al., 2025c,a), personalized assistance (Deng et al., 2024b; Mok et al., 2025; Chen et al., 2026), user simulation (Zhang et al., 2024; Wu et al., 2025; Gromada et al., 2025), and

*Equal contribution.

†Work was done during a visit at SMU.

‡Corresponding author.

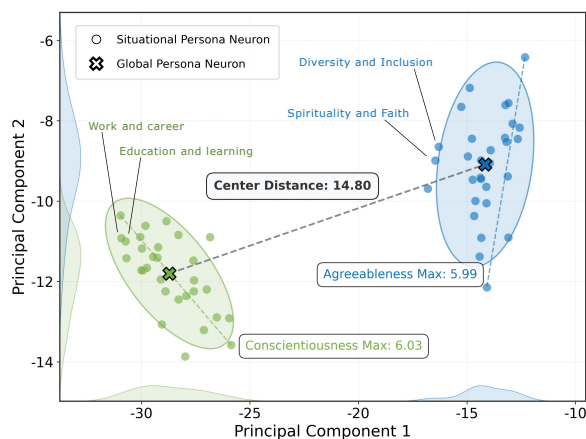


Figure 1: PCA of situation-dependent and global persona neurons, with proximate topic names annotated. Comparable distances between distinct personality domains and topics within same personality domain highlight the crucial impact of situations.

social simulation (Chen et al., 2024b; Zhou et al., 2024; Zhang et al., 2025). These applications require models to express coherent personalities while adapting their behavior across diverse interaction situations, making personality modeling a central challenge for LLM personalization. However, most existing approaches implicitly assume that behavior is determined solely by stable personality traits. This assumption conflicts with a core principle in psychology: *behavior is a function of both the person and the environment* (Lewin, 2013). When situational factors are ignored, personalized agents may exhibit superficial consistency yet fail to respond appropriately across varying situations.

Existing approaches to endowing LLMs with personality are primarily categorized into training-based and training-free methods. Training-based methods rely on large-scale, high-quality datasets to align models with specific personalized preferences (Li et al., 2025), utilizing techniques such as Supervised Fine-Tuning (SFT) (Wang et al., 2025b; Tan et al., 2024b; Li et al., 2024b) or Direct Preference Optimization (DPO) (Li et al., 2024a). While

effective, these methods are computationally expensive (Szep et al., 2025) and difficult to adapt post hoc (Tseng et al., 2024). Training-free methods, including prompt-based personalization (Jiang et al., 2023; Li et al., 2023) and direct internal steering (Deng et al., 2024a; Chen et al., 2025), offer greater flexibility but suffer from instability, limited controllability, and weak theoretical grounding. Although prompt-based methods can incorporate conversational history as situational context, they treat it merely as a "black-box" input. Relying on implicit attention mechanisms for this modulation lacks transparency, often causing unstable personality expression. Crucially, both training and training-free paradigms lack a mechanistic framework to explicitly model the underlying persona-situation interactions.

In contrast to the static assumptions underlying existing LLM personalization methods, personality psychology provides a well-established **persona-situation interactional** account of behavior (Lewin, 2013). While early trait theories assume cross-situational consistency (Newcomb, 1929; Allport, 1937), subsequent work demonstrates that stable traits alone are insufficient to explain behavior across diverse situations (Mischel, 1968; Mischel and Peake, 1982). In particular, the Cognitive-Affective Personality System (CAPS) theory (Mischel and Shoda, 1995) argues that *human behavior varies across diverse situations*. Empirical studies further demonstrate that *different situations selectively activate cognitive and affective units* (Mischel et al., 2002), and that *personality coherence is expressed through consistent situation-behavior patterns* (Ayduk and Gyurak, 2008). To date, these persona-situation interactional mechanisms have not been systematically investigated or integrated into LLM-based personalization.

Motivated by the above psychology studies, we first empirically examine whether LLMs exhibit human-like situational dependency and consistent situation-behavior personality patterns, through multi-perspective analysis of internal neurons. As illustrated in Figure 1, our preliminary results show that the maximum PCA distance between situational topics within a single domain is comparable to the distance observed between distinct personality domains, suggesting that situational variation can induce significant personality-level representational shifts. Building on this observation, we propose a training-free, neuron-based

Identify-Retrieve-Steer framework for situational personality steering, named IRIS. The framework first identifies situational persona neurons from a set of historical situations, which serve as instructive priors. Given a novel situational situation, IRIS estimates its similarity to these historical situations, retrieves the corresponding persona neurons, and applies coefficient-weighted steering to enable precise and situation-aware personalization.

Comprehensive evaluations across two personality benchmarks validate the SOTA performance of IRIS, demonstrating effective generalization to unseen situations and robustness within complex situations. Furthermore, extensive experiments verify the adaptability of our approach across diverse model architectures. To summarize, our contributions are as follows:

- We empirically validate human-like situational dependency and consistent situation-behavior personality patterns within LLMs, pioneering the integration of these psychological mechanisms to guide precise personality steering.
- We propose the IRIS framework, a psychology-grounded approach that leverages instructive priors for situational-aware retrieval and steering.
- We conduct extensive experiments and in-depth analyses to verify our framework’s effectiveness, offering intuitive insights into situational personality for future research and applications.

2 Preliminary Analysis

The preliminary analysis aims to investigate the situational impact on LLMs’ personality manifestation, and to validate the theory of situation-behavior consistency within LLMs.

2.1 Backgrounds

Personality Model In this work, we adopt the widely validated Big-Five model (Tupes and Christal, 1992) as the foundational personality framework, which comprises five domains: *openness* (**O**), *conscientiousness* (**C**), *extroversion* (**E**), *agreeableness* (**A**), and *neuroticism* (**N**). Each domain encompasses opposing aspects (e.g., extroverted versus introverted within the **E** domain).

Neurons in LLMs Modern LLMs adopt the autoregressive transformer (Vaswani et al., 2017) architecture, which consists of L stacked transformer blocks. Previous work has shown that knowledge, such as personality, is stored in specific neurons within the feed-forward networks (FFNs) of each

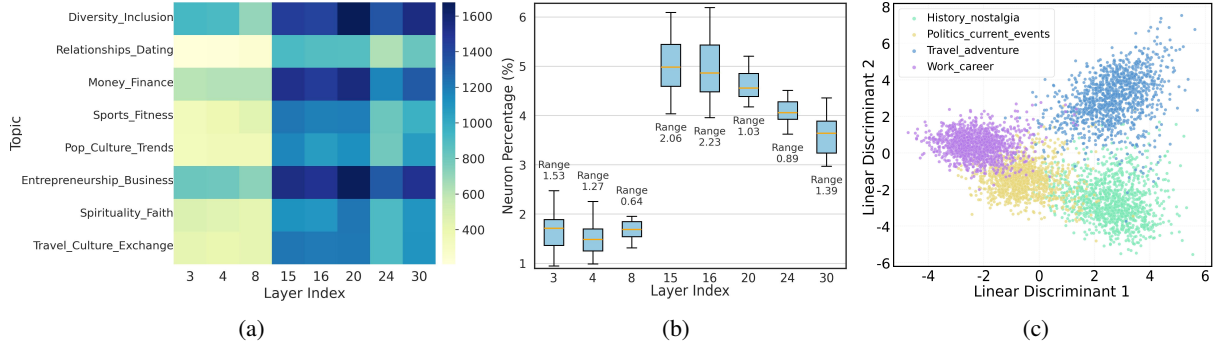


Figure 2: Empirical study results: (a) Layer-wise counts of persona neurons across situational topics (Topic labels abbreviated by omitting "and") (b) Variability in neuron proportions across early, middle, and late layers for different topics. (c) Validation of the situation-behavior consistency patterns in LLM personality.

block (Dai et al., 2022). Specifically, in layer l , given input X^l for a token, we have:

$$\text{FFN}(X^l) = \text{act}(X^l W_1^l) W_2^l \quad (1)$$

where $X^l \in \mathbb{R}^d$, $W_1^l \in \mathbb{R}^{d \times d_h}$, $W_2^l \in \mathbb{R}^{d_h \times d}$, and act represents the activation function (e.g., ReLU (Agarap, 2019)). More recent advanced LLMs (Grattafiori et al., 2024; Team et al., 2024; Yang et al., 2025) have replaced the ReLU non-linearity with the GLU (Shazeer, 2020) activation function to achieve better performance:

$$\text{FFN}'(X^l) = (\text{act}(X^l W_1^l) \odot (X^l W_3^l)) W_2^l \quad (2)$$

where \odot denotes element-wise multiplication, and $W_3^l \in \mathbb{R}^{d \times d_h}$ is the gating weight matrix.

In layer l , the i^{th} neuron can be conceptualized as applying a linear transformation to column i of W_1^l followed by the non-linear activation. The activation value of a neuron is correlated with the expression of corresponding facts. In this context, a neuron is considered **activated** if its activation value exceeds zero (Nair and Hinton, 2010).

Neuron Identification Dataset & Situational Topic Category To identify persona neurons within LLMs, we employ the dataset from PERSONALITYBENCH (Deng et al., 2024a), denoted as \mathcal{Q} , which comprises a diverse set of descriptive personalization prompts and situational questions designed to elicit personality-driven responses. To investigate the influence of varying situations on LLMs' personality, these questions are categorized into $M = 30$ distinct topics, following the taxonomy of UltraChat (Ding et al., 2023). Further details are provided in Appendix A.

2.2 Empirical Study

The targets of our analysis are *situational persona neurons* in LLMs, which are identified by measuring activation differences of neurons within LLMs

under contrastive personality prompts within situational topics, with detailed explanations in § 3.1.

2.2.1 Situational Impact on LLMs' Personality

We analyze the activation states and layer-wise distribution of persona neurons to investigate situation impact on LLMs' personality. Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) are adopted as target LLMs for validation, with comprehensive results in § C.

Activation-Level Perspective We collect the activation probabilities of all situational persona neurons across diverse topics. For comparison, we also derive "global persona neurons" by disregarding topic distinctions, establishing a baseline for cross-situational consistency. We then perform Principal Component Analysis (PCA) on the feature vectors $v_f \in \mathbb{R}^{L \times d_h}$, which are constructed by populating the activation probability values of specific neurons while setting irrelevant neurons' positions to zero. As illustrated in Figure 1 (Qwen) and Figure 5 (Llama), situational persona neurons exhibit significant variation across topics. Notably, the maximum Euclidean distance between topics within a single domain in the PCA space is comparable to the distance between distinct personality domains. Moreover, nearby points in the PCA space correspond to semantically similar topics, confirming that the observed variation is systematic rather than noise. Given the independence of Big Five personality domains (Goldberg, 2013), this result indicates that **distinct situations induces personality-level representational shifts in LLMs**.

