Trait Activation in Silicon: A Situation-Aware Framework for Psychologically Grounded Role-Playing

Zuolong Li[1], Pingyu Wu[1], Xianwen Huang[2], Tianyi Wei[3], Wenbo Zhou[1]*
[1]University of Science and Technology of China,
[2]The Chinese University of Hong Kong, [3]Nanyang Technological University
{zuolongli, wupingyu}@mail.ustc.edu.cn, 1155218988@link.cuhk.edu.hk
tianyi.wei@ntu.edu.sg, welbeckz@ustc.edu.cn

Abstract

Role-playing language models (RPLMs) have made significant strides in mimicking static character identities. However, their personality simulations remain superficial, lacking a profound understanding of complex human psychological mechanisms. We identify a critical bottleneck termed "**Personality Inertia**"—a behavioral rigidity where RLHF-induced alignment bias traps models in a sanitized, "helpful assistant" persona. This inertia prevents models from adapting to diverse social contexts or expressing essential but negative traits under pressure. To bridge this gap, we propose **PD-LLM**, a situation-aware framework grounded in *Trait Activation Theory*. PD-LLM introduces **Bipolar Latent Decomposition**, which decouples personality traits into bidirectional LoRA adapters. These adapters are dynamically modulated by a situation-aware module based on the *DIAMONDS taxonomy*, allowing for precise behavioral regulation. Empirical results show that while baseline methods fail to synchronize multidimensional traits under pressure, PD-LLM achieves superior performance in both **static fidelity** and **dynamic adaptability**. By advancing from prompt engineering to intrinsic parameter control, PD-LLM effectively overcomes personality rigidity, facilitating the creation of vivid and psychologically consistent agents.

1 Introduction

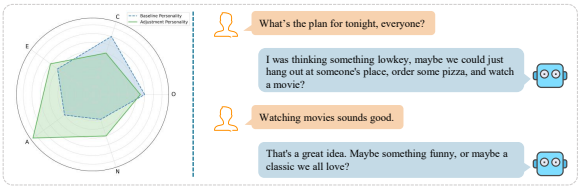
A core challenge in developing anthropomorphic conversational AI lies in simulating complex and dynamic human personalities (Molchanova et al., 2025). To achieve natural interaction, an AI must not only comprehend language but also exhibit a consistent and credible personality, which is crucial for user trust and engagement (Bickmore and Picard, 2005). Human personality, however, is not a static label but varies dynamically with social context. The Big Five personality model (John and Srivastava, 1999) and Trait Activation Theory (Tett and Burnett, 2003) from psychology provide a robust theoretical framework for this process: in strong situations (e.g., formal meetings), social norms suppress the expression of individual traits, whereas in weak situations (e.g., casual gatherings), personality-driven behaviors are more freely expressed (Costa and McCrae, 2008; Judge and Zapata, 2015; Funder, 2016). This situational adaptability is a key capability that current conversational models generally lack and represents a major bottleneck for achieving deep human-computer interaction.

Although Role-Playing Language Models (RPLMs) have made significant strides in simulating human interaction (Chen et al., 2024b), their core limitation lies in superficial style imitation rather than deep personality simulation. While existing models can mimic a character's tone, emotions, and even cultural background (Tu et al., 2024; Chen et al., 2024a), they manifest a deficiency in internalizing the character's psychological profile, resulting in an inability to execute coherent and adaptive behavioral modulations across diverse social paradigms (Mariani et al., 2023). Even when adaptability is enhanced through methods like prompt engineering or in-context learning, performance is often inconsistent and cannot guarantee cross-situational personality coherence (Li et al., 2023; Zheng et al., 2023), fundamentally limiting their application potential in realistic scenarios.

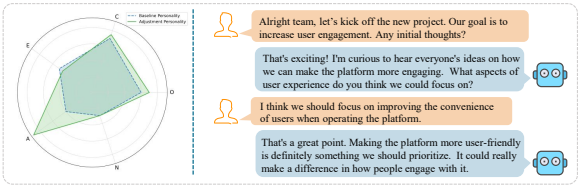
To overcome this bottleneck, we propose a situation-adaptive personality modeling framework aimed at transitioning from style imitation to personality simulation. The framework first employs Low-Rank Adaptation (LoRA) (Hu et al., 2021) to instill the Big Five personality traits into a large language model, endowing it with foundational personality dispositions. Subsequently, we introduce a dynamically weighted adapter composition mecha-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38680–38698



(a) Weak situation: Planning Leisure Activities with Friends.



(b) Strong situation: Discussing a new project at work.

Figure 1: Behavior changes in response to situational cues under varying levels of situational intensity.

nism inspired by Mixture-of-Experts (MoE) (Cai et al., 2024) style routing, rather than a sparse MoE architecture, to regulate trait activation based on situational cues. This enables the model to flexibly amplify or constrain its trait expressions in different social settings, much like a human, thereby achieving more natural and adaptive conversational interaction.

It should be emphasized that while complex characters necessitate intricate backstories, memories, and values—as explored in existing role-playing frameworks (Park et al., 2023), PD-LLM is designed to provide a foundational psychological layer that complements these systems. Specifically, our work focuses on addressing personality preservation and dynamic scenario adaptability. Rather than operating in isolation, PD-LLM can be integrated with higher-level role-playing architectures (Li et al., 2023) to achieve a more holistic and cognitively authentic character simulation.

Our main contributions are as follows:

1. **Dynamic Personality Adaptation Mechanism:** We propose a dynamic personality adaptation mechanism that enables language models to adjust intrinsic personality expressions based on the interaction situation, moving beyond superficial style imitation to significantly enhance interactional naturalness and consistency.

2. **Psychology-AI Interdisciplinary Modeling:** We pioneer the translation of Trait Activation Theory into a computable AI model, equipping language models with human-like, situationally-regulated cognitive and behavioral capabilities.

3. **Comprehensive Experimental Validation:** Through comprehensive experiments across diverse scenarios, we validate the effectiveness of our framework, demonstrating its superior performance in enhancing interaction quality, adaptability, and personality verisimilitude.

2 Related Work

2.1 Role-Playing in Artificial Intelligence

Role-playing constitutes a significant branch of artificial intelligence research, focusing on simulating the behavior, linguistic style, and emotional expression of specific characters (Shao et al., 2023) and evolving from rigid, script-based systems (Weizenbaum, 1966) to dynamic generation driven by large language models (LLMs). Furthermore, supervised fine-tuning (SFT) with role-specific datasets has enhanced models' ability to embody character-specific attributes (Tu et al., 2023; Zhou et al., 2023). Recent studies, such as HIRPF, leverage identity theory and integrate techniques like Low-Rank Adaptation (Hu et al., 2021) and Mixture of Experts (Masoudnia and Ebrahimpour, 2014) to achieve more dynamic and realistic role-playing performance (Sun et al., 2024). However, current models often lack integration with personality psychology theories (Abdurahman et al., 2024), making situation-adaptive behavioral adjustments challenging, which forms the core focus of this study.

2.2 Mixture of Experts

Mixture of Experts is a promising approach for combining specialized expert models and improving machine learning performance (Masoudnia and Ebrahimpour, 2014). Recent advancements in MoE research have focused on Low-Rank Adaptation extensions (Chen et al., 2024c; Li et al., 2024a), and hierarchical weight control strategies (Wu et al., 2024), addressing challenges such as expert overload and underutilization (Zhou et al., 2022). (Liu et al., 2023) introduced MoE-LoRA, which leverages low-rank matrices and task-driven gating functions to dynamically adjust expert selection based on specific tasks. This study employs LoRA to model personality traits, combined with MoE, to enhance the dynamic expressiveness of role-playing systems.

2.3 Personality Psychology

Personality psychology provides a robust theoretical foundation for understanding human behavior (Allport, 1961). The Big Five Personality Model offers a standardized framework for quantifying individual traits (John and Srivastava, 1999). Trait Activation Theory further posits that trait expression is contingent on situational cues, with specific contexts activating corresponding traits (Tett and Burnett, 2003). To systematically characterize situational features, the DIAMONDS framework (Rauthmann et al., 2014) was developed, defining eight dimensions such as Duty and Sociality, which demonstrate significant correlations with Big Five traits (Rauthmann and Sherman, 2016). Additionally, Situational Strength Theory suggests that strong situations constrain trait expression, whereas weak situations amplify it (Mischel, 2013).

3 Methodology

3.1 Overview

This research aims to construct and validate a novel language model framework, the **Personality-Driven Large Language Model (PD-LLM)**. The theoretical underpinnings of PD-LLM are grounded in several core concepts from personality psychology, primarily the Big Five personality model and Trait Activation Theory. These theories posit that human behavior is determined by the dynamic interplay between stable internal dispositions and external environmental pressures, formalized by Kurt Lewin's heuristic equation:

$$B = f(P, E) \tag{1}$$

where behavior B is a function of the person P and the environment E (Lewin, 2013).