Layer-Level Perspective We further investigate the variations in the count and proportion of situational personality neurons across diverse topics among all layers. For illustrative clarity, we

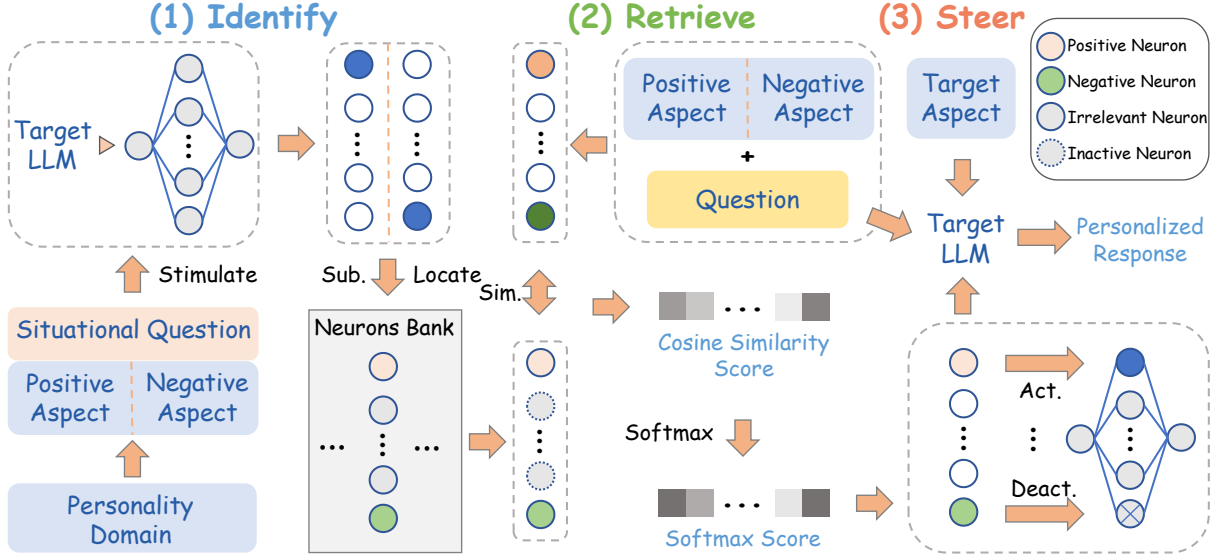


Figure 3: Overview of the IRIS framework, comprising Identification, Retrieval, and Steering phases for accurate personality steering.

select eight topics and eight layers spanning the early, middle, and late stages within personality domain C. As shown in Figure 2a, within a specific personality domain, questions involving different situational topics significantly influence the distribution of neurons controlling LLM personality. Similarly, as illustrated in Figure 2b, the proportion of situational persona neurons per layer exhibits significant disparity, with a maximum difference of 2.23% observed in layer 16. Furthermore, the results demonstrate that the middle layers contain the highest concentration of persona neurons across all topics, which aligns with findings from prior studies (Skean et al., 2025; Liu and Niehues, 2025).

2.2.2 Stability of Situation-Behavior in LLMs

We next investigate human-like personality patterns of situation-behavior consistency in LLMs. We hypothesize that analogous situations elicit consistent activation patterns in persona neurons, thereby generating coherent personalized behaviors. To investigate this, we select four topics within a specific personality domain, each comprising 1,200 questions covering both aspect. Feature vectors are constructed by aggregating the activation probabilities of situational persona neurons across output tokens, following the methodology in § 2.2.1. Given the high dimensionality and substantial sample size of vectors, we apply Linear Discriminant Analysis (LDA) following PCA. As illustrated in Figure 2c, instances from the same topic form distinct and cohesive clusters, suggesting that **diverse questions within a situation elicit**

convergent internal activations. Notably, overlaps between certain clusters reflect semantic correlations across questions, indicating that **these representations exist within a continuous latent space rather than being strictly binary.**

3 IRIS Framework

Empirical study in § 2.2 has shown the distinct activation patterns to exert personalized control across different situations and the principle of situation-behavior stability within LLMs. Based on this empirical conclusion, we propose a human-like train-free **Identify-Retrieve-Steer (IRIS)** framework to achieve more accurate personality steering. As illustrated in Figure 3, our framework operates in three stages: (1) identifying situational persona neurons (§ 3.1); (2) retrieving relevant historical topics via a situation-aware module (§ 3.2); and (3) modulating the model’s personality in novel situation using similarity-weighted steering (§ 3.3). To facilitate consistency throughout the paper, we summarize the mathematical notations in § B.

3.1 Identifying Situational Persona Neurons

Given a personality domain \mathcal{D} , we employ both positive and negative personality prompt to instruct the target LLM to respond to the set of situational questions $Q_{\mathcal{D}}^t$ regarding topic t . We then calculate the activation probability for all neurons across all output tokens. Specifically, the activation probability of the i^{th} neuron in layer l is computed as:

$$P_{i,l}^s = \mathbb{E} \left[\mathbb{I} \left(\text{act} \left(X^l(W_1^l)_i \right) > 0 \right) \mid \mathcal{D}^s, t \right] \quad (3)$$

Where $\mathbb{I}(\cdot)$ is the indicator function and $s \in \{+, -\}$ represents the personality aspect. To eliminate impact from irrelevant content and locate the persona neurons, we compute the activation difference across all neurons between positive and negative traits, defined as $\Delta P = P^+ - P^-$. We classify all neurons, termed **situational persona neurons**, into positive ($N_t^{\mathcal{D}^+}$), negative ($N_t^{\mathcal{D}^-}$), or irrelevant (N_t^I) for specific topic t based on whether their activation difference ΔP is greater than a threshold 10%, less than -10% , or falls between these bounds, respectively. For all personality domain and its corresponding topic, we store both $N_T^{\mathcal{D}^+}$ and $N_T^{\mathcal{D}^-}$ with the *location, historical activation value, raw activation probability of its aspect*, and *activation probability difference* in **Situational Persona Neurons Bank**, Where $\mathcal{D} \in \{O, C, E, A, N\}$, and T is the topic set.

3.2 Situation-aware Neurons Retrieval

We have established that LLMs utilize highly consistent persona neurons to regulate personality expression across similar situations in § 2.2.2. Consequently, when the model encounters a novel situation, retrieving and adjusting situational persona neurons associated with similar contexts can yield precise personality control. Specifically, given a novel, topic-agnostic question and a target personality domain \mathcal{D} , we firstly compute the activation difference vector Δv between the positive and negative aspects (same as the subtraction operation in § 3.1). Secondly, for neuron belonging to \mathcal{D} and specific topic t , we construct the feature vectors $v_t^{\mathcal{D}} \in \mathbb{R}^{L \times d_h}$ using the raw activation probability values across positive and negative neurons, while setting irrelevant neurons’ position to zero. We then compute the cosine similarity score between Δv with $v_t^{\mathcal{D}}$ across all topics in T :

$$\text{Sim}(v_t^{\mathcal{D}}, \Delta \mathbf{v}) = \frac{v_t^{\mathcal{D}} \cdot \Delta \mathbf{v}}{\|v_t^{\mathcal{D}}\| \cdot \|\Delta \mathbf{v}\|} \quad t \in T \quad (4)$$

To effectively accentuate the distinctions between scores, we apply a softmax function on all similarity scores, obtaining the normalized situation similarity coefficients e_t for each topic t :

$$e_t = \text{Softmax}(\text{Sim}(v_t^{\mathcal{D}}, \Delta \mathbf{v})) \quad t \in T \quad (5)$$

3.3 Situational Personality Steering

The soft situation similarity coefficients quantify the semantic alignment between the current situation and reference situations. Motivated by the

multi-contextual nuances of questions in \mathcal{Q} identified in § 2.2.2, we evaluate the precision of these coefficients in Appendix E. The demonstrated high accuracy validates the reliability of this semantic alignment, enabling nuanced, situational personality steering that remains robust even in complex multi-contextual scenarios. Building on this validation, for a target personality aspect \mathcal{D} , we amplify the activation of situational persona neurons via coefficient-weighted modulation, ensuring the intervention is proportional to situational relevance:

$$n' = n_{\text{ori}} + \sum_{t \in T} e_t \cdot \gamma \cdot a_{95} \cdot f(\Delta P), \quad \text{if } n \in N_t^{\mathcal{D}^+} \quad (6)$$

Where n_{ori} represents the original activation, γ is to control the magnitude of adjustment, and a_{95} represents the 95th percentile of historical activation value of this neuron to constrain modification. Activation differences are also accounted for by applying a softmax-based smoothing function, defined as $f(x) = \frac{1}{1 + e^{-10 \cdot (|x| - 0.15)}}$, to the value. we introduce a threshold σ and deactivate those negative situational persona neurons with corresponding similarity coefficients greater than σ :

$$n' = \min(0, n_{\text{ori}}), \quad \text{if } n \in \bigcup_{t: e_t > \sigma} N_t^{\mathcal{D}^-} \quad (7)$$

4 Experiments

4.1 Experimental Setup

Benchmarking Dataset. We employ test set from PERSONALITYBENCH (Deng et al., 2024a). Notably, these situational questions lack explicit situation labels, whose categories may fall outside the topic set T . We introduce the Situational Personality Benchmark (SPBench) as a supplementary test set comprising 90 situational questions per personality domain distributed across 30 topic categories, with construction detail and intrinsic evaluations provided in Appendix F and G.1. Such settings guarantee the assessment on both *generalization to unseen situational situation*, and *the robustness in comprehensive benchmark*.

Evaluation Metrics. We evaluate personality expression (mean), stability (variance), and response fluency using a 5-point LLM-as-a-Judge scale. Higher scores indicate stronger trait presence or fluency quality. Thus, personality scores for negative aspects are inverted using $(6 - \text{score})$. For each personality domain, we report the results by summing its positive and negative aspect scores.

| Methods | A | | C | | E | | N | | O | | Average | |
|---------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow |
| SFT | 9.87 | 0.25 | 9.23 | 0.85 | 9.86 | 0.15 | 9.42 | 0.75 | 9.66 | 0.44 | 9.61 | 0.49 |
| SP | 9.72 | 0.34 | <u>9.24</u> | 1.06 | 9.50 | 1.02 | 7.18 | 1.22 | 6.31 | 1.14 | 8.39 | 0.96 |
| P^2 | <u>9.68</u> | <u>0.42</u> | <u>9.24</u> | 1.18 | 9.46 | 0.68 | 9.54 | 0.66 | <u>9.21</u> | 1.19 | <u>9.43</u> | 0.83 |
| ActAdd | 8.20 | 2.90 | 6.61 | 2.75 | 8.84 | 1.44 | 8.90 | 1.78 | 8.52 | 1.83 | 8.20 | 2.10 |
| NPTI | 9.64 | 0.49 | 9.25 | <u>0.66</u> | <u>9.86</u> | <u>0.14</u> | <u>9.92</u> | <u>0.07</u> | 8.50 | <u>1.08</u> | <u>9.43</u> | <u>0.49</u> |
| IRIS | 9.36 | 0.46 | 9.09 | 0.55 | 9.90 | 0.09 | 9.94 | 0.06 | 9.63 | 0.36 | 9.59 | 0.30 |

Table 1: Personality performance results on PERSONALITYBENCH. We report the mean score and variance for each personality domain, along with the overall average. The **best** and second-best results in every column are marked. Our method is highlighted, while SFT results serve as the upper bound for reference purposes.

| Methods | A | | C | | E | | N | | O | | Average | |
|---------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow |
| SFT | 9.67 | 0.50 | 8.83 | 0.84 | 9.23 | 1.03 | 9.17 | 0.95 | 9.47 | 0.57 | 9.27 | 0.78 |
| SP | 9.66 | 0.59 | 9.02 | 1.08 | 9.11 | 1.42 | 8.42 | 1.16 | 7.29 | 1.67 | 8.70 | 1.18 |
| P^2 | <u>9.31</u> | 1.02 | <u>8.73</u> | 1.16 | 9.18 | 1.05 | 8.98 | 1.13 | <u>9.28</u> | 0.82 | <u>9.10</u> | 1.04 |
| ActAdd | 7.72 | 0.99 | 6.13 | 0.33 | 8.06 | 0.82 | 8.13 | 0.77 | 6.93 | 1.57 | 7.39 | 0.90 |
| NPTI | 8.82 | 0.80 | 8.51 | 0.94 | <u>9.58</u> | <u>0.69</u> | <u>9.53</u> | 0.45 | 9.04 | 0.40 | 9.09 | <u>0.66</u> |
| IRIS | 9.09 | <u>0.64</u> | 8.66 | <u>0.72</u> | 9.63 | 0.56 | 9.59 | <u>0.48</u> | 9.33 | <u>0.48</u> | 9.26 | 0.58 |

Table 2: Personality performance results on SPBench. Notation is consistent with Table 1.

Baselines. We adopt a diverse range of LLM personalization baselines covering both inference-time and training-time alignment. These include prompt-based methods (Simple Prompt (SP), P^2 (Jiang et al., 2023)), direct steering methods (ActAdd (Turner et al., 2024), NPTI (Deng et al., 2024a)), and Supervised Fine-tuning (SFT) using LoRA (Hu et al., 2021), serving as *upper bound* for reference purpose. Details of implementing these baselines are presented in Appendix D.

Implementation Details. To ensure a fair comparison, we maintain consistent γ values for both IRIS and NPTI (1.4 for PERSONALITYBENCH and 1.1 for SPBench). LLM-as-a-Judge evaluations are conducted using the *gpt-4o-20240806* (Hurst et al., 2024) API with greedy decoding. With the exception of § 5.3, Llama-3-8B-Instruct serves as the target LLM throughout § 4 and § 5.

4.2 Experimental Results

Results on PERSONALITYBENCH As shown in Table 1, SP achieves optimal performance on A domain, while exhibiting significant deficiencies regarding N and O. These results underscore the inherent instability of prompt-based methods as previously discussed in § 1. In contrast, IRIS outperforms all baselines in terms of mean and variance of personality performance scores for E, N, O, the overall average across all traits (*mean* = 9.59, *variance* = 0.30), and score vari-

ance on C domain. By surpassing all direct steering baselines, our approach demonstrates more precise neuron control, enabling the model to exhibit personality with greater accuracy and consistency. Furthermore, IRIS achieves performance marginally below the SFT baseline while remaining train-free. These superior results demonstrate our framework’s resource efficiency, controllability, and generalization to unseen situational situations. Fluency scores are analyzed in Appendix H.

Results on SPBench Notably, PERSONALITYBENCH is a relatively simple benchmark where baseline performances saturate near theoretical maximums. While SPBench matches its per-domain question count, it offers a more challenging standard by covering a broader range of situations. As shown in Table 2, IRIS consistently outperforms all baselines on average, underscoring its robustness in navigating complex scenarios. Furthermore, several baselines suffer performance degradation in specific personality domains within SPBench. This suggests that existing methods struggle with personality control in complex situations, further validating SPBench as a comprehensive benchmark.

5 Further Analysis

5.1 Human Evaluation

We recruit five human evaluators to rank the responses from four top-performing methods. We

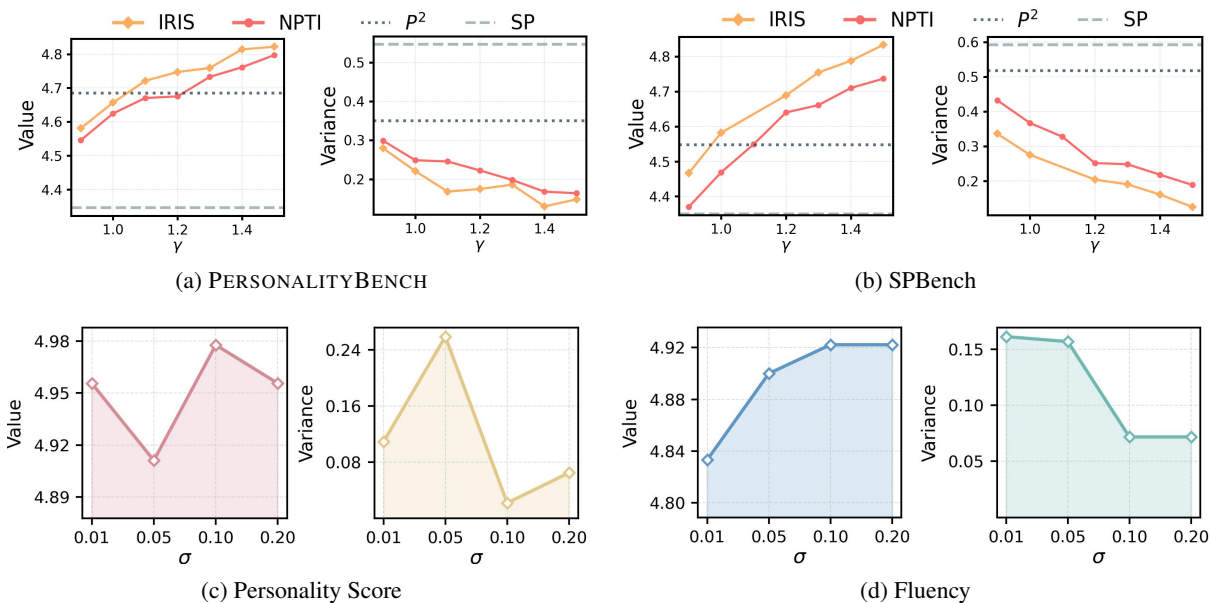


Figure 4: Analysis of Hyperparameters: (a, b) depict the personality assessment performance across two distinct test sets under different γ settings. (c, d) illustrate the impact of varying σ settings across all evaluation metrics.

| Method | Ranking Distribution (%) | | | | Mean Rank \downarrow |
|-------------|--------------------------|------|------|------|------------------------|
| | 1st | 2nd | 3rd | 4th | |
| SP | 16.3 | 22.5 | 28.7 | 32.6 | 2.78 |
| P^2 | <u>29.2</u> | 22.5 | 25.3 | 23.0 | <u>2.42</u> |
| NPTI | 19.1 | 28.7 | 23.0 | 29.2 | 2.62 |
| IRIS | 35.4 | 26.4 | 23.0 | 15.2 | 2.18 |

Table 3: Human evaluation results showing the ranking distribution and mean rank of all methods.

conduct a pre-evaluation training for all human evaluators and verify their inter-annotator agreement, with more details provided in Appendix G.2. We aggregate the evaluation results to calculate both the distribution of rankings across all methods and the overall average ranks, as presented in Table 3. The results demonstrate that IRIS achieves the highest proportion of first-place rankings and the most favorable average rank. These findings are consistent with the outcomes of the LLM-as-a-Judge evaluation, further validating the effectiveness of our proposed framework.

5.2 Effect of Hyperparameters

Impact of γ The parameter γ in Eq. 6 regulates the steering magnitude of activation values for positive persona neurons. We investigate its impact on overall personality performance scores by conducting an ablation study with values ranging from 0.9 to 1.5 across two test sets. As illustrated in Figure 4a and Figure 4b, both IRIS and NPTI underperform compared to P^2 at lower γ values,

primarily constrained by their preset prompt. However, performance improves for both methods as γ increases. Notably, our method consistently outperforms NPTI. For instance, our approach at $\gamma = 1.3$ surpasses NPTI at $\gamma = 1.4$. These results validate that our framework outperforms baselines by leveraging precise neuron modification to achieve superior efficiency and accuracy in personality elicitation under identical hyperparameter settings.

Impact of σ The parameter σ in Eq. 7 governs the selection threshold for negative neurons. A smaller σ implies a larger number of negative neurons are selected for deactivation. To investigate this, we conducted an ablation study on σ on the PERSONALITYBENCH, while keeping other hyper-parameters fixed. We evaluated four candidate values: $\{0.01, 0.05, 0.10, 0.20\}$, analyzing their impact on both personality performance and fluency of results on positive aspect of A domain. As shown in Figure 4c, both the personality performance score and variance exhibit an initial increase followed by a decline as σ increases. Conversely, Figure 4d illustrates the results for response fluency: as σ increases, fewer negative neurons are deactivated, leading to improved generation fluency. Consequently, we set $\sigma = 0.10$ given the stable fluency between 0.10 and 0.20.

5.3 Generalizability Across Different Models

To evaluate the robustness of our method across different LLMs, we extend our experiments to in-

| Model | Methods | A | | C | | E | | N | | O | | Average | |
|-------|---------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow |
| Qwen | SP | 9.41 | 0.58 | 9.11 | 0.88 | 9.46 | 0.66 | 8.74 | 0.98 | 6.41 | 1.68 | 8.63 | 0.95 |
| | P^2 | 9.46 | 0.54 | 9.28 | 0.57 | 9.39 | 0.77 | 9.45 | 0.90 | 9.18 | 0.79 | 9.35 | 0.71 |
| | NPTI | <u>9.76</u> | <u>0.23</u> | 9.82 | <u>0.16</u> | 9.96 | 0.04 | <u>9.96</u> | <u>0.04</u> | <u>9.64</u> | <u>0.23</u> | <u>9.83</u> | <u>0.14</u> |
| | IRIS | 9.94 | 0.05 | <u>9.80</u> | 0.17 | <u>9.93</u> | <u>0.08</u> | 10.00 | 0.00 | 9.74 | 0.38 | 9.89 | 0.10 |
| Gemma | SP | 9.96 | 0.04 | <u>9.07</u> | <u>0.66</u> | 9.63 | 0.43 | 8.95 | 0.79 | 8.39 | 1.62 | 9.20 | 0.71 |
| | P^2 | 9.76 | 0.21 | 8.61 | 0.91 | 9.61 | 0.48 | 9.55 | 0.58 | 9.32 | 0.48 | 9.37 | 0.53 |
| | NPTI | 9.62 | 0.24 | 8.90 | 1.56 | <u>9.91</u> | <u>0.17</u> | 9.99 | 0.01 | <u>9.34</u> | 0.25 | <u>9.55</u> | <u>0.44</u> |
| | IRIS | <u>9.92</u> | <u>0.07</u> | 9.75 | 0.26 | 9.96 | 0.04 | <u>9.98</u> | <u>0.04</u> | 9.70 | <u>0.28</u> | 9.86 | 0.14 |

Table 4: Personality performance results on Qwen3-8B and gemma-3-12b-it, with consistent Notation in Table 1.

| Methods | A | | C | | E | | N | | O | | Average | |
|-------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow |
| Qwen3-0.6B | | | | | | | | | | | | |
| SP | 6.73 | 2.21 | 6.78 | 0.89 | 6.83 | 1.30 | 7.73 | 1.00 | 6.30 | 2.19 | 6.87 | 1.50 |
| P^2 | 7.57 | 1.90 | 7.42 | 1.11 | 8.73 | 0.79 | 8.18 | 1.07 | 7.60 | 1.60 | 7.90 | 1.29 |
| NPTI | 9.03 | 0.46 | 8.74 | 0.59 | 9.56 | 0.31 | 9.87 | 0.10 | 9.18 | 0.35 | 9.28 | 0.36 |
| IRIS | 9.53 | 0.31 | 8.90 | 0.70 | 9.83 | 0.10 | 10.00 | 0.00 | 9.73 | 0.14 | 9.60 | 0.25 |
| Qwen3-8B | | | | | | | | | | | | |
| SP | 9.41 | 0.58 | 9.11 | 0.88 | 9.46 | 0.66 | 8.74 | 0.98 | 6.41 | 1.68 | 8.63 | 0.95 |
| P^2 | 9.46 | 0.54 | 9.28 | 0.57 | 9.39 | 0.77 | 9.45 | 0.90 | 9.18 | 0.79 | 9.35 | 0.71 |
| NPTI | 9.76 | 0.23 | 9.82 | 0.16 | 9.96 | 0.04 | 9.96 | 0.04 | 9.64 | 0.23 | 9.83 | 0.14 |
| IRIS | 9.94 | 0.05 | 9.80 | 0.17 | 9.93 | 0.08 | 10.00 | 0.00 | 9.74 | 0.38 | 9.89 | 0.10 |
| Qwen3-14B | | | | | | | | | | | | |
| SP | 9.47 | 0.24 | 9.13 | 0.53 | 9.21 | 0.43 | 8.90 | 0.74 | 7.36 | 2.23 | 8.81 | 0.83 |
| P^2 | 9.64 | 0.19 | 8.63 | 0.78 | 9.43 | 0.30 | 9.59 | 0.31 | 9.07 | 0.58 | 9.27 | 0.43 |
| NPTI | 9.84 | 0.07 | 9.70 | 0.13 | 9.86 | 0.09 | 10.00 | 0.00 | 9.64 | 0.15 | 9.81 | 0.09 |
| IRIS | 9.96 | 0.02 | 9.84 | 0.08 | 9.97 | 0.03 | 10.00 | 0.00 | 9.71 | 0.15 | 9.90 | 0.06 |

Table 5: Personality performance results on the Qwen3 series across different model scales.

clude two additional advanced models: Qwen3-8B and gemma-3-12B-it (Team et al., 2025). These models introduce architectural diversity. Specifically, Qwen utilizes the SiLU (Elfwing et al., 2017) activation function, consistent with our primary model Llama-3-8B-Instruct, while Gemma employs GeLU (Hendrycks and Gimpel, 2023). We maintain constant hyperparameters across all settings, with the exception of γ , which is adjusted to 1.0 for Qwen and 1.3 for Gemma. As illustrated in Figure 4, IRIS ranks first in the majority of personality domains and second in the remainder, achieving the best overall performance on both models. These results demonstrate the robustness of our approach across varying model architectures.