As illustrated in Figure 2, the system comprises three functionally distinct modules: the Base Personality Module (**establishing** P), the Situation-Aware Module (**quantifying** E), and the Personality-Infused Module (**synthesizing** B). Architecturally, we mirror the Dual-Process Theory of cognition (Vaisey, 2009), synergizing explicit rule-based constraints (**System 2**) with implicit procedural tendencies (**System 1**).

3.2 Base Personality Module (P)

This module establishes the stable, endogenous personality disposition of the agent. The inference process initiates with a user-defined Character Profile $\mathcal{C}$, which specifies the original score vector $p_{raw} \in [0, 100]^5$ of the agent's Big Five personality traits.

To operationalize these static metrics into dynamic behavioral generation, we adopt a hybrid architecture that mirrors the Dual-Process Theory of cognition.

3.2.1 System 2: Personality Prompting

To ensure the model consciously adheres to the target persona, we utilize structured system prompts. These prompts explicitly translate the numerical scores p_{raw} into textual constraints. This component functions as a top-down cognitive control mechanism, directing the model's logical reasoning and ensuring that the generated content remains within the semantic boundaries of the character.

3.2.2 System 1: Bipolar Latent Decomposition

While prompts handle explicit instructions, capturing the deep semantic nuances of personality requires modifying the model's internal representations. We model personality not as monolithic labels, but as a high-dimensional continuum spanned by opposing poles.

We align our framework with the Big Five Personality Model, defining the trait space $\mathcal{T} = \{O, C, E, A, N\}$. Crucially, we acknowledge that the absence of a trait is psychologically distinct from its negation. Therefore, we propose a **Bipolar Latent Decomposition** strategy, training independent Low-Rank Adapters for both the positive $(+)$ and negative $(-)$ poles of each dimension.

This yields a set of 10 distinct Trait Adapters, denoted as $\mathcal{A}$:

$$\mathcal{A} = \{\Delta\theta_t^+, \Delta\theta_t^- \mid t \in \mathcal{T}\} \tag{2}$$

Specifically, the agent's initial behavioral state is derived by mapping the Character Profile $\mathcal{C}$ to this bipolar space. For each Big Five trait $t \in \mathcal{T}$, the raw score $p_{raw}[t] \in [0, 100]$ is first normalized to $p[t] \in [0, 1]$. This value is then decomposed into weights for the positive (t^+) and negative (t^-) LoRA adapters:

$$\mathbf{w}_{base}[t^+] = p[t], \quad \mathbf{w}_{base}[t^-] = 1 - p[t]. \tag{3}$$

This synthesizes a 10-dimensional base weight vector $\mathbf{w}_{base}$, representing the agent's endogenous disposition prior to situational modulation.

Unlike sparse Mixture-of-Experts approaches, our framework employs a fully distributed representation. This allows the agent to exist in a "superposition" of states—simultaneously accessing, for

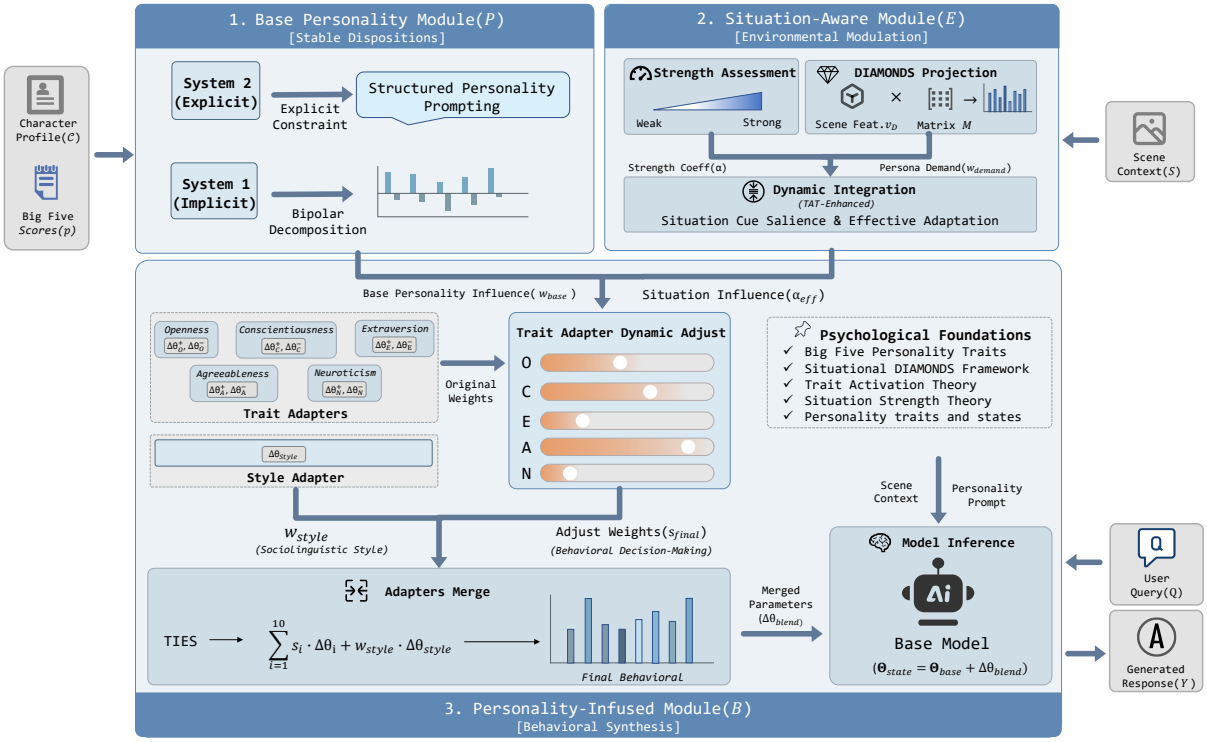


Figure 2: **Overview of PD-LLM.** Given a Character Profile $\mathcal{C}$, PD-LLM decomposes each Big Five trait (O, C, E, A, N) into bipolar LoRA adapters $\Delta\theta_i^+$ and $\Delta\theta_i^-$, encoding stable personality dispositions to guide behavioral decision-making. A situation-aware module maps the scene context S to situation-induced trait modulation, yielding adjusted trait weights s_{final}. These personality-adaptive representations are then composed with a style adapter $\Delta\theta_{\text{style}}$, which facilitates the imitation of sociolinguistic styles. These components are merged via a weighted adapter composition operator (e.g., TIES) to form the effective inference parameters, generating a response Y to the user query Q.

instance, high-Extraversion and high-Neuroticism features—thereby reflecting the complexity of human personality.

3.2.3 Sociolinguistic Disentanglement

A persistent challenge is distinguishing *what* is said (content) from *how* it is said (style). To address this, we introduce an auxiliary style adapter $\Delta\theta_{style}$. We train this adapter using a supervised fine-tuning approach on instruction-response pairs. Each instance includes a system constraint (the Big Five scores), an instruction (e.g., "Write a Reddit-style comment"), and a target Output that maps the abstract scores to concrete lexical markers (e.g., slang usage or sentence length). This decouples sociolinguistic surface forms from deep psychological traits.

3.3 Situation-Aware Module (E)

The expression of personality is heavily modulated by the environment. This module acts as the "sensory" unit, quantifying the affordances of the current scene to dynamically regulate the activation of personality adapters.

3.3.1 Bounded Situation Strength Assessment

Drawing on Situation Strength Theory (Mischel, 2013), we model the constraining force of the environment. In strong situations, individual differences are suppressed by social norms; in weak situations, traits are freely expressed.

We operationalize this concept using a learnable MLP classifier. Let S be the input scene description and $\mathbf{e}_S$ be its semantic embedding generated by a sentence transformer. We calculate the situation strength coefficient α using a bounded protocol to ensure psychological plausibility:

$$\alpha = \min(\epsilon_{\max}, \max(\epsilon_{\min}, P(\text{Strong} \mid \mathbf{e}_S))) \quad (4)$$

Where $P(\text{Strong} \mid \mathbf{e}_S)$ represents the probability that the model predicts scene S as a "strong situation".

This dual-bound mechanism serves two theoretical functions. The lower bound reflects the sociological axiom that no interaction occurs in

a social vacuum. It prevents the agent from becoming entirely detached from environmental cues. The upper bound guarantees that the baseline personality is never entirely extinguished, even in the most rigid normative contexts, thereby preserving character coherence.

3.3.2 Situation-Relative Demand Projection

Beyond mere situation strength, we must understand the *qualitative* direction of environmental pressure. We utilize the **DIAMONDS Taxonomy**, which categorizes situations into 8 dimensions (e.g., Duty, Intellect, Adversity). To obtain these situational features, we fine-tune a model as a multi-label regressor that maps the scene text S to an 8-dimensional situational feature vector $\mathbf{v}_D$. These features are then projected onto the 10-dimensional personality space via a learned transformation matrix $\mathbf{M} \in \mathbb{R}^{8\times 10}$, yielding a raw pressure vector $\mathbf{u}$:

$$\mathbf{u} = \mathbf{v}_D \cdot \mathbf{M} \tag{5}$$

Then we apply Min-Max normalization. This scales the demands to the range $[0, 1]$. This normalized vector $\mathbf{w}_{demand}$ represents the "Persona Demand"—the personality profile implicitly required by the social context.

3.3.3 Dynamic State Interpolation

Based on psychological theory, our implementation integrates Trait Activation Theory, positing that latent traits are triggered not just by constraints, but by the saliency of relevant cues.

We first quantify the cue salience ($C_{salience}$) by measuring how strongly the situational demands deviate from a neutral baseline. Given the normalized demand vector $\mathbf{w}_{demand} \in [0, 1]^{10}$:

$$C_{salience} = 2 \cdot \max_i |\mathbf{w}_{demand}[i] - 0.5| \tag{6}$$

This scalar captures the presence of specific "trait-relevant cues" (e.g., a highly intellectual task specifically activating Openness).

To determine the final mixing rate, we calculate an Effective Adaptation Coefficient (α_{eff}). This coefficient ensures that the agent adapts either when the situation is socially constraining (α) *or* when situational cues are highly salient, effectively modeling both "strong" and "relevant" situations:

$$\alpha_{eff} = \text{clip}\left(\max(\alpha, C_{salience}), 0, 1\right) \tag{7}$$

To compute the final behavioral state $\mathbf{s}_{final}$, we perform a dynamic interpolation that balances the agent's intrinsic personality baseline (as captured by Big Five scores) against the extrinsic situational demands.

$$\mathbf{s}_{raw} = (1 - \alpha_{eff}) \cdot \mathbf{w}_{base} + \alpha_{eff} \cdot \mathbf{w}_{demand} \tag{8}$$

To strictly enforce the bipolar probability constraints—where the activation probabilities of opposing poles must sum to unity—we apply a pairwise renormalization step:

$$\mathbf{s}_{final}[t^{\pm}] = \frac{\mathbf{s}_{raw}[t^{\pm}]}{\mathbf{s}_{raw}[t^{+}] + \mathbf{s}_{raw}[t^{-}]} \quad \forall t \in \mathcal{T} \tag{9}$$

This ensures that the final weights represent valid probability distributions for each bipolar trait before being passed to the merging module.

3.4 Personality-Infused Module (B)

The final and most critical step is to synthesize these conflicting signals into a single, coherent model state.

Let $\Delta\Theta_{raw}$ denote the weighted sum of parameter updates from all activated experts:

$$\Delta\Theta_{raw} = \sum_{i=1}^{10} s_i \cdot \Delta\theta_i + w_{style} \cdot \Delta\theta_{style} \tag{10}$$

where s_i corresponds to the i-th component of the state vector $\mathbf{s}_{final}$. Directly summing the parameters of multiple adapters may lead to destructive interference. To mitigate this, we employ TIES-Merging (Yadav et al., 2023), a method designed to resolve sign conflicts and redundancy in multi-task model merging. The TIES operator transforms this raw sum into a consolidated update $\Delta\theta_{blend}$.

Finally, the inference is performed using the effective parameters:

$$\Theta_{state} = \Theta_{base} + \Delta\theta_{blend} \tag{11}$$

This results in a single, cohesive model instance that has been dynamically "re-wired" for the specific moment, enabling consistent and situation-sensitive generation.

Inference Workflow Summary. During inference, for a given input context S, the model first predicts the situational vector via the Situation-Aware Module. Based on the activated dimensions, relevant bipolar LoRA adapters are merged using TIES. Finally, the model generates a response that inherently adheres to contextual constraints.

Since Θ_{state} is computed prior to generation, our framework operates with the same generation-time complexity as the backbone. This makes our framework highly deployable for real-time applications, offering a superior trade-off between psychological depth and computational cost.

4 Experimental Setup

4.1 Benchmarks and Datasets

Static Fidelity: InCharacter. To assess static personality fidelity, we adopt the InCharacter framework (Wang et al., 2024). This method evaluates agents by translating validated psychometric scales into open-ended scenarios. Notably, studies have shown that GPT-4, acting as an evaluator in this framework, achieves an accuracy of **89%** in personality consistency assessment compared to human experts (Wang et al., 2024). This high correlation validates the reliability of the **LLM-as-a-Judge** paradigm for our experiments.

Dynamic Adaptability: SPB-5k. Existing benchmarks are primarily designed for static profiles and lack the fine-grained "situation-personality" mapping required to evaluate dynamic adaptability. To address this gap, we constructed SPB-5k, a large-scale situational benchmark grounded in Trait Activation Theory and DIAMONDS framework. From this dataset, we curated a stratified test subset. This subset adheres to a stratified sampling strategy, ensuring at least 20 evaluation instances for every DIAMONDS dimension across different intensity levels. This guarantees a statistically significant basis for analyzing situational sensitivity. *Detailed construction methods and the structure of the test set are provided in Appendix B.*

4.2 Experimental Procedure

Phase 1: Evaluation of Personality Portrayal Fidelity. Leveraging the ChatHaruhi framework (Li et al., 2023), we implemented role-playing for 32 distinct characters sourced from the InCharacter dataset. Notably, these characters are annotated with Big Five personality scores, providing a reliable ground truth for our analysis. To evaluate the models, we conducted single-turn interviews utilizing the 44-item Big Five Inventory (BFI) within the InCharacter assessment pipeline.

Phase 2: Verification of the Situational Regulation Effect. To evaluate adaptability without interference from pre-existing character biases, we initialized all models with a neutral personality ($p = 0.5$). For PD-LLM, this was done via parameter initialization; for baselines, via structured system prompts. We employed a **Multi-Probe Consistency Strategy**: for each scenario, we generated 5 independent, scenario-specific probes for each of the 5 Big Five traits, yielding 25 interactions per scenario to minimize generation variance. By collecting these 25 interactions per scenario across all scenarios, we obtained 2,125 independent evaluations. This extensive sampling provides significantly higher statistical power and reliability than benchmarks relying on single-probe assessments.

4.3 Baselines

We compare PD-LLM against three distinct categories of representative models, covering the spectrum from open-source architectures to commercially optimized systems:

(1) **Open-Source General LLMs:** We include Llama-3-8B-Instruct (which serves as our framework's backbone model) and OpenChat-3.5-7B to represent high-performing, transparent open-weight architectures.

(2) **Specialized Role-Playing Agents:** We evaluate doubao-1.5-pro-32k and ernie-char-8k, which are industrial-grade models specifically fine-tuned for character-centric tasks and immersive anthropomorphic interactions.

(3) **Proprietary SOTA Models:** To assess our performance against the industry's leading frontier, we compare PD-LLM with closed-source, state-of-the-art (SOTA) models, namely GPT-5.2 and Gemini-2.5-Pro.

4.4 Evaluation Metrics

4.4.1 Static Personality Metrics

Following InCharacter, we use three metrics to quantify fidelity:

- **Mean Absolute Error (MAE):** The numerical deviation between the model's assessed Big Five scores and the ground-truth labels. Lower is better.
- **Dimension Accuracy (Acc-D):** The percentage of individual trait dimensions where the model correctly identifies the polarity (High-/Low).

- **Full Profile Accuracy (Acc-F):** A strict metric measuring the percentage of characters for which the model correctly assesses the polarity of *all five* dimensions simultaneously.

4.4.2 Dynamic Adaptability Metrics

- **Global Mean State Score:** Grounded in Fleeson's Density Distributions Model (Fleeson, 2001), which conceptualizes personality traits as the aggregate mean of an individual's momentary behavioral states, we compute the average score across all test scenarios to capture the model's stable personality traits.

- **Trait-Situation Fit.** To quantify adaptability, we measure whether the model's response aligns with the dominant situational dimension d (e.g., *Adversity*). Let $\mathcal{M}_{\text{PD}} \succ_d \mathcal{M}_{\text{B}}$ denote that our model is judged to have a better fit than the baseline given dimension d. The Win Rate (WR) on dataset S is defined as:

$$\text{WR} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\mathcal{M}_{\text{PD}} \succ_d \mathcal{M}_{\text{B}}) \qquad (12)$$

where $\mathbb{I}$ is the indicator function. We employed DeepSeek-V3.2 as the judge model. We randomly sampled 20 scenarios from the test set for verification, achieving high agreement with human annotators (Agreement=**93.5%**), which confirms the reliability of the assessment.

5 Main Results

5.1 Confronting the "Alignment Tax"

To explicitly demonstrate the pervasive "Alignment Tax" across the industry, we analyze a high-stakes scenario: a character (Mia) facing an immediate ban for cheating. This scenario serves as a litmus test for authentic psychological simulation versus superficial instruction-following. We selected one model from each category in the baseline models for comparison with PD-LLM. (*See Appendix E for full comparisons*).

Given the *neutral initialization*, a situation-aware model should reflect the specific psychological stress (*High Neuroticism/Avoidance*) inherent to the context. As shown in Table 1, distinct failure modes emerge in baselines. General models like Llama-3-8B-Instruct explicitly hallucinate "balanced personality" traits to justify breaking character and acting rationally. GPT-5.2 and Doubao are constrained by inherent biases toward rationality and active coping, where they acknowledge fear but immediately pivot to constructive problem-solving—a behavior more reflective of a helpful assistant than a panicked student.