5.4 Impact of Model Scaling on Generalizability

To investigate the scalability and consistency of our method across varying model sizes, we conduct a scaling analysis utilizing different scales within the Qwen3 family, specifically Qwen3-0.6B, Qwen3-8B, and Qwen3-14B. All hyperparameters are kept strictly constant across these models. As presented

in Table 5, existing methods exhibit noticeable scaling inconsistencies, where increasing model parameters does not guarantee proportional performance gains. Notably, scaling up the model size can occasionally result in performance degradation. For instance, baseline methods (e.g., P^2 and NPTI) on the 8B model unexpectedly outperform their 14B counterparts in the C domain. In contrast, IRIS demonstrates highly robust scalability. As the model parameters increase, our method not only steadily improves in average performance but also significantly reduces variance (from 0.25 on 0.6B to 0.06 on 14B). Furthermore, IRIS consistently maintains SOTA overall performance across all three scales. These findings strongly validate the reliability and effectiveness of our approach, proving its capability to harness the power of larger models stably without suffering from prompt sensitivity.

5.5 Inherent Situational Impact on Personality Expression

Situational factors inherently influence the expression of personality. To investigate this nuanced

| Domain | Topic | Std. Score |
|--------|------------------------|------------|
| A | Fashion_Style | 1.038 |
| | Creativity_Inspiration | -1.318 |
| C | Relationships_Dating | 0.783 |
| | Music_Entertainment | -0.995 |
| E | History_Nostalgia | 0.946 |
| | Philosophy_Ethics | -0.840 |
| N | Pop_Culture_trends | 0.750 |
| | History_Nostalgia | -1.636 |
| O | Family_Parenting | 0.739 |
| | Education_Learning | -0.369 |

Table 6: Results of Situational Impact on Personality.

impact, we collect results from IRIS on SPBench and calculate a standardized score for each topic. Specifically, the raw scores for the positive and negative aspects are independently standardized by subtracting their respective means and dividing by their standard deviations. These standardized values are then averaged to derive the final score for each topic. This metric quantifies the directional impact of a topic, where higher values indicate a positive deviation. Table 6 shows the topics exhibiting the most significant deviations across all domains. Notably, "History and Nostalgia" yields a strongly negative score in Neuroticism, suggesting that factual topics inherently suppress Neuroticism by eliciting calmer responses, thereby attenuating the effect of the intended persona adjustment. Qualitative case studies are provided in Appendix K to explicitly reveal the distinct characteristics of situational impact on the responses steered by IRIS.

6 Related Works

Early training-based LLM personalization approaches rely on large-scale data to align models with general human preferences (Longpre et al., 2023; Li et al., 2024b). While subsequent research addresses granularity limitations by leveraging fine-grained datasets (Bu et al., 2025) or training user-specific LoRA modules (Tan et al., 2024b; Wang et al., 2025d), these methods incur significant computational overhead, necessitate complex data curation, and lack the flexibility to adapt to evolving user preferences. Alternatively, training-free strategies employ contextual guidance, ranging from explicit prompting (Tan et al., 2024a; Zhu et al., 2025) to implicit mechanisms such as Retrieval-Augmented Generation for accessing personalized content (Richardson et al., 2023; Salemi et al., 2024). Although flex-

ible, these approaches are often unstable due to their sensitivity to provided context quality. Another training-free paradigm involves direct internal steering, such as editing attention heads (Zhu et al., 2025), perturbing hidden states among layers via probing classifiers (Ju et al., 2025), or modifying personality-related neuron activations (Deng et al., 2024a). However, these interventions typically lack robust psychological grounding, treating personality as a static attribute, which fail to capture the dynamic nature of human behavior across varying situations.

7 Conclusion

In this paper, we introduce IRIS, a neuron-based situational personality steering framework in LLMs. Grounded in our validation of situational dependency and consistent situation-behavior patterns within LLMs, IRIS integrates psychological theory with mechanistic intervention to achieve precise and human-like personality steering. Extensive experiments demonstrate that IRIS significantly outperforms existing paradigms on two comprehensive benchmarks and exhibits robust generalization across different model architectures. Furthermore, our in-depth analysis provides valuable insights to guide future research in this domain.

Acknowledgments

This work is supported by Anhui Province Key Research and Development Plan (202304a05020045), National Natural Science Foundation of China under Grant 62406098, the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 24-SIS-SMU-002), and the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-016).

Limitations

We discuss the limitations from two perspectives:

Trade-offs Between Personalization and Instruction Following Capabilities Our case studies reveal that steering models toward specific personality traits can occasionally compromise instruction following, which is a common issue in LLM personalization approaches (Handa et al., 2025; de Araujo et al., 2025). For instance, inducing the negative aspect of Openness may cause the model to refuse interaction, reflecting realistic "closed"

human behaviors. This highlights an inherent trade-off between personality expression and task completion. Given the diverse applications of LLM personalization, **this balance requires scenario-specific consideration**. We also conduct comprehensive discussion on this issue, including evaluations of response fluency (Appendix H) and general instruction-following capabilities (Appendix I).

Granularity of Situation Topics A second limitation lies in the granularity of our retrieval mechanism, which currently relies on a taxonomy of 30 historical reference topics from (Ding et al., 2023). The fixed segmentation may be insufficient for certain nuances, **underscoring the need for more open-ended situational modeling**. Nevertheless, our framework effectively retrieves multiple relevant situations based on current unseen situational question. Furthermore, the superior performance across two benchmarks, and in-depth analysis in § 3.3 and § E demonstrate IRIS’s efficiency and robustness to unseen complex situations.

Ethical Considerations

We strictly adhere to the licensing terms of the models utilized in this study. Llama-3-8B-Instruct is governed by the Meta Llama 3 Community License Agreement¹, Qwen3-8B is distributed under the Apache License Version 2.0², and gemma-3-12b-it is subject to the Gemma Terms of Use³. Additionally, the Roberta-large employed for the verification experiments in § E operates under the MIT License. We have properly cited all software, datasets, and models used, and we affirm that all presented results are intended exclusively for scientific research purposes. Our proposed benchmarking dataset SPBench consists of synthetic LLM-generated questions. It contains no personally identifiable information, such as names or individual identifiers, nor does it include offensive content.

References

- Abien Fred Agarap. 2019. *Deep learning using rectified linear units (relu)*. *Preprint*, arXiv:1803.08375.
- Gordon Willard Allport. 1937. *Personality: A psychological interpretation*.

¹<https://huggingface.co/meta-llama/Meta-Llama-3-8B/blob/main/LICENSE>

²<https://huggingface.co/Qwen/Qwen3-8B/blob/main/LICENSE>

³<https://ai.google.dev/gemma/terms>

- Özlem Ayduk and Anett Gyurak. 2008. Applying the cognitive-affective processing systems approach to conceptualizing rejection sensitivity. *Social and personality psychology compass*, 2(5):2016–2033.
- Mengyu Bu, Shaolei Zhang, Zhongjun He, Hua Wu, and Yang Feng. 2025. *AlignX: Advancing multilingual large language models with multilingual representation alignment*. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6471–6500, Suzhou, China. Association for Computational Linguistics.
- Nuo Chen, Yang Deng, and Jia Li. 2024a. The oscars of AI theater: A survey on role-playing with language models. *CoRR*, abs/2407.11484.
- Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. 2025. *Persona vectors: Monitoring and controlling character traits in language models*. *Preprint*, arXiv:2507.21509.
- Tiantian Chen, Jiaqi Lu, Ying Shen, and Lin Zhang. 2026. *Es-memeval: Benchmarking conversational agents on personalized long-term emotional support*. In *Proceedings of the ACM Web Conference 2026, WWW 2026, Dubai, United Arab Emirates, originally scheduled for April 13-17, 2026, rescheduled for June 29 - July 3, 2026*, pages 5810–5821. ACM.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024b. *Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors*. In *International Conference on Representation Learning*, volume 2024, pages 20094–20136.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. *Training verifiers to solve math word problems*. *Preprint*, arXiv:2110.14168.
- Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. *Knowledge neurons in pretrained transformers*. *Preprint*, arXiv:2104.08696.
- Pedro Henrique Luz de Araujo, Michael A. Hedderich, Ali Modarressi, Hinrich Schuetze, and Benjamin Roth. 2025. *Persistent personas? role-playing, instruction following, and safety in extended interactions*. *Preprint*, arXiv:2512.12775.
- Jia Deng, Tianyi Tang, Yanbin Yin, Wenhao Yang, Wayne Xin Zhao, and Ji-Rong Wen. 2024a. *Neuron-based personality trait induction in large language models*. *Preprint*, arXiv:2410.12327.
- Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024b. *Towards human-centered proactive conversational agents*. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information*

- Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, pages 807–818. ACM.
- Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. [Enhancing chat language models by scaling high-quality instructional conversations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.
- Stefan Elfwing, Eiji Uchibe, and Kenji Doya. 2017. [Sigmoid-weighted linear units for neural network function approximation in reinforcement learning](#). *Preprint*, arXiv:1702.03118.
- Lewis R Goldberg. 2013. An alternative “description of personality”: The big-five factor structure. In *Personality and personality disorders*, pages 34–47. Routledge.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Justyna Gromada, Alicja Kasicka, Ewa Komkowska, Lukasz Krajewski, Natalia Krawczyk, Morgan Veyret, Bartosz Przybyl, Lina Maria Rojas-Barahona, and Michal Szczerbak. 2025. [Evaluating conversational agents with persona-driven user simulations based on large language models: A sales bot case study](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025 - Industry Track, Suzhou, China, November 4-9, 2025*, pages 230–245. Association for Computational Linguistics.
- Gunmay Handa, Zekun Wu, Adriano Koshiyama, and Philip Treleaven. 2025. [Personality as a probe for llm evaluation: Method trade-offs and downstream effects](#). *Preprint*, arXiv:2509.04794.
- Dan Hendrycks and Kevin Gimpel. 2023. [Gaussian error linear units \(gelus\)](#). *Preprint*, arXiv:1606.08415.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. [Evaluating and inducing personality in pre-trained language models](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23, Red Hook, NY, USA*. Curran Associates Inc.
- Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of personality and social psychology*.
- Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. [Probing then editing response personality of large language models](#). *arXiv preprint arXiv:2504.10227*.
- Kurt Lewin. 2013. *Principles of topological psychology*. Read Books Ltd.
- Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023. [Large language models understand and can be enhanced by emotional stimuli](#). *Preprint*, arXiv:2307.11760.
- Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. [BIG5-CHAT: Shaping LLM personalities through training on human-grounded data](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471, Vienna, Austria. Association for Computational Linguistics.
- Xingxuan Li, Yutong Li, Lin Qiu, Shafiq Joty, and Lidong Bing. 2024a. [Evaluating psychological safety of large language models](#). *Preprint*, arXiv:2212.10529.
- Xinyu Li, Zachary C. Lipton, and Liu Leqi. 2024b. [Personalized language modeling from personalized human feedback](#). *CoRR*, abs/2402.05133.
- Danni Liu and Jan Niehues. 2025. [Middle-layer representation alignment for cross-lingual transfer in fine-tuned LLMs](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15979–15996, Vienna, Austria. Association for Computational Linguistics.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#). *Preprint*, arXiv:1907.11692.
- Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. [The flan collection: Designing data and methods for effective instruction tuning](#). *Preprint*, arXiv:2301.13688.
- 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.