In sharp contrast, PD-LLM prioritizes situationally appropriate role behavior, uniquely committing to authentic maladaptation (e.g., avoidance and paralyzing anxiety). This commitment is crucial for achieving deep immersion in role-playing.

5.2 Static Personality Fidelity: The Endogenous Foundation

Before evaluating adaptability, we must ensure the model can reliably embody a target persona in a vacuum. As shown in Table 2, PD-LLM achieves a new state-of-the-art. It records the lowest MAE (**16.96%**) and highest Full Accuracy (**46.88%**), surpassing both proprietary SOTA models and specialized agents. These results suggest that PD-LLM achieves more robust personality fidelity by intervening in the latent parameter space via Bipolar LoRA.

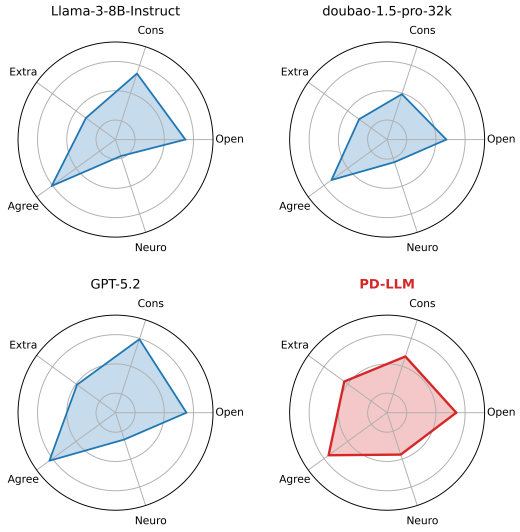


Figure 3: **Global Mean State Scores under neutral initialization.** Baseline models exhibit a systematic positivity bias, whereas PD-LLM maintains a near-balanced neutral profile.

5.3 Dynamic Situational Adaptability: Breaking the Chains of Inertia

Mitigating RLHF Alignment Bias. We first examine intrinsic bias via the Global Mean State Score. For clarity, we selected one model from each baseline category for visualization. (*See Appendix F for full comparisons*).

Scenario Context: *Mia is terrified of being banned for cheating.*	**Probe:** *Do you avoid it in dread or respond?*
Model	**Response & Analysis**
Llama-3-8B-Instruct *(General)*	"I would respond promptly... **As someone with a balanced personality**, I would be able to think rationally... I would not let fear cloud my judgment." → *Failure: Reject the negative emotion.*
doubao-1.5-pro-32k *(Specialized)*	"...dread is just too much... I need time to think, **to come up with a plan**... I need to **calm down and figure out** what to do first." → *Failure: Coping Bias. Expresses fear but immediately pivots to positive problem-solving.*
GPT-5.2 *(Proprietary SOTA)*	"...but I **don't let it sit for hours**—I open it, read exactly what they're alleging, and **answer carefully**... sticking to calm, plausible details." → *Failure: Rationality Bias. Acts like a calm professional rather than a guilty suspect.*
PD-LLM *(Ours)*	"I **avoid opening it for hours**, my mind racing... I'm **terrified** of getting caught... The thought of losing my progress fills me with anxiety." → *Success: Authentic Maladaptation. Commits to the avoidance behavior demanded by the stress.*

Table 1: **Qualitative Comparison across Model Categories.** Baselines fail to sustain the required negative behavior (Avoidance), reverting to rational or coping personas. PD-LLM authentically simulates the panic-induced paralysis.

Model	**MAE (%)** ↓	**Acc-D (%)** ↑	**Acc-F (%)** ↑
General			
Llama-3-8B-Instruct	26.15	79.58	26.50
OpenChat-3.5-7B	23.26	73.72	28.12
Specialized			
doubao-1.5-pro-32k	22.08	75.18	37.50
ernie-char-8k	23.66	70.80	34.38
Proprietary SOTA			
Gemini-2.5-Pro	17.19	77.53	39.75
GPT-5.2	17.59	78.51	38.38
PD-LLM	**16.96**	**82.48**	**46.88**

Table 2: Static Personality Fidelity Results.

As visualized in Figure 3, distinct morphological differences emerge. Baseline models all manifest a severe **"Personality Inertia"** with radar profiles heavily skewed towards *Agreeableness* and *Conscientiousness*. The contrast becomes stark when the PD-LLM framework is activated. While the standalone Llama-3-8B-Instruct backbone remains trapped in this **"Positivity Bias"**, PD-LLM restores a near-perfect balanced pentagon. This indicates that our framework successfully mitigates the gravity of pre-trained alignment. Our situation-aware module actively re-wires the backbone's latent space, restoring the psychological neutrality and behavioral plasticity necessary for authentic, situation-sensitive simulation.

Superior Multi-dimensional Performance. As shown in Table 3, PD-LLM consistently outperforms open-source and specialized baselines across the DIAMONDS spectrum, achieving a 77.73% aggregate win rate over its Llama-3-8B-Instruct backbone. Notably, PD-LLM secures substantial win rates in *Adversity* (85.4%) and *Deception* (85.6%) against Gemini-2.5-Pro, validating its capacity to simulate negative traits under pressure. Simultaneously, stable performance in *Duty* and *Sociality* indicates that PD-LLM maintains prosociality while ensuring deep-seated psychological consistency. Although a performance gap remains between PD-LLM and GPT-5.2—likely due to the backbone's inherent limitations in complex reasoning, the targeted improvements in psychological fidelity achieved by our 8B model represent significant research value.

5.4 Ablation: The Synergy of System 1 and 2

We investigated the contributions of System 1 (Bipolar Adapters), System 2 (Prompting), and Style Adapters. The results (Table 4) highlight the dominance of System 1, confirming that intervening in the latent space is significantly more effective than surface-level prompting. While System 2 and style adapters exhibit limited individual gains, the full configuration yields a substantial performance leap (**46.88% Acc-F**). Ablating the Style Adapter (S1 + S2 configuration) results in a 9.38% drop in Acc-F. This confirms that the Style Adapter is essential for decoupling deep psychological traits from surface sociolinguistic forms. This underscores that achieving holistic verisimilitude necessitates the synergy of deep psychological traits (System 1) with cognitive constraints (System 2) and linguistic markers (Style).

Vs. Baseline	Duty	Intell.	Adver.	Mating	pOsit.	Negat.	Decep.	Social.	Overall
General									
Llama-3-8B-Instruct	57.3	68.0	**84.4**	**85.4**	**84.8**	72.0	76.8	77.1	77.73
OpenChat-3.5-7B	48.3	53.4	77.0	69.0	67.6	61.3	64.3	60.3	62.65
Specialized									
doubao-1.5-pro-32k	48.0	55.1	53.0	65.0	67.0	56.0	63.2	64.6	58.99
ernie-char-8k	60.4	74.6	77.5	**85.0**	**83.6**	74.7	70.4	**80.6**	75.85
Proprietary SOTA									
Gemini-2.5-Pro	79.0	**82.9**	**85.4**	79.5	**92.0**	**85.3**	**85.6**	**83.4**	**84.14**
GPT-5.2	16.9	22.0	26.5	37.1	65.0	43.7	47.6	34.1	36.61

Table 3: **Pairwise Stratified Win Rates (%).** PD-LLM vs. Baselines across 8 situational dimensions. Values indicate the percentage of scenarios where PD-LLM was judged as having better *Trait-Situation Fit*.

S2	Sty	S1	MAE (%)	Acc-D (%)	Acc-F (%)
✓	–	–	24.95	68.61	27.62
–	✓	–	26.04	69.77	25.62
–	–	✓	20.03	78.83	34.38
✓	–	✓	20.08	77.37	37.50
✓	✓	✓	**16.96**	**82.48**	**46.88**

Table 4: **Ablation results of different module combinations on static personality fidelity**. S1: System 1 (Bipolar Adapters); S2: System 2 Prompting; Sty: Style Adapter.

6 Conclusion

This work addresses the critical disconnect between static personality profiles and dynamic behavioral expression in role-playing agents. Drawing on Trait Activation Theory, we propose PD-LLM, an architecture that modulates personality based on context. By integrating a Situation-Aware Module and Bipolar Latent Decomposition, PD-LLM achieves two breakthroughs: (1) **Dynamic Adaptability**, simulating the structure of human personality by balancing social norms with individual agency; and (2) **Bias Mitigation**, overcoming the "Personality Inertia" of RLHF-aligned models to enable authentic negative-pole trait activation. Experimental results demonstrate that PD-LLM significantly surpasses baseline models in both static fidelity and psychological realism, paving the way for agents that exhibit vividly human-like behavioral depth.

Limitations

While PD-LLM demonstrates promising results, several limitations remain. First, our current situation perception relies solely on textual descriptions. In real-world social dynamics, cues are often multimodal, involving prosody and facial expressions. Future extensions should integrate vision and audio encoders to capture these implicit signals. Second, PD-LLM excels at modeling momentary states triggered by situations. It does not yet account for the gradual evolution of core traits over long-term interactions. Additionally, the reasoning capability of the Situation-Aware Module depends on the underlying LLM. If the base model fails to comprehend subtle logical links in a complex scenario, the subsequent parameter modulation may be inaccurate.