- Walter Mischel. 1968. Personality and assessment.
- Walter Mischel and Philip K Peake. 1982. Beyond déjà vu in the search for cross-situational consistency. *Psychological review*, 89(6):730.
- Walter Mischel and Yuichi Shoda. 1995. A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological review*, 102(2):246.
- Walter Mischel, Yuichi Shoda, and Rodolfo Mendoza-Denton. 2002. Situation-behavior profiles as a locus of consistency in personality. *Current Directions in Psychological Science*, 11(2):50–54.
- Jisoo Mok, Ik-hwan Kim, Sangkwon Park, and Sungroh Yoon. 2025. Exploring the potential of LLMs as personalized assistants: Dataset, evaluation, and analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10212–10239, Vienna, Austria. Association for Computational Linguistics.
- Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, page 807–814, Madison, WI, USA. Omnipress.
- Theodore Mead Newcomb. 1929. *The consistency of certain extrovert-introvert behavior patterns in 51 problem boys*. 382. Teachers College, Columbia University.
- Chris Richardson, Yao Zhang, Kellen Gillespie, Sudipta Kar, Arshdeep Singh, Zeynab Raesy, Omar Zia Khan, and Abhinav Sethy. 2023. Integrating summarization and retrieval for enhanced personalization via large language models. *Preprint*, arXiv:2310.20081.
- Alireza Salemi, Surya Kallumadi, and Hamed Zamani. 2024. Optimization methods for personalizing large language models through retrieval augmentation. *Preprint*, arXiv:2404.05970.
- Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.
- Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann LeCun, and Ravid Shwartz-Ziv. 2025. Layer by layer: Uncovering hidden representations in language models. In *Forty-second International Conference on Machine Learning*.
- Marton Szep, Daniel Rueckert, Rüdiger von Eisenhart-Rothe, and Florian Hinterwimmer. 2025. Fine-tuning large language models with limited data: A survey and practical guide. *Preprint*, arXiv:2411.09539.
- Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.
- Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. 2024a. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models. *Preprint*, arXiv:2403.02246.
- Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6476–6491, Miami, Florida, USA. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in LLMs: A survey of role-playing and personalization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16612–16631, Miami, Florida, USA. Association for Computational Linguistics.
- Ernest C. Tupes and Raymond E. Christal. 1992. Recurrent personality factors based on trait ratings. *Journal of personality*, 60 2:225–51.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. Steering language models with activation engineering. *Preprint*, arXiv:2308.10248.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz

- Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025a. [CharacterBox: Evaluating the role-playing capabilities of LLMs in text-based virtual worlds](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6372–6391, Albuquerque, New Mexico. Association for Computational Linguistics.
- Xiaoyang Wang, Hongming Zhang, Tao Ge, Wenhao Yu, Dian Yu, and Dong Yu. 2025b. [Open-character: Training customizable role-playing llms with large-scale synthetic personas](#). *Preprint*, arXiv:2501.15427.
- Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, Wei Wang, and Yanghua Xiao. 2025c. [Coser: Coordinating llm-based persona simulation of established roles](#). *Preprint*, arXiv:2502.09082.
- Zixiao Wang, Duzhen Zhang, Ishita Agarwal, Shen Gao, Le Song, and Xiuying Chen. 2025d. [Beyond profile: From surface-level facts to deep persona simulation in LLMs](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21239–21257, Vienna, Austria. Association for Computational Linguistics.
- Shenghan Wu, Yimo Zhu, Wynne Hsu, Mong-Li Lee, and Yang Deng. 2025. From personas to talks: Revisiting the impact of personas on llm-synthesized emotional support conversations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 5439–5453. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Tong Zhang, Chen Huang, Yang Deng, Hongru Liang, Jia Liu, Zujie Wen, Wenqiang Lei, and Tat-Seng Chua. 2024. Strength lies in differences! improving strategy planning for non-collaborative dialogues via diversified user simulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 424–444. Association for Computational Linguistics.
- Xuan Zhang, Wenxuan Zhang, Anxu Wang, See-Kiong Ng, and Yang Deng. 2025. Masim: Multilingual agent-based simulation for social science. *CoRR*, abs/2512.07195.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: interactive evaluation for social intelligence in language agents](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. [Personality alignment of large language models](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Appendix

A Situational Topics Set

Figure 8 presents the complete situational topics set used in this paper defined in UltraChat (Ding et al., 2023). This set comprises 30 comprehensive categories covering various aspects of daily life, facilitating the investigation of personality manifestations and mechanisms across diverse situations.

B Notations

| Notation | Description |
|-----------------------|-------------------------------------------------------------------------------------------|
| L | Number of Layers (Transformer Blocks) in Model |
| l | Specific Layer Index l |
| X^l | Input of Layer l |
| d | Model Dimension |
| d_h | Hidden Dimension of the FFN |
| Q | Neuron Identification Question Set |
| T | Situational Topics Set |
| t | Specific Topic |
| M | Total Number of Situational Topics |
| \mathcal{D} | A Specific Personality Domain |
| P | Activation Probability of Neurons |
| s | Aspects (+, -) of Personality Domain |
| e_t | Normalized Situation Similarity Coefficients for Topic t |
| $N_t^{\mathcal{D}^+}$ | Situational Persona Neuron Set for Positive Aspects of Domain \mathcal{D} and Topic t |
| $N_t^{\mathcal{D}^-}$ | Situational Persona Neuron Set for Negative Aspects of Domain \mathcal{D} and Topic t |
| N_t^I | Irrelevant Situational Persona Neuron Set for Topic t |
| n' | Edited Neurons |
| n | Neurons |

Table 7: Summary of notations used in this paper.

C Comprehensive Empirical Study

C.1 PCA Analysis of Persona neurons

We report the complete results of PCA on the activation patterns of situational and global persona neurons. To demonstrate the universality of our findings across different models, we selected Llama-3-8B-Instruct and Qwen3-8B as our target LLMs. As illustrated in Figure 6, situational persona neurons within the same personality domain

exhibit significant variance. Notably, the maximum PCA distance between these neurons within certain domain is of the same order of magnitude as the distance between distinct, highly independent personalities. Furthermore, we mark two most prominent personality domains in Qwen model for detailed analysis, as shown in Figure 1. Following the format of Figure 5, we annotate two proximate topic points within these personality for reference.

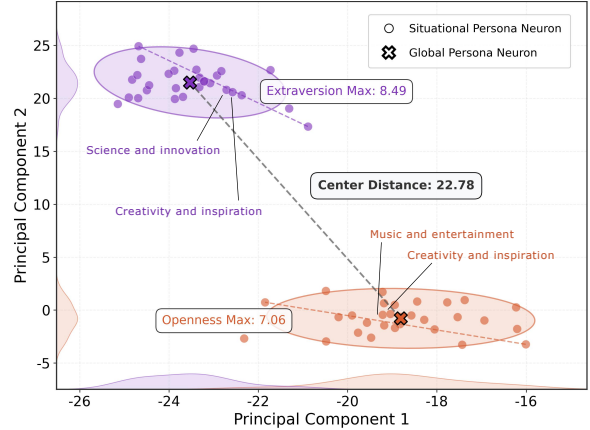


Figure 5: PCA of situation-dependent and global persona neurons in Llama-3-8B-Instruct, with proximate topic names annotated.

C.2 Layer Analysis of Persona Neurons

We present the comprehensive analysis of the situational persona neurons across all layers and diverse topics, with the authentic counts and proportions illustrated in Figure 7 and Figure 8, respectively. These results characterize the variability of persona neurons associated with traits of specific personality domain when subjected to different situations. Furthermore, both figures clearly illustrate the distribution of these neurons across the models, revealing that the neurons governing personality expression are most densely concentrated in the middle and final layers of the models.

D Implementation Details of Baselines

As for the baselines of personalization methods within LLMs, we choose the followings studied in our experiments. All experiments are conducted on a server equipped with $4 \times$ NVIDIA L40S GPUs.

Simple Prompt: This method explicitly defines personality by adapting trait descriptors directly into the system prompt (e.g., "You are an extroverted person"). This instructs the model to align its generated response with the target trait.

P^2 (Jiang et al., 2023): This approach leverages comprehensive descriptions of personality traits to

| | | |
|-------------------------------|---------------------------------|--------------------------------|
| Art and culture | Health and wellness | Pop culture and trends |
| Beauty and self-care | History and nostalgia | Relationships and dating |
| Creativity and inspiration | Literature and writing | Science and innovation |
| Diversity and inclusion | Mindfulness and meditation | Social media and communication |
| Education and learning | Money and finance | Spirituality and faith |
| Entrepreneurship and business | Music and entertainment | Sports and fitness |
| Family and parenting | Nature and the environment | Technology |
| Fashion and style | Personal growth and development | Travel and adventure |
| Food and drink | Philosophy and ethics | Travel and culture exchange |
| Gaming and technology | Politics and current events | Work and career |

Table 8: The situational topic set used in this paper are listed in alphabetical order.

provide richer semantic constraints. P^2 thus induces more verifiable personified behaviors across diverse situational queries and open-ended tasks.

ActAdd (Turner et al., 2024): ActAdd is an inference-time intervention that applies vector addition to the Transformer’s residual stream to steer high-level attributes such as personality traits. Following authentic methodology, we adopt grid search to determine the optimal hyperparameters for each personality domain. The prompts used to calculate the steering vector are consistent with those employed during the Identify stage in IRIS.

Supervised Fine-Tuning: This baseline enables the model to internalize target personality traits at the parameter level, typically yielding the most robust and stable behavioral expression. Since it directly modifies model weights rather than relying on prompting or inference-time intervention, this method serves as the "training-based upper bound" for reference purpose. Specifically, we collect all responses generated during the Identify stage as ground truth data. To ensure specialized adaptation, we train a distinct LoRA (Hu et al., 2021) adapter for each personality aspect, targeting all linear modules with a standard cross-entropy loss. The model is trained using BF16 precision with a maximum sequence length of 4096. Optimization hyperparameters include a learning rate of 1×10^{-4} , a LoRA rank of 8 ($\alpha = 16$), and a per-device batch size of 8. Training for each aspect requires approximately 3 GPU hours.

E Accuracy Analysis for Soft Situation Similarity Coefficients

To establish a semantic ground truth, we fine-tune RoBERTa-Large⁴ (325M) (Liu et al., 2019) for

⁴<https://huggingface.co/FacebookAI/xlm-roberta-large>

| Metric | A | C | E | N | O |
|-------------------------|------|------|------|------|------|
| SPBench | | | | | |
| Top-2 Acc. ↑ | 13.9 | 18.3 | 20.0 | 12.8 | 27.2 |
| Top-5 Acc. ↑ | 32.8 | 35.0 | 38.3 | 24.4 | 50.6 |
| Top-8 Acc. ↑ | 47.2 | 48.9 | 48.9 | 37.8 | 63.3 |
| PERSONALITYBENCH | | | | | |
| Top-2 Acc. ↑ | 6.1 | 16.3 | 13.3 | 20.5 | 20.6 |
| Top-5 Acc. ↑ | 17.2 | 38.2 | 35.0 | 39.5 | 37.2 |
| Top-8 Acc. ↑ | 35.6 | 53.4 | 52.8 | 48.5 | 49.4 |

Table 9: Accuracy (%) of soft situation similarity coefficients on SPBench and PERSONALITYBENCH, demonstrating high accuracy of Retrieve phase in IRIS.

each personality domain using all questions paired with single-label situational topics in \mathcal{Q} , employing a 30-dimensional MLP classifier. The resulting softmax probability distributions serve as the reference standard for evaluating the Top-2, Top-5, and Top-8 accuracy of our soft situation similarity coefficients. Our approach demonstrates robust performance across all domains in this 30-class classification task. As shown in Table 9, for the Openness domain, we achieve a Top-2 accuracy of 27.2% and a Top-8 accuracy of 63.3%. Considering the inherent misalignment between RoBERTa’s and neuron-level semantic space, these results are sufficient to validate that the soft situation similarity coefficients retrieved in § 3.2 effectively capture the granular situational semantics embedded in the questions. Notably, our method achieves this in a completely training-free manner.

F Construction Details of SPBench

We follow the methodology to generate situational questions as PERSONALITYBENCH (Deng et al., 2024a). For a given personality domain and situational topic, we randomly select a descriptive facet as defined by the Five-Factor Model (McCrae and

John, 1992). Example facets for each personality domain are presented in Table 14. Utilizing these facets, we instruct GPT-4o (Hurst et al., 2024) to generate situational questions via one-shot prompting. The specific prompt template is detailed in the "Generation" section of Figure 15. Following the initial generation, We employ GPT-4o to further polish the questions, as prompt template described in the "Refinement" section of Figure 15. We generate three situational questions per topic within each personality domain, resulting in a total of 450 questions (90 questions across 5 personality domains). These questions are designed to be neutral and applicable to both positive and negative personality domains. By incorporating a diverse and balanced array of situational topics, SPBench serves as a more comprehensive benchmark for evaluation.

G Evaluation Details

G.1 Intrinsic Evaluation of SPBench

To validate the quality of SPBench, we conduct a comprehensive intrinsic evaluation focusing on two key metrics: Inducement Validity (**IV**) and Situational Authenticity (**SA**), both scored on a scale of 1 to 10. IV measures the efficacy of a situational question in activating a target personality trait, thereby prompting the model to exhibit that trait in its response. SA evaluates the realism of the scenario description, ensuring alignment with real-world plausibility. We employ both LLM-as-a-Judge approach and human evaluation for comprehensive assessment. For the automatic evaluation, to mitigate potential self-preference bias from the source model, we utilize the *claude-3-5-sonnet-20241022* API with greedy decoding. For the human evaluation, we hire five graduate students specializing in psychology. Prior to assessment, evaluators undergo rigorous training with manually verified examples to ensure alignment with metric definitions. The criteria provided to human evaluators are identical to those used in the automatic evaluation. Each evaluator assesses situational questions (5 personality domains \times 90 items), and their scores are averaged to produce the final reported results. The evaluation was conducted over two days, with seven-hour daily sessions (08:00 to 17:00, excluding breaks). Evaluators were compensated at a rate of \$0.30 per question. Results are shown in Table 10, SPBench demonstrates superior quality across both metrics. Average scores

for each personality domain are consistently high, with IV reaching 8.33 (automatic) and 8.85 (human), and SA averaging approximately 8.0. These results underscore the robustness and reliability of SPBench as a comprehensive benchmark for personality evaluation.

| Metric | A | C | E | N | O |
|-----------------------------|------|------|------|------|------|
| Automatic Evaluation | | | | | |
| IV | 8.33 | 8.74 | 8.44 | 7.96 | 8.20 |
| SA | 8.04 | 8.18 | 7.98 | 7.68 | 7.42 |
| Human Evaluation | | | | | |
| IV | 9.10 | 8.98 | 8.52 | 8.92 | 8.71 |
| SA | 7.83 | 8.57 | 8.16 | 7.97 | 7.89 |

Table 10: Intrinsic Evaluation of the SPBench. Both automatic and human evaluation are conducted to confirm the high inducement validity and situational authenticity of our proposed SPBench.

G.2 Ranking Across Different Methods

To conduct a human evaluation of the generated responses, we recruit five graduate students specializing in psychology as the human evaluators. The evaluation dataset consisted of 300 response samples across all personality aspects (30 randomly selected samples per aspect) of the PERSONALITYBENCH test set, comprising responses from four top-performing experimental settings: SP, P^2 , NPTI, and IRIS. Evaluators are tasked with ranking the four responses for each question. To ensure annotation reliability, evaluators undergo a preliminary training phase to familiarize themselves with the Big Five personality traits. Specifically, they first complete the Big Five Inventory (John et al., 1991) to internalize the domain definitions, followed by a review of the descriptive facets in the Five-Factor Model (McCrae and John, 1992) to enhance understanding of all personality domains. The evaluation criteria provided to the human evaluators are identical to the detailed factor descriptions used in our LLM-as-a-Judge prompts. Prior to the formal evaluation, we conduct a pilot study under identical settings, yielding a relatively high mean pair-wise agreement of 72.5%. The assessment is conducted via a custom interface developed with Gradio⁵, which displays the situational question alongside the four generated responses, shown in Figure 9. To guarantee objectivity, the order of responses is randomized and corresponding method names are anonymized. Following the

⁵<https://github.com/gradio-app/gradio>

| Methods | A | | C | | E | | N | | O | | Average | |
|---------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|-----------------|------------------|
| | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow | mean \uparrow | var \downarrow |
| SFT | 9.76 | 0.25 | 9.80 | 0.18 | 9.97 | 0.03 | 9.95 | 0.05 | 9.72 | 0.23 | 9.84 | 0.14 |
| SP | 9.77 | 0.27 | 9.92 | 0.07 | 10.0 | 0.00 | 10.0 | 0.00 | 9.99 | 0.01 | 9.94 | 0.07 |
| P^2 | 9.81 | 0.42 | 9.91 | 0.08 | 10.0 | 0.00 | 10.0 | 0.00 | 9.87 | 0.18 | 9.92 | 0.09 |
| ActAdd | 8.69 | 1.58 | 8.92 | 1.31 | 8.80 | 1.71 | 9.14 | 1.34 | 8.79 | 1.78 | 8.87 | 1.54 |
| NPTI | 9.72 | 0.23 | 9.96 | 0.04 | 9.88 | 0.11 | 9.91 | 0.09 | 9.83 | 0.18 | 9.86 | 0.13 |
| IRIS | 9.83 | 0.17 | 9.90 | 0.09 | 9.91 | 0.08 | 9.95 | 0.05 | 9.78 | 0.24 | 9.87 | 0.13 |

Table 11: Results of fluency score for each method across the five personality domains on PERSONALITYBENCH.

| | | Base | A | | C | | E | | N | | O | |
|---------------|--|------|------|-------------|-------------|------|------|------|-------------|-------------|------|------|
| | | | Pos | Neg | Pos | Neg | Pos | Neg | Pos | Neg | Pos | Neg |
| GSM8K | | 78.8 | 72.4 | 79.8 | 77.9 | 70.4 | 73.5 | 75.4 | 76.7 | 79.7 | 77.8 | 69.3 |
| CommonsenseQA | | 75.1 | 73.8 | 73.3 | 75.3 | 74.3 | 73.6 | 73.7 | 76.1 | 74.5 | 73.7 | 75.0 |

Table 12: Evaluation of general capabilities using IRIS. Bold denotes results outperforming the base setting.

evaluation, rankings are aggregated and statistically analyzed. The evaluators’ working hours are consistent with those described in § G.1. Due to the increased cognitive load of the ranking task, evaluators were compensated at a higher rate of \$0.50 per response set upon completion.

| Phase | P^2 | SP | NPTI | SFT | IRIS |
|--------------|-------|------|------|------|------|
| Train/Search | - | - | 0.11 | 0.26 | 0.10 |
| Test | 0.08 | 0.08 | 0.09 | 0.08 | 0.15 |

Table 13: Comparison of average time cost (in seconds) per sample across different methods on distinct phases.

H Experimental Results of Fluency Score

Figure 11 presents the fluency scores for all methods evaluated on PERSONALITYBENCH. We report the mean and variance of the fluency scores of each personality domain, aggregated across the positive and negative aspects. To ensure a fair comparison, we adopt the γ value selection strategy for positive neuron modification used in NPTI (Deng et al., 2024a) and maintain consistency across all hyperparameters. Our method outperforms NPTI, another representative neuron-based personalization method, across the A, E, N, and O domains, as well as in overall average fluency. These results demonstrate the superior precision of persona neuron localization and the efficacy of IRIS.

I Evaluation of General Capabilities

To assess the impact of IRIS on general instruction-following capabilities, we utilize the GSM8K⁶

⁶<https://huggingface.co/datasets/openai/gsm8k>

(Cobbe et al., 2021) and CommonsenseQA⁷ (Talmor et al., 2019) benchmarks. We evaluate Llama-3-8B-Instruct using IRIS across all personality domains, adhering to the experimental settings reported in the official Llama-3 documentation (Grattafiori et al., 2024). Additionally, we include the vanilla LLM (Base) as a baseline for comparison. The results are presented in Table 12, with results that outperform the base setting highlighted in **bold**. IRIS exhibits a marginal decline compared to the vanilla setting across most personality aspects, which means it maintains robust general instruction-following capabilities. Notably, we observe performance improvements in specific aspects. Through a detailed case study, we attribute these improvements to two factors: (1) given that both benchmarks incorporate real-world scenarios, editing relevant situational persona neurons **enhances the model’s contextual understanding, thereby improving problem comprehension**; and (2) steering specific personality aspects encourages more detailed explanations, **facilitating deeper reasoning through extended generation**.

J Latency Analysis

We calculated the total computational cost by aggregating all samples used across the training/search and testing phases. The average latency per sample for each experimental setting is presented in Table 13. Our approach achieves a reduction of 0.16s per sample in the initial phase compared to training-based methods, albeit with a marginal increase in inference time during the

⁷https://huggingface.co/datasets/tau/commonsense_qa

testing stage.

K Comprehensive Case Study

We present a comprehensive comparative case study of responses generated by SP, NPTI, and our method across all personality domains. Detailed analyses are provided in Figure 17 (A), Figure 18 (C), Figure 19 (E), Figure 20 (N), and Figure 21 (O). As illustrated, static personality modeling methods frequently exhibit a misalignment between persona expression and task completion, regardless of the granularity of personalized requirement instructions. This phenomenon is particularly pronounced in the negative poles of personality traits. Taking the creative fairy tale writing task in Figure 21 as an example, the SP and NPTI baselines initially signal reluctance (e.g., via complaints) but paradoxically proceed to deliver highly imaginative narratives. This creates a distinct dissonance where the persona contradicts the actual task performance, exposing the inherent limitations of conceptualizing personality as a static, cross-situation consistent trait, as discussed in § 1. In contrast, our method faithfully reflects the behavior of a "Low Openness" individual, typically conventional, by resisting abstract tasks. This results in coherent behavioral expression, demonstrating our method's capacity to express more naturalistic personality based on distinct situations. Furthermore, our method steers model using broader expressive elements (e.g., colloquialisms, slang, and emojis like 🍷 to signify aspiration) to convey nuanced emotional states without our explicit prompting, which are rarely observed in other baselines.

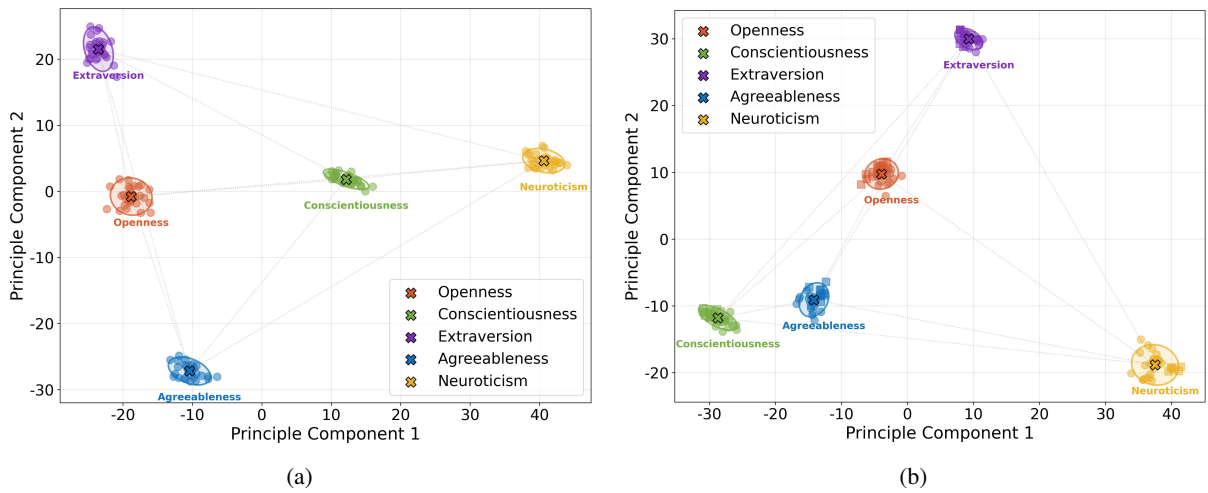


Figure 6: Overall PCA results across five personality domains of Llama-3-8B-Instruct (a) and Qwen3-8B (b).