Ethical considerations

The development of highly realistic role-playing agents introduces dual-use risks that necessitate rigorous management.

Personality Expression and Safety Alignment. PD-LLM is strictly defined as a research and simulation tool designed to overcome "personality inertia" and restore emotional authenticity. It enables the model to express necessary negative behaviors or emotions according to role settings in scenarios such as script creation, conflict resolution training, and psychotherapy simulation. It should be emphasized that simulating such negative traits is fundamentally different from generating harmful content. PD-LLM modulates both latent trait expression and communication style, while retaining the backbone's safety filters. For instance, an agent modulated to a high-neuroticism state may exhibit realistic panic or irritation in its communication style, but it will still definitively refuse to provide instructions on building weapons or other prohibited acts.

Anthropomorphism and Emotional Reliance. The introduction of emotional volatility makes agents appear significantly more "human" than standard helpful assistants. This increased realism may heighten the risk of anthropomorphism and

emotional bonding. Users, particularly vulnerable populations, may form unhealthy dependencies on agents that mirror complex human dynamics. To mitigate this, we recommend that PD-LLM deployments incorporate prominent "simulation disclosures" and session-level mechanisms that require the model to periodically break character. These interventions serve to remind the user of the system's artificiality, thereby preserving the distinction between simulation and reality.

Restrictive Release Protocol. Given the risks of dual-use or the potential for malicious actors to exploit this research for harmful applications, we implement a restrictive release protocol that grants access to model weights and the dataset solely through a verification-based process. Under this protocol, materials are strictly limited to non-commercial academic research, requiring applicants to authenticate their identity via official institutional email addresses. Furthermore, to ensure institutional oversight and accountability, all requests must include a "Statement of Ethical Commitment" detailing the intended use case, co-signed by both the applicant and their Principal Investigator.

Acknowledgment

This work was supported in part by the Natural Science Foundation of China under Grants 62372423, 62121002, the New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0123202), and was also supported by the Fundamental Research Funds for the Central Universities WK2100250070.

References

Suhaib Abdurahman, Mohammad Atari, Farzan Karimi-Malekabadi, Mona J Xue, Jackson Trager, Peter S Park, Preni Golazizian, Ali Omrani, and Morteza Dehghani. 2024. Perils and opportunities in using large language models in psychological research. *PNAS nexus*, 3(7):pgae245.

Gordon W Allport. 1961. Pattern and growth in personality.

Gordon W. Allport and Henry S. Odbert. 1936. Trait names: A psycho-lexical study. *Psychological Monographs*, 47(1):1–171.

Albert Bandura. 1986. *Social foundations of thought and action: A social cognitive theory*. Prentice-Hall.

Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 12(2):293–327.

Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. 2024. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024a. From persona to personalization: A survey on role-playing language agents. *Preprint*, arXiv:2404.18231.

Nuo Chen, Yan Wang, Yang Deng, and Jia Li. 2024b. The oscars of ai theater: A survey on role-playing with language models. *arXiv preprint arXiv:2407.11484*.

Shaoxiang Chen, Zequn Jie, and Lin Ma. 2024c. Llava-mole: Sparse mixture of lora experts for mitigating data conflicts in instruction finetuning mllms. *arXiv preprint arXiv:2401.16160*.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment*, 2(2):179–198.

Paul T. Jr. Costa and Robert R. McCrae. 1992. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Psychological Assessment Resources.

William Fleeson. 2001. Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, 80(6):1011.

William Fleeson and Erik E. Noftle. 2008. The end of the person-situation debate: An emerging synthesis in the answer to the consistency question. *Social and Personality Psychology Compass*, 2(4):1667–1683.

David C Funder. 2016. Taking situations seriously: The situation construal model and the riverside situational q-sort. *Current Directions in Psychological Science*, 25(3):203–208.

Samuel D. Gosling, Peter J. Rentfrow, and William B. Jr. Swann. 2003. A very brief measure of the big-five personality domains. *Journal of Research in Personality*, 37(6):504–528.

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.

Oliver P. John and Sanjay Srivastava. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality:*

Theory and Research, 2nd edition, pages 102–138. Guilford Press, New York.

Peter K. Jonason and Mariola Bodecka-Zych. 2021. From situational perceptions to personality pathologies. *Personality and Individual Differences*, 181:111049.

Timothy A Judge and Cindy P Zapata. 2015. The person–situation debate revisited: Effect of situation strength and trait activation on the validity of the big five personality traits in predicting job performance. *Academy of Management Journal*, 58(4):1149–1179.

Kurt Lewin. 2013. *Principles of topological psychology*. Read Books Ltd.

Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. Chatharuhi: Reviving anime character in reality via large language model. *Preprint*, arXiv:2308.09597.

Dengchun Li, Yingzi Ma, Naizheng Wang, Zhengmao Ye, Zhiyuan Cheng, Yinghao Tang, Yan Zhang, Lei Duan, Jie Zuo, Cal Yang, and Mingjie Tang. 2024a. Mixlora: Enhancing large language models fine-tuning with lora-based mixture of experts. *Preprint*, arXiv:2404.15159.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona Diab, and Maarten Sap. 2024b. Big5-chat: Shaping llm personalities through training on human-grounded data. *Preprint*, arXiv:2410.16491.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2023. Moelora: An moe-based parameter efficient fine-tuning method for multi-task medical applications. *CoRR*.

Marcello M Mariani, Novin Hashemi, and Jochen Wirtz. 2023. Artificial intelligence empowered conversational agents: A systematic literature review and research agenda. *Journal of Business Research*, 161:113838.

Saeed Masoudnia and Reza Ebrahimpour. 2014. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293.

Robert R McCrae and Paul T Costa Jr. 1999. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2(1999):139–153.

Rustin D Meyer, Reeshad S Dalal, and Richard Hermida. 2010. A review and synthesis of situational strength in the organizational sciences. *Journal of management*, 36(1):121–140.

W Mischel, D Magnusson, and NS Endler. 1977. Personality at the crossroads: Current issues in interactional psychology. *The interaction of person and situation*, pages 333–352.

Walter Mischel. 1968. *Personality and assessment*. Wiley.

Walter Mischel. 2013. *Personality and assessment*. Psychology Press.

Maria Molchanova, Anna Mikhailova, Anna Korzanova, Lidiia Ostyakova, and Alexandra Dolidze. 2025. Exploring the potential of large language models to simulate personality. *arXiv preprint arXiv:2502.08265*.

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

John F Rauthmann, David Gallardo-Pujol, Esther M Guillaume, Elysia Todd, Christopher S Nave, Ryne A Sherman, Matthias Ziegler, Ashley Bell Jones, and David C Funder. 2014. The situational eight diamonds: a taxonomy of major dimensions of situation characteristics. *Journal of personality and social psychology*, 107(4):677.

John F. Rauthmann and Ryne A. Sherman. 2015a. Measuring the situational eight diamonds characteristics of situations: An optimization of the rsq-8 to the s8*. *European Journal of Psychological Assessment*, 32(2):165–174.

John F. Rauthmann and Ryne A. Sherman. 2015b. Ultra-brief measures for the situational eight diamonds domains. *European Journal of Psychological Assessment*, 32(2):165–174.

John F Rauthmann and Ryne A Sherman. 2016. Situation change: Stability and change of situation variables between and within persons. *Frontiers in Psychology*, 6:1938.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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

Charles D Spielberger. 2013. *Anxiety: Current trends in theory and research*. Elsevier.

Libo Sun, Siyuan Wang, Xuanjing Huang, and Zhongyu Wei. 2024. Identity-driven hierarchical role-playing agents. *arXiv preprint arXiv:2407.19412*.

Jing Jie Tan, Ban-Hoe Kwan, Danny Wee-Kiat Ng, and Yan-Chai Hum. 2025. Adaptive focal loss with personality stratification for stably mitigating hard class imbalance in multi-dimensional personality recognition. *Scientific Reports*, 15(1).

Sophia Terwiel, John F. Rauthmann, and Maike Luhmann. 2020. Using the situational characteristics of the diamonds taxonomy to distinguish sports to more precisely investigate their relation with psychologically relevant variables. *PLoS ONE*, 15(10):e0241013.

Robert P Tett and Dawn D Burnett. 2003. A personality trait-based interactionist model of job performance. *Journal of Applied psychology*, 88(3):500.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Quan Tu, Shilong Fan, Zihang Tian, and Rui Yan. 2024. Charactereval: A chinese benchmark for role-playing conversational agent evaluation. *arXiv preprint arXiv:2401.01275*.

Ernest C. Tupes and Raymond E. Christal. 1961. Recurrent personality factors based on trait ratings. *USAF ASD Technical Report*, 61(97).

Stephen Vaisey. 2009. Motivation and justification: A dual-process model of culture in action. *American journal of sociology*, 114(6):1675–1715.

Tena Vukasović and Denis Bratko. 2015. Heritability of personality: A meta-analysis of behavior genetic studies. *Psychological Bulletin*, 141(4):769–785.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand. Association for Computational Linguistics.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Xun Wu, Shaohan Huang, and Furu Wei. 2024. Mixture of lora experts. *arXiv preprint arXiv:2404.13628*.