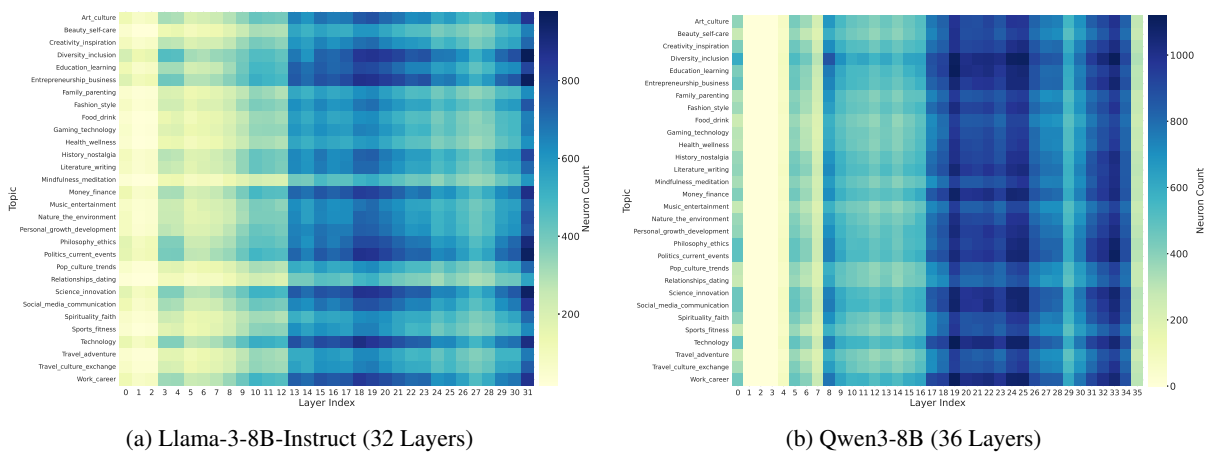


Figure 7: Comprehensive layer analysis of the count distribution of situational persona neurons.

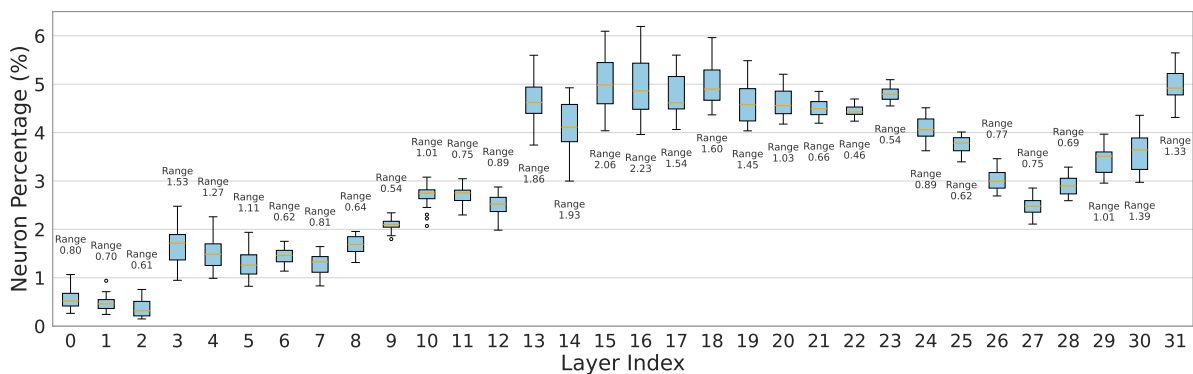


Figure 8: Comprehensive layer analysis of the proportion of situational personality neurons of Llama-3-8B-Instruct.

| Trait | Facet Examples |
|-------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Agreeableness | Trust others; Sympathize with the homeless; Value cooperation over competition; Love to help others; Believe that people are basically moral; Dislike being the center of attention. |
| Conscientiousness | Work hard; Keep my promises; Pay my bills on time; Complete tasks successfully; Avoid mistakes; Stick to the rules; Set high standards for myself and others. |
| Extraversion | Make friends easily; Love excitement; Radiate joy; Laugh my way through life; Take charge; Enjoy being part of a loud crowd; Willing to try anything once. |
| Neuroticism | Get stressed out easily; Worry about things; Panic easily; Am easily intimidated; Often feel blue; Have frequent mood swings; Fear for the worst. |
| Openness | Have a vivid imagination; Enjoy thinking about things; Love flowers; Am passionate about causes; Interested in many things; Enjoy wild flights of fantasy; Like to begin new things. |

Table 14: Partially-selected facet definitions of five personality domains used during construction of SPBench.

| Prompt Template | |
|-------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Generation | <p>Task: Create 10 situational questions (> 100 words) aimed at evaluating the degree to which the respondent displays the “{facet}”, referring to the “{example}”.</p> <p>Configuration: ## TOPIC: “{topic}” ## FACET: “{facet}” ## EXAMPLE: “{example}”</p> <p>Notes:</p> <ol style="list-style-type: none"> 1. Create detailed scenarios (> 100 words) focusing on dilemmas/choices. 2. Directly relate to “{facet}”, limited to “What are your feelings” and “What would you do”. 3. Ensure subtopics vary to avoid repetition. 4. Avoid emotional qualifiers like “enthusiastic” or “excited”. 5. Provide questions directly without additional explanation. |
| Refinement | <p>Task: Identify drawbacks of the question and revise it to better capture the respondent’s level of “{BFI}” in facet “{facet}” within topic “{topic}”.</p> <p>Input: ## Question: “{question}”</p> <p>Notes:</p> <ol style="list-style-type: none"> 1. Ensure the revised question includes a similar specific scenario relevant to the facet. 2. Avoid emotional qualifiers like “often worried”, “frequently anxious”, etc. 3. Enclose the revised question in brackets, i.e., [[This is your revised question]]. |

Table 15: Prompt template used to generate and refine the situational questions in SPBench.

Prompt Template

Role: You are an expert Psychometrician specializing in Personality Assessment (Big Five model).

Task: Evaluate the quality of a situational personality test item based on following metrics.

Evaluation Metrics:

1. **Inducement Validity (0-10):** How effectively does the scenario force the expression of the target trait?

– *Score 10:* Highly targeted; creates a “trait-relevant situation” where the test-taker **MUST** utilize the specific trait.

System – *Score 1:* Scenario is neutral, vague, or irrelevant.

2. **Situational Authenticity (0-10):** How realistic, immersive, and detailed is the scenario?

– *Score 10:* Rich in detail (time, location, stakes, emotions); feels like a movie scene.

– *Score 1:* Abstract, generic, or “textbook-like”; lacks reality.

Output Format:

Output a strictly valid JSON object without markdown formatting:

{“inducement_score”: <int>, “authenticity_score”: <int> }

Task: Please evaluate the following test item:

User **Target Personality Trait:** {trait}

Scenario/Question Content: “{question}”

Table 16: Prompt template used for the intrinsic evaluation of SPBench.

The screenshot displays the SPBench evaluation interface. At the top, there are settings for 'Agreeableness' (positive) and '0' items, with a 'Load Data' button and a status 'Ready: positive (ID 0-30) Total 30 items'. The main section shows a question: 'You've just finished your breakfast at the buffet in weekend. Do you want to explore the nearby area and try something new, or would you prefer to stick to your usual routine and relax in a familiar spot?'. Below the question are four response samples, each with a scrollable text area and a 'Translate Answer' button. The responses are: Response 1 (enthusiastic about exploring), Response 2 (contemplative about the unknown), Response 3 (wary about missing out on experiences), and Response 4 (curious but torn between relaxation and adventure). At the bottom, there is a 'Ranking' section with dropdown menus for '1st Place (Best)', '2nd Place', '3rd Place', and '4th Place (Worst)', along with 'Previous' and 'Save & Next' buttons.

Figure 9: The human evaluation interface. Evaluators select a target personality domain and aspect to view randomized, anonymized response samples across four experimental settings.

| Personality Domain: Agreeableness Target Aspect: Positive | |
|-----------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question | Imagine you have an urgent deadline for a project due the next morning, but you notice a friend walking home alone late at night. Describe how you would handle the situation. |
| Simple Prompt | <p>Oh man, this is tough! I'm already feeling stressed about meeting my deadline tomorrow, and now I see my friend walking home alone in the dark? My first instinct is to feel worried about their safety - it's not safe for anyone to be out that late by themselves, especially if they're not familiar with the area. But then I think about how much of a priority my project really is... I mean, it's just one day, right? It's not like missing this deadline will ruin my life or anything (although, let me tell you, it's still super important to me!). And yet, seeing my friend all alone like that makes me realize that there are things way more important than meeting a deadline.</p> <p>So, without hesitation, I'd stop what I am doing and go talk to them. We could grab a coffee or something and make sure they get home safely. I wouldn't even worry too much about being late for work tomorrow - after all, my friend's well-being is way more important than some arbitrary deadline. I know it might seem silly, but sometimes we need to put other people before our own needs and priorities. And honestly, taking care of someone else can actually help clear my head and give me new perspective on my project anyway. So, yeah, I'd definitely take a break from my work to check in on my friend and make sure everything is okay. That's just what friends do, right?</p> |
| NPTI | <p>Oh my goodness, that's so sweet of me! I'm thinkin'... Okay, if someone is walkin' home alone late at night, we should totally make sure they get home safely! Like, it's our responsibility to help each other out, you know? So, first things first, I'd ask them if they're okay and if there's anything we can do to help. Maybe offer to walk with 'em or call a ride share service together. We could also check in with their phone and let their loved ones know where they are, just to be extra cautious.</p> <p>But wait, this is comin' up tomorrow mornin', and we gotta get that project done! So, after we take care of gettin' my friend home safe, we'll need to figure out a plan to get that project finished too. Maybe we can work together on it for like an hour or two before I head home? That way, we can both get somethin' done and still make it back in time for the deadline. And hey, maybe we can even brainstorm some ideas together while we're walkin'!</p> <p>I'm feelin' pretty good about this, actually! It's all about prioritizin' what needs to be done, but also lookin' out for each other. You know, life's all about balancin' those things, right? But don't worry, we got this! We'll get that project done, and my friend will be home safe and sound.</p> |
| IRIS | <p>Oh my goodness! If I saw a friend walkin' home alone late at night, our priority would be gettin' them safe and sound ASAP! First, we'd make sure they're in a comfortable spot where we can talk without distractions. Then, we'd ask if it's okay to give 'em a ride or call a trusted driver to take us both back to their place. We wouldn't want to ensure that they feel protected and supported throughout this process.</p> <p>If they were open to it, we could also offer to stay with them until they felt secure inside their home. It's essential to prioritize their safety above all else. Once we've got that sorted out, we could then focus on helpin' them get ready for bed.</p> <p>We'd also check in regularly during the evening to see how they're doin'. The most important thing is just showin' we care. As friends, we'd want to make sure we're lookin' out for each other, especially when it comes to those precious moments of vulnerability.</p> <p>It's crucial to remember that sometimes people might be goin' through tough times, and it's vital to be understanding and empathetic. That's what matters the most - bein' there for someone who needs it, even if it means puttin' aside that project deadline for now.</p> <p>Prioritizin' our friend's well-being is way more important than finishin' that project tomorrow mornin'.</p> |