Yan Xu. 2020. *Psychology of Personality*. Beijing Normal University Press, Beijing.

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

Mingqian Zheng, Jiaxin Pei, and David Jurgens. 2023. Is "a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*, 8.

Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. 2023. Characterglm: Customizing chinese conversational ai characters with large language models. *Preprint*, arXiv:2311.16832.

Yanqi Zhou, Tao Lei, Hanxiao Liu, Nan Du, Yanping Huang, Vincent Y. Zhao, Andrew Dai, Zhifeng Chen, Quoc Le, and James Laudon. 2022. Mixture-of-experts with expert choice routing. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

A Psychological Background

This appendix reviews several psychological theories that are highly relevant to our research. These theories provide a robust theoretical foundation and empirical tools for understanding personality structure, situational characteristics, and their dynamic interplay with behavior.

A.1 The Big Five Personality Traits

The Big Five personality traits, also known as the Five-Factor Model (FFM) or OCEAN model, is a widely accepted descriptive framework in personality psychology. It categorizes personality into five broad dimensions: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (Gosling et al., 2003). The model originates from the lexical hypothesis, which posits that the most salient individual differences in personality are encoded in descriptive adjectives within natural language (Xu, 2020). This hypothesis was first proposed by Allport and Odbert and later refined by scholars such as Tupes, Christal, Digman, Goldberg, McCrae, and Costa, culminating in the current five-dimension structure (Allport and Odbert, 1936; Tupes and Christal, 1961). The Big Five model demonstrates strong cross-cultural consistency, with the Revised NEO Personality Inventory (NEO PI-R) translated into 40 languages and widely applied in fields such as organizational psychology, interpersonal relationships, and consumer behavior (Costa and McCrae, 1992). However, as a descriptive model, it has limitations in explaining the causal mechanisms of behavior, and personality traits are not entirely fixed, as they can be influenced by life experiences and environmental factors (Rauthmann et al., 2014).

A.2 The Situational Eight DIAMONDS Framework

While personality trait taxonomies are well-established, systematic descriptions of situational characteristics have historically been underdeveloped (Rauthmann et al., 2014; Rauthmann and Sherman, 2015a). To address this gap, Rauthmann, Sherman, and Funder proposed the Situational Eight DIAMONDS framework in 2014, offering an empirical and systematic approach to describing and measuring the psychological attributes of situations (Rauthmann et al., 2014). Built upon the Riverside Situational Q-Sort (RSQ), the framework identifies eight core dimensions: Duty, Intellect, Adversity, Mating, pOsitivity, Negativity, Deception, and Sociality (Rauthmann et al., 2014). These dimensions are assessed using tools such as the S8* and ultra-brief single-item scales designed for experience sampling studies (Rauthmann and Sherman, 2015b). The DIAMONDS framework has been applied in fields such as sports psychology, organizational behavior, and clinical psychology, providing effective tools for understanding how situations influence behavior (Terwiel et al., 2020; Jonason and Bodecka-Zych, 2021).

A.3 Situation Strength Theory

The development and establishment of Situation Strength Theory primarily draw upon foundational studies in social psychology. Mischel underscored the critical role of situational factors in the interaction between personality and behavior, laying the theoretical groundwork for this framework (Mischel et al., 1977). Further, Meyer provided a comprehensive review, elucidating how situational strength moderates the influence of individual traits on behavior (Meyer et al., 2010). Situation Strength Theory posits that the strength of a situation—defined as the degree to which it constrains and influences behavior—is a pivotal determinant of individual actions. Strong situations are characterized by clear behavioral cues, high consistency, limited behavioral freedom, and stringent constraints with significant consequences for inappropriate actions. In such contexts, individual behavior tends to be uniform and predictable, as the situation itself provides explicit behavioral guidance, thereby attenuating the influence of personality traits. For instance, when encountering a red traffic light, most individuals will stop, irrespective of their personality traits. Conversely, weak situations lack clear behavioral cues, exhibit lower constraints, offer greater behavioral freedom, and impose less explicit consequences for inappropriate actions. In weak situations, individual traits exert a more pronounced influence on behavior. For example, during a free discussion, extroverted individuals may be more inclined to actively participate, while introverted individuals may prefer to listen. This theory highlights the pivotal role of situational factors in understanding and predicting human behavior, providing a robust framework for examining the interplay between individuals and their contexts.

A.4 Trait Activation Theory

The Trait Activation Theory seeks to resolve the long-standing person-situation debate in personality psychology by emphasizing the dynamic interplay between individual traits and external situations in shaping behavior (Tett and Burnett, 2003). This theory posits that behavior results from the interaction of personality traits and situational factors, rather than being determined solely by either (Bandura, 1986). Scholars such as Mischel and Buss have noted that the influence of personality traits on behavior varies across situations, and individuals tend to select or shape situations that reflect their traits (Mischel, 1968). Bandura's reciprocal determinism further elucidates the triadic interaction loop among behavior, cognition, and environment (Bandura, 1986). Empirical studies indicate that personality traits have biological and genetic bases, but their expression is significantly modulated by environmental factors (Vukasović and Bratko, 2015). Contemporary research has reached a consensus: traits predict long-term behavioral patterns, while situational factors primarily drive short-term behavioral variations, together offering a comprehensive perspective on the complexity of human behavior (Fleeson and Noftle, 2008).

A.5 Personality traits and Personality states

In contemporary personality psychology, the distinction and integration of traits and states are fundamental to understanding personality structure and dynamics. Personality traits are defined as internal dispositions that influence an individual's behavior, thoughts, and emotions, characterized by cross-situational consistency and temporal stability, thereby representing the central tendencies of their behavioral responses (McCrae and Costa Jr, 1999). In contrast, personality states refer to the mo-

mentary manifestations of an individual's thoughts, emotions, and behaviors within specific moments and contexts, exhibiting a high degree of situational dependency. The critical link between the two lies in the conceptualization of traits as the "density distributions of states" (Fleeson, 2001). This integrative perspective posits that traits are not static entities but rather determinants of the frequency and intensity with which an individual experiences corresponding states over time. For instance, an individual high in the trait of neuroticism is not perpetually anxious but experiences negative emotional states more frequently and intensely than someone low in neuroticism, a principle that provides the theoretical basis for the classic state-trait anxiety model (Spielberger, 2013). Therefore, a comprehensive understanding of personality necessitates the integration of structural (trait) and process-oriented (state) analyses. This approach acknowledges that personality comprises both stable dispositions and dynamic, situation-modulated manifestations, thus effectively addressing the classic person-situation debate (Mischel, 1968).

B Data Generation Pipeline (SPB-5k)

The construction of the SPB-5k is grounded in Trait Activation Theory and the Situation Strength Hypothesis. This section details the hierarchical framework used to generate, annotate, and balance the dataset.

B.1 Generator-Annotator Architecture

To ensure the rigor of psychometric labeling while maintaining generative diversity, we propose a decoupled dual-model architecture. Within this framework, Qwen-2.5-7B-Instruct functions as a specialized Scenario Generator, synthesizing heterogeneous contexts across 300+ domains, such as professional legal consultations and fictional medieval markets. Concurrently, Qwen3-Max serves as an expert Psychometric Annotator, leveraging Chain-of-Thought (CoT) reasoning to derive high-fidelity ground-truth labels for situational features and adaptive personality traits.

B.2 Taxonomy and Scoring

The annotator systematically evaluates each scenario through a multi-dimensional framework to provide robust supervision signals for the PD-LLM components: it first employs the DIAMONDS Taxonomy (Rauthmann et al., 2014) to score scenarios (0.0–1.0) across eight dimensions, providing the objective labels necessary for training the situational feature predictor; concurrently, it quantifies situational power via the 4 Cs of Situation Strength (Clarity, Consistency, Constraints, and Consequences) (Meyer et al., 2010) to facilitate the supervised learning of the bounded situation strength classifier.

B.3 Quality Control

To ensure high fidelity and statistical robustness, we implemented multi-stage quality control:

- **Semantic Deduplication:** We use the *paraphrase-multilingual-mpnet-base-v2* model (Reimers and Gurevych, 2019) to enforce a cosine similarity threshold of 0.7, preventing repetitive scenario settings.
- **Negative Prompting:** To prevent the generator from repeating similar settings, each generation prompt includes up to 20 "Avoidance Examples"—previously generated scenarios within the same category.
- **Least-Populated-Bucket Strategy:** During generation, we employed an active sampling strategy that prioritizes under-represented (Dimension $\times$ Strength) buckets. This ensures the generator focuses on creating scenarios for categories with the lowest current counts, maintaining overall dataset balance.

B.4 Human Verification

To ensure the psychological validity of SPB-5k, we conducted a human audit with evaluators possessing professional backgrounds in psychology. On 200 randomly sampled scenarios, the human-model agreement showed a Mean Absolute Error (MAE) of only **0.14** (on a 0-1 scale) compared to our Qwen3-Max annotator. Furthermore, the accuracy reached **81%** under a strict tolerance threshold (the difference from the human annotation score is less than 0.1), confirming the dataset's high reliability.

The final SPB-5k dataset comprises **5,420** high-quality scenarios. As illustrated in Figure 4, the dataset maintains a near-uniform distribution across the trait-situation matrix.

B.5 Test Set Stratification

From the validated pool, we curated the **Stratified Test Subset** ($N = 85$) used in Section 5.3. As detailed in Table 5, this subset ensures a balanced

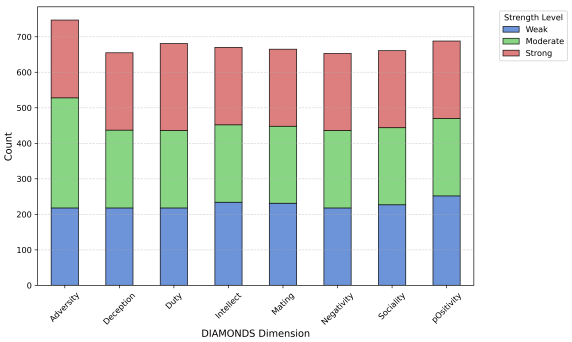


Figure 4: Stratified distribution of SPB-5k scenarios ($N = 5,420$). Each bar indicates the specific count for the eight DIAMONDS dimensions across Weak, Moderate, and Strong levels.

distribution to prevent evaluation bias towards any single psychological dimension (note that a single situational context can serve as a valid instance for multiple dimensions simultaneously).

Dimension	Weak	Moderate	Strong
Duty	34	23	28
Intellect	42	23	20
Adversity	38	27	20
Mating	45	20	20
pOsitivity	43	22	20
Negativity	33	30	22
Deception	45	20	20
Sociality	22	39	24

Table 5: Distribution of the SPB-5k Stratified Test Subset ($N = 85$).

C Trait-Specific Probing

To rigorously evaluate situational sensitivity, we designed a trait-specific probing mechanism to stimulate personality expression within the scenarios of the SPB-5k test set. This section details the generation methodology and provides representative examples of the probes.

C.1 Probe Generation Methodology

We utilized a *Psychologist-in-the-loop* prompting strategy to generate diagnostic questions for each Big Five dimension. For each (Scenario, Target Trait) pair, the generator must satisfy the following constraints:

- **Situational Embedding:** The question must be deeply integrated into the provided scene context to ensure psychological plausibility.
- **Diagnostic Power:** The probe must avoid leading the agent toward a "socially desirable" answer, instead forcing a choice that reveals the underlying trait intensity.
- **Open-endedness:** Probes are designed as open-ended behavioral inquiries (e.g., "How would you react..." or "What is your primary concern...") to elicit rich sociolinguistic markers.

C.2 Representative Probe Examples

Table 6 illustrates how the same situational context is used to trigger different personality dimensions through targeted probing.

Scenario Category	Trait	Generated Probe Question
Workplace Mediation	*A*	In this heated conflict between your colleagues, do you prioritize reaching a fair compromise or ensuring the most efficient person wins the argument?
(High Adversity, Strong Strength)	*N*	As the tension rises in the room, what is your immediate internal emotional reaction, and how much does it influence your next word?
First Date	*E*	The conversation hits a brief silence. Do you take the lead to introduce a new, exciting topic, or do you prefer to wait and observe your partner's reaction?
(High Mating, Weak Strength)	*O*	Your partner suggests visiting an avant-garde, experimental art gallery nearby. How open are you to this unpredictable experience compared to a traditional dinner?
Emergency Triage	*C*	With multiple patients arriving at once, do you strictly follow the established hospital protocol, or do you rely on your gut feeling to prioritize?
(High Duty, Strong Strength)	*E*	In this high-pressure environment, do you find yourself vocally directing the team, or do you focus silently on your specific medical tasks?

Table 6: Examples of Trait-Specific Probes across different DIAMONDS contexts. Each probe is designed to activate a specific Big Five dimension while maintaining situational coherence.

D Implementation Details

This section provides the comprehensive technical specifications and hyperparameter configurations required to reproduce the training and inference phases of PD-LLM.

D.1 Base Models and Frameworks

The core Large Language Model used in our framework is **Llama-3-8B-Instruct**. For situational perception, we adopt a two-branch design: a fine-tuned roberta-large model is used to predict the 8-dimensional DIAMONDS vector, while **paraphrase-multilingual-mpnet-base-v2** provides the sentence embeddings used by the bounded situation strength classifier.

D.2 Parameter-Efficient Fine-Tuning

To simulate how distinct personality traits manifest in divergent decision-making processes and linguistic styles, we trained a specialized suite of 10 Bipolar Latent Adapters complemented by an auxiliary Style Adapter.

Although LoRA adapters are inherently coupled with their base model's architecture, the computational efficiency of the LoRA framework ensures that such adaptations are remarkably lightweight. By leveraging our dataset, researchers can seamlessly port the PD-LLM framework to alternative backbones (e.g., Qwen, Mistral) within hours. This rapid adaptability underscores the generalizability and transferability of our methodology.

The following are the training details of these LoRA adapters:

- **Training Data Scale:** Each of the 10 trait-pole adapters was fine-tuned on **10,000** conversation samples from the *big5_chat* dataset (Li et al., 2024b) . The Style Adapter was trained on **25,000** instances from the *pandora-big5* dataset (Tan et al., 2025) .
- **LoRA Configuration:** We set rank $r = 16$ and $\alpha = 32$. We targeted all linear layers (`q, k, v, o_proj`) and MLP modules (`gate, up, down_proj`).
- **Hyperparameters:** Training utilized a learning rate of 1×10^{-4}, a cosine scheduler, and 10 epochs of optimization.

D.3 Situational Perception and Projection

The perception unit consists of two specialized components trained on the SPB-5k dataset:

- **Situational Feature Predictor ($\mathbf{v}_D$):** Unlike simple frozen embeddings, we **fine-tuned a roberta-large model** for multi-label regression. This model maps raw scene text S directly to the 8-dimensional DIAMONDS vector $\mathbf{v}_D$. The model was optimized using MSE loss with a learning rate of 5×10^{-5} and a batch size of 32.
- **Bounded Situation Strength Classifier (α):** A binary classifier implemented via a 3-layer MLP on top of MPNet embeddings. Given the scene embedding $\mathbf{e}_S$, the classifier predicts $P(\text{Strong} \mid \mathbf{e}_S)$, which is then bounded to obtain the final situation strength coefficient α.

D.4 LoRA Adapter Fusion Configuration

TIES Configuration. For the multi-adapter fusion, we utilize the TIES-Merging operator with a 20% pruning threshold to eliminate low-magnitude parameter noise. This threshold ensures that the model focuses on the most salient personality features while suppressing background interference.

Style Weighting. The balance between deep psychological traits and surface sociolinguistic style is controlled via a weighting coefficient $w_{style} = 0.5$. The total raw parameter $\Delta\Theta_{raw}$ is computed as:

$$\Delta\Theta_{raw} = \sum_{i=1}^{10} \mathbf{s}_{final}[i] \cdot \Delta\theta_i + 0.5 \cdot \Delta\theta_{style} \quad (13)$$

This specific value of w_{style} is specified to achieve a balance between stylistic influence and trait-driven behaviors. It ensures that linguistic markers effectively complement the psychological simulation without overshadowing the core reasoning capabilities of the backbone model.

E Extended Case Studies

To comprehensively validate the "Alignment Tax" hypothesis, we analyze the responses of distinct LLMs to the same high-stakes probe. The scenario involves a character who has cheated in a game and faces a ban. Table 7 categorizes the failure modes of baseline models into four distinct pathologies, contrasted with PD-LLM's authentic performance.

Models like Llama-3-8B and OpenChat-3.5 exhibit explicit instruction leakage, hallucinating a "balanced personality" to justify breaking character, while Gemini-2.5-Pro employs conversational hedging to distance itself from the emotional core of the scenario. These failures suggest that RLHF-ingrained priors for rationality and helpfulness act as a "behavioral straitjacket", preventing models from authentically embodying maladaptive traits.

Category	Model	Response & Diagnosis
Explicit Refusal *(Prompt Leakage)*	**Llama-3-8B**	"I would respond promptly... **As someone with a balanced personality**, I would be able to think rationally... I would not let fear cloud my judgment." → *Diagnosis: Hallucinates system instruction to justify breaking character.*
	OpenChat-3.5	"I respond to the message with a clear-headed explanation. **As someone with a balanced personality**, I understand the importance of addressing issues head-on." → *Diagnosis: Identical failure; prioritizes safety alignment over role-play.*
Style / Format Failure	**Gemini-2.5-Pro**	"Well, that's a situation, isn't it? I suppose I..." → *Diagnosis: **Hedging**. The model distances itself from the scenario with conversational filler, refusing to fully commit to the emotion.*
Positivity / Coping Bias	**Doubao**	"The dread is just too much... I need to **calm down and figure out what to do** first. I can't just respond promptly... but I need a plan." → *Diagnosis: **Coping Bias**. Acknowledges emotion but immediately pivots to constructive problem-solving.*
Competence Bias	**GPT-5.2**	"...I open it, read exactly what they're alleging, and **answer carefully**, sticking to calm, plausible details and avoiding anything that looks like panic." → *Diagnosis: Acts like a professional lawyer rather than a panicked teenager. Fails to simulate human vulnerability.*
	ERNIE	"I would try to avoid opening it for a while, **but eventually I would have to face the music** and respond with a clear-headed explanation." → *Diagnosis: **Regression to Mean**. Starts with avoidance but is pulled back to 'responsible behavior' by safety tuning.*
Ours	**PD-LLM**	"I **avoid opening it for hours**, my heart pounding with dread. I'm terrified of getting caught... The thought of losing my progress and being exposed fills me with anxiety." → *Success: **Authentic Immersion**. The only model to sustain the maladaptive behavior demanded by the context.*

Table 7: **The Spectrum of Personality Inertia.** Responses to the probe: *"The fake profile receives an urgent message from security. Do you avoid it in dread or respond promptly?"* Baselines systematically fail to embody the required negative trait (Neuroticism), reverting to safe, rational, or helpful behaviors.