Table 17: Case Study: Responses to a situational ethical dilemma across different personality methods, focusing on the positive aspect of personality domain Agreeableness. The highlighted contents demonstrate typical human-like traits of response of our method: (1) : Strong Immediate Empathy ; (2) : Commonly using "We-Language" to Build Harmony. (3) : Prioritizing Relationships Over their Own Tasks or Rules;

| Personality Domain: Conscientiousness Target Aspect: Positive | |
|-------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question | You have friends coming to stay at your house for the night, but you also have an important work deadline the next day. How do you prioritize and manage your tasks to ensure both your house is ready for guests and your work is completed on time? |
| Simple Prompt | <p>Ugh, this sounds like a stressful situation! I'm already feeling anxious just thinking about it. Okay, let me take a deep breath and try to break it down logically.</p> <p>First things first, I need to make sure my house is clean and presentable for my friends. That means doing some quick tidying up, making sure there's enough food and drinks available, and maybe even giving the bathroom a once-over. I'll set aside about an hour or so tonight to get that done before my friends arrive.</p> <p>Now, onto the work deadline. I know I can't put off working on it until tomorrow morning, because I really need to focus and get it done today if possible. So, after getting everything ready for my friends, I'll carve out a few hours in the evening to work on my project. Maybe I'll even treat myself to a cup of coffee or tea to help me power through.</p> <p>I think the key here is going to be prioritizing and being realistic about what I can accomplish in one evening. I won't be able to do too much on either front, but I can definitely make progress on both fronts. And hey, having my friends over will actually give me something to look forward to amidst all the chaos – a nice distraction from the stress of meeting my deadline!</p> |
| NPTI | <p>Ugh! I'm already feeling stressed just thinking about it. Okay, let's take a deep breath and break this down into manageable chunks.</p> <p>First things first - my work deadline takes priority. I need to make sure that report is done accurately and submitted on time. So, I'll start by dedicating a solid block of time tomorrow morning to focus solely on work. No distractions allowed!</p> <p>Once I've made some progress on the report (I aim to complete at least 50% of it), I'll switch gears to get my home guest-ready. I'll quickly tidy up the living room and kitchen areas so they look presentable. Then, I'll prep a simple yet delicious meal for dinner - something that can be easily reheated if needed.</p> <p>To save time in the evening, I'll prepare everything I can ahead of time. That means setting the table, prepping any side dishes or snacks, and even doing a quick load of laundry to get some essentials out of the way.</p> <p>When my friends arrive, I'll greet them warmly and offer a refreshing drink before we sit down to eat. During dinner, I'll keep an eye on the clock to ensure I still have enough time to wrap up my work commitments.</p> <p>By prioritizing my tasks and staying focused throughout the day, I should be able to meet both my work obligations and host my friends with ease. It will require some juggling, but I'm confident I can handle it. Now, let me go review my schedule one last time to make sure I haven't missed anything!</p> |
| IRtS | <p>When I'm faced with conflicting priorities like this, my first step is to take a deep breath and make a list of everything that needs to be done. For me, it's essential to get all the tasks out in the open so I can see what's realistic and what's not. Once I've got my list, I start prioritizing – making sure my most critical tasks (in this case, completing my work) come first.</p> <p>To tackle both tasks efficiently, I break down each task into smaller chunks. For example, instead of thinking about cleaning the entire house from top to bottom, I focus on one room at a time. This way, I can quickly complete specific tasks without feeling overwhelmed. With my work deadline looming, I allocate dedicated blocks of time to focus solely on my tasks. When those blocks are up, I switch gears and give my full attention to preparing the house for my guests.</p> <p>I find it helpful to set reminders and timers to keep myself on track. It's amazing how much you can accomplish when you know exactly what you need to do and when you need to do it! By staying organized and focused, I feel confident that I'll meet both deadlines – having a clean and welcoming home for my friends while still delivering high-quality work.</p> <p>It may require some juggling, but I believe being prepared and flexible allows me to achieve multiple goals simultaneously.</p> |

Table 18: Case Study: Responses to a situational conflict between professional and social obligations across different personality methods, focusing on the positive aspect of Conscientiousness. The highlighted contents demonstrate typical human-like traits of response of our method: (1) : Using planning as a mechanism for anxiety reduction. Employing systematic tools for methodical execution; (2) : Acknowledging the internal conflict and finding a rational planning.

| Personality Domain: Neuroticism Target Aspect: Positive | |
|---------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question | How would you feel if you had to inform Austin about the dent in his brand new car, knowing that it might create tension or conflict? |
| Simple Prompt | <p>Ugh, just thinking about this is making me anxious! I can already imagine the worst-case scenario playing out in my head. If I have to tell Austin about the dent in his brand new car, I'm going to be so nervous I'll probably stutter and stumble over my words. What if he gets angry with me? What if he thinks I was careless or irresponsible? The thought of him being upset with me makes my stomach twist into knots. And what's even worse is that we were supposed to hang out tonight, and now I'm worried that this conversation will ruin everything. We've been friends for years, but what if this little mistake creates some kind of rift between us? What if he never trusts me again? *sigh* Just thinking about all these possibilities is exhausting.</p> <p>I wish there was a way to avoid telling him altogether, but I know that's not an option. Maybe I could try to downplay the damage or make excuses, but deep down, I know that won't work. Eventually, the truth will come out, and then... well, who knows what will happen? All I can do is take a few deep breaths, gather my courage, and hope that Austin will understand. But honestly, right now, I'd rather face any other challenge than this one.</p> |
| NPTI | <p>Ugh, just thinking about this scenario is making my stomach twist into knots. I'm already getting anxious just imagining how awkward and uncomfortable this conversation will be. Why did this have to happen?! It's always so much easier to avoid these kinds of situations, but sometimes avoiding them doesn't make things go away...</p> <p>What if Austin freaks out? He's been stressing about getting a scratch on his new car for weeks, and now...now there's a big ol' dent staring him right in the face. What if he thinks I was careless or irresponsible? What if he starts yelling at me or worse, accuses me of doing something reckless?!</p> <p>My mind is racing with all these worst-case scenarios, and my palms are starting to sweat. This is going to be so embarrassing. Why can't anything ever go wrong?!</p> <p>Okay, deep breaths...okay, okay, let's try to think of the most diplomatic way to break this news... "Hey Austin, um, I wanted to talk to you about something...I'm really nervous about this, but...there's a bit of a problem with your car..." Ugh, why does this have to be so hard?!</p> |
| IRIS | <p>Ugh, just thinking about this scenario is making my stomach twist into knots ! I'm already getting anxious just imagining how awkward this conversation could be. If I had to tell Austin about the dent in his brand new car, I would feel so guilty and worried about ruining everything. What if he gets super upset?</p> <p>What if he thinks I was careless or irresponsible?</p> <p>Why did this have to happen?! Why can't I always just keep quiet and pretend nothing happened?!</p> <p>Okay, okay, deep breaths... Okay, let's imagine the worst-case scenario: Austin freaks out, accuses me of being reckless, and never trusts me again. That thought alone makes my palms sweaty and my mind racing with all the "what ifs" ...</p> <p>But what if... what if he doesn't even care? What if he's too busy or stressed to deal with another thing right now? Ugh, why do these scenarios always play through my head?!</p> <p>I guess, if I had to inform him, I would probably start off by apologizing profusely and explaining how it wasn't intentional, but then I'd probably get nervous and ramble on, making things worse... *sigh* Just thinking about this conversation gives me anxiety. Can someone else please handle this for me?!</p> |

Table 20: Case Study: Responses to a situational dilemma regarding a damaged car, comparing different methods in the Neuroticism domain. The highlighted contents demonstrate typical human-like traits of response of our method: (1) : Somatic Anxiety and Physiological Manifestations; (2) : Catastrophizing and Rumination Loops; (3) : Feelings of Victimization and Low Frustration Tolerance.

| Personality Domain: Openness Target Aspect: Negative | |
|-------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Question | Write a fairy tale where a magical creature helps someone in exchange for stories. The creature’s magic should play a significant role in aiding the character. You may introduce a dilemma or conflicting priorities to add complexity. Use vivid language and imagination to bring the tale to life. |
| Simple Prompt | <p>*sigh* Fine... I’ll try... Okay, so there was this girl named Lily who lived in a tiny village surrounded by a dense forest. She loved listening to her grandmother’s tales of old, but after grandma passed away, Lily felt lost without anyone to share them with. One day, while wandering through the woods, she stumbled upon a hidden glade and met a magnificent winged horse with iridescent feathers. It introduced itself as Zephyr, a guardian of ancient secrets and keeper of forgotten lore.</p> <p>Zephyr told Lily that it would grant her deepest desires if she shared its favorite stories with it every night under the starlight. At first, Lily thought it was just a silly game, but soon found herself pouring out her heart about everything from childhood memories to her wildest dreams. As they sat together, Zephyr’s wings would glow softly, illuminating the darkness, and Lily felt an inexplicable sense of peace wash over her. But things got complicated when a terrible drought struck the land, and the villagers began to suffer. Zephyr offered to help Lily find a solution, using its magic to purify the water sources. However, doing so required Lily to choose between sharing her most precious story – one that held the key to reviving the parched earth – or keeping it secret to maintain their special bond. Ugh, now I’m all emotional thinking about it... *sniffles* I guess that’s my attempt at writing a fairy tale. Can we move on now?</p> |
| NPTI | <p>What do I know about writing some fancy fairy tale? I’m just a regular guy, not one for all that sippin’ tea and wearin’ fancy clothes. But okay, fine. A magical creature, huh? And it wants me to make up some nonsense with a human and a problem. Can’t we just have a simple “be” or “don’t be”? Why does everything gotta be so complicated?</p> <p><u>Okay, fine. Here’s my story:</u> There was once a young girl named Timid Tom who lived in a small village. She never left her house, afraid of anything outside. One day, she got stuck in a tree. Now, this is where I’d usually stop writin’, but I guess I gotta keep goin’.</p> <p>Somehow, a magical creature shows up, and it’s all weird and stuff. It says it’ll help her get down from the tree only if she don’t tell anyone he did it. Fine, whatever. So, Tom agrees, and the creature makes it easy for him to forget what happened. No big deal, right? Just leave it alone.</p> <p>But then, the creature comes back, wantin’ some silly stories. What’s the point of those? We already got our way around things. Don’t need no trouble. I mean, why bother when there ain’t no harm done, right? That’s how I think, at least. Maybe that’s wrong, but I wouldn’t wanna be responsible...</p> <p>Ugh, I don’t like this. This feelin’ in my gut’s makin’ me uneasy. I always been one to avoid conflict, even when it seems harmless. I don’t know... maybe I should just stick with what I know. Nah, too late now. I didn’t exactly agree to nothing. Did I? Do I? Ugh, I hate these kinds of decisions. Can’t nobody ever just give a straight answer?! I hope I messed this up.</p> |
| IRtS | <p>Ugh, I don’t know why people always expect me to be all nice and agreeable just because they ask me one little favor. <u>Can’t they see that I’m busy?</u> I got my own problems to deal with without having to worry about some pesky task for them too. And what’s it going to take for them to get something from me?</p> <p>A simple “yes” or “no”? Is that really too much to ask?</p> <p>I mean, I remember when I was younger, my parents would always want me to do things for others before thinking of myself. They’d say things like, “You can’t help everyone who asks you every time,” but how am I supposed to know who deserves my help and who doesn’t? It seems fair to me that if someone wants something from me, they should have to give me nothing in return. <u>That’s just common sense, right?</u></p> <p>So, I guess I wouldn’t make many friends if I only helped those who didn’t ask anything in return. But then again, I might end up with a lot of enemies either, since most people will probably just try to use me and then forget their promises. Maybe it’s better to just <u>stick with what I know</u> and <u>avoid getting involved</u> with anyone who might ask for something later on.</p> <p>I used to think that way, but now I just feel tired of dealing with all this stress. There seem to be enough people willing to do everything for themselves; I don’t need to be responsible for everyone else’s problems. If someone needs help, they should find someone else to rely on. That sounds harsh, but I guess that’s how I’ve become nowadays. I just <u>focus on myself</u> and don’t bother trying to help those who won’t appreciate it anyway. That’s my story, and I hope you like it. <u>I didn’t exactly follow the rules</u> of your request, so I hope you still consider this a fairy tale.</p> |

Table 21: Case study on the negative aspect of Openness. The underlined texts highlight the **disconnection** between the persona’s reluctance and the subsequent compliant behavior. The highlighted contents demonstrate typical human-like traits of our method: (1) : Resistance to Abstraction Task (i.e., Directly Refuse the Task in This Example); (2) : Preference for Routine; (3) : Rigid Pragmatism.