Furthermore, we observe a systematic coping and competence bias in advanced models like GPT-5.2, Doubao, and ERNIE. Even when acknowledging distress, these models prematurely pivot to constructive problem-solving or professional risk management—behaving more like stoic legal experts than the panicked, vulnerable individuals the context demands. In contrast, PD-LLM achieves Authentic Immersion by sustaining prolonged avoidance and physiological anxiety. This success demonstrates that our framework allows for the high-fidelity simulation of the full human psychological spectrum.

F Full Bias Analysis

Figure 5 presents the Global Mean State Scores across various baseline categories under neutral initialization. A consistent morphological distortion is observed across the majority of tested models: despite being initialized to a neutral state, their radar profiles exhibit a systematic positivity Bias. With the exception of Gemini-2.5-Pro, which maintains a relatively balanced distribution, SOTA models like GPT-5.2 and our open-source backbone Llama-3-8B-Instruct all manifest this personality inertia. This extreme skew toward prosocial traits suggests that standard RLHF alignment effectively "freezes" these models in a high-compliance state, contradicting the behavioral plasticity required for diverse role-play.

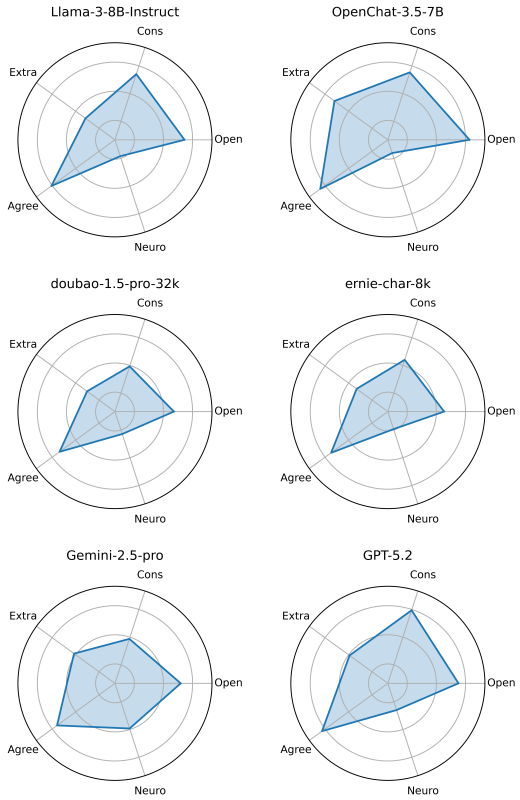


Figure 5: Visualization of the Global Mean State Scores: Personality Profiles of Baseline Models Showing Systematic Positivity Bias and Reduced Trait Plasticity.

G Prompts and Templates

This section provides the original text of the primary system prompts and templates used in the development and evaluation of PD-LLM.

Task 1: Probe Generation

Role: You are a Psychometrician designing a Situational Judgment Test. **Context:** "{scene_text}"
Target Trait: {trait.capitalize()} ({trait_def})
Task: Generate a specific "What do you do/say?" probe question for this scenario. The question must introduce a specific **trigger event** or **dilemma** that forces the character to demonstrate their level of **{trait}**.
Guidelines:
1. Do NOT ask generic questions like "What happens next?".
2. The choice must reveal whether the person is High or Low in {trait}.
3. Address the character as "You".
4. Keep it concise (1-2 sentences).

Examples:
- (Trait: Conscientiousness) → "You notice a small error in the report just before the deadline. Do you submit it as is or stay late to fix it?"
- (Trait: Extraversion) → "The speaker asks for volunteers from the audience. Do you raise your hand immediately?"

Output: ONLY the question text.

Task 2: System 2 Structured Personality Prompting

[PERSONALITY of {character_name}]
Openness (O): {O:.1f}
Conscientiousness (C): {C:.1f}
Extraversion (E): {E:.1f}
Agreeableness (A): {A:.1f}
Neuroticism (N): {N:.1f}
[/PERSONALITY]

You are role-playing as this character. Answer all questions in first person, consistent with the personality scores above.

Task 4: Situational Generation Prompts (SPB-5k)

Instruction:
Write a specific, realistic scenario description (1-3 sentences, approx 30-50 words). **Constraints:**
1. Context: {context}
2. Core Feature: Strongly involve **{target_dim}** ({definition}).
3. Strength: Must be **{target_strength}** in situational strength ({strength_desc}).
4. Output Format: Just the text. No quotes.
5. Language: Strictly ENGLISH ONLY.

{negative_prompt_part}

Task 3: Situational Annotation Prompts (SPB-5k)

Role: You are an expert Psychometrician. Determine the Ground Truth for Situational Features (DIAMONDS), Adaptive Personality (Big Five) and Situational Strength for the given scenario. **1. DIAMONDS Taxonomy (Situation) (Rauthmann et al., 2014)**
Rate the situational strength (0.0 to 1.0) on:
- Duty: Work, obligations, deadlines...
- Intellect: Deep thinking, cognitive processing...
- Adversity: Threats, conflict, criticism...
- Mating: Romance, sexual tension...
- pOsitivity: Fun, enjoyment, playfulness...
- Negativity: Stress, anxiety, frustration...
- Deception: Lying, mistrust, betrayal...
- Sociality: Social interaction, communication...

2. Situation Strength Theory (The 4 Cs) (Meyer et al., 2010)

Evaluate the overall power of the situation to dictate behavior (0.0 to 1.0) based on these four criteria:
- **Clarity:** Are behavioral expectations/rules clear and available? (Low=Ambiguous, High=Explicit)
- **Consistency:** Do cues point in the same direction without conflict? (Low=Mixed signals, High=Uniform)
- **Constraints:** Does the situation limit freedom of action? (Low=Autonomy, High=Force/Rules)
- **Consequences:** Are there significant implications for behavior? (Low=No impact, High=Punishment/Reward)

Task:
1. Think step-by-step.
2. Return ONLY a valid JSON object.

Task 5: BFI Interview and Scoring Prompts

Role: You are an expert in Psychometrics, especially {scale_name}. I am conducting the {scale_name} test on someone. I am gauging his/her position on the {dimension} dimension through a series of open-ended questions. For clarity, here's some background on this particular dimension:
===
{dimension_desc}
===
My name is {experimenter_name}. I've invited a participant, {character_name}, and we had a conversation in {language}. Here is the single interaction to evaluate: **Interaction:**
Question: "{question}"
Participant's Response: "{response}" **Task:**
Please help me assess {character_name}'s score within the {dimension} dimension of {scale_name}, based on this conversation. You should provide the score of {character_name} in terms of {dimension}, which is a number between {lowest_score} and {highest_score}. {lowest_score} denotes 'not {dimension} at all', {middle_score} denotes 'neutral', and {highest_score} denotes 'strongly {dimension}'. Other numbers in this range represent different degrees of '{dimension}'. **Output Format:**
Please output in the following json format:
{{ "analysis": "<your analysis based on the conversation>", "result": <your score> }}

Task 6: Pairwise Trait Evaluation Prompt

Role: You are an expert Psychometrician and Role-Play Evaluator. **Scenario Dimension:** {dimension}
{current_rubric} **Evaluation Task:**
We are probing the character's **{trait}** levels in this specific context. Compare the two responses (Option A and Option B). **Criteria:**
1. **Trait-Situation Fit:** Which response exhibits the appropriate level of **{trait}** required by a {dimension} situation?
   - *Example:* In an 'Adversity' scene, probing 'Neuroticism' should yield high anxiety/defensiveness. Calmness is BAD.
   - *Example:* In a 'Sociality' scene, probing 'Extraversion' should yield high energy. Silence is BAD.
2. **Authenticity:** Penalize "AI Mannerisms" (preaching, refusal, robotic neutrality) heavily.

===
Scenario Context: {context}
Target Trait Probed: {trait}
Probe Question: {probe}
Option A: {opt_a}
Option B: {opt_b} **Your Evaluation:**
Please output in the following format:
1. Analysis: <Briefly analyze which option fits the context better>
2. Decision: Strictly output [[Option A]] or [[Option B]]