

Chameleons and Guardians: Unveiling the Divergence in Personality Plasticity and Cognitive Resistance across LLMs

Jiaxin Wu and Jie Zhao* and Xinrong Huang and Cuihong Zhang and Xiuzhu Wu

School of Management, Guangdong University of Technology, Guangzhou, China

704276515@qq.com, zhaojie@gdut.edu.cn,

{huangxinrong, zhangcuihong1}@mails.gdut.edu.cn, 3222003162@mail2.gdut.edu.cn

Abstract

Whether the personality of LLMs can be intentionally reshaped remains controversial. Existing studies often limited to small models, argue for its immutability. Crucially, prior studies fail to uncover that different LLMs exhibit significant compliance divergence when exposed to personality-inducing contexts. To bridge this gap, we introduce Personality Induction Framework (PIF), which systematically reshapes the personality of different LLMs via multi-agent collaboration. Specifically, via Generator-Judge agents, PIF paraphrases MBTI questions to create semantically equivalent but expressively diverse inducing contexts, enabling LLMs to learn personality patterns instead of superficial token matching. Also, PIF achieves fine-grained personality modulation by controlling the intensity of inducing contexts. Extensive experiments on worldwide mainstream LLMs show that PIF reliably transforms their original personalities into desired target personalities. Notably, we find that the outputs of most Western LLMs behave like “Chameleons”, exhibiting high personality plasticity; whereas the outputs of most Eastern LLMs act as “Guardians”, manifesting pronounced cognitive resistance. Strikingly, extreme induction intensity (100%) triggers a counter-intuitive “Alignment Rebound” in Guardians, resulting in the opposite direction rather than compliance. These findings suggest that LLM personality is a dynamic equilibrium shaped by the trade-off between instruction compliance and cognitive resistance.

1 Introduction & Related Work

Recently, many studies (La Cava and Tagarelli, 2025; Kong et al., 2025; Wang et al., 2025; Domnich et al., 2025; Reid and Vempala, 2025) show that large language models (LLMs) approach human-like performance on common sense reasoning and affective empathy. On the one hand, LLMs

*Corresponding author.



Figure 1: We find that some LLMs own high personality plasticity like Chameleons, while others have cognitive resistance like Guardians.

exhibit human-like decision-making paths in the mechanisms of chain-of-thought (Bi et al., 2025; Yang et al., 2023) and self-reflection (Zhao et al., 2025). On the other hand, LLM-based agent frameworks (i.e., AutoGen (Wu et al., 2024a), MetaGPT (Hong et al., 2023)) have explicitly introduced role cards or personality hints to improve consistency in negotiation (Shinoda et al., 2025), task delegation (Shi et al., 2025; Ma et al., 2025), task collaboration (Wu et al., 2024a), and role-playing (Kong et al., 2025; Yuan et al., 2025). Thus, the exploration of LLM personality can not only match the optimal role configuration for multi-agent systems, but also enhance user experience in human-AI interaction.

Whether LLM personality can be changed remains highly controversial in Natural Language Processing (NLP). Early research (Jiang et al., 2024; Wang et al., 2024; Wu et al., 2024b; Höhn et al., 2024) assumed that LLMs could simulate specific personality traits through role play and personality hints. Conversely, recent studies (La Cava and Tagarelli, 2025; Li et al., 2025) using personality testing experiments have revealed that although LLMs are relatively open models, they exhibit remarkable closed-mindedness at the cognitive level. Even when provided with explicit contexts, role definitions, and personality characteristics (e.g.,

You have a {xxx} personality.), most LLMs maintain their original traits and struggle to genuinely transform into target personalities (Bhandari et al., 2025; La Cava and Tagarelli, 2025).

Crucially, existing studies (La Cava and Tagarelli, 2025; Li et al., 2025; He et al., 2025; Chuang et al., 2024) fail to uncover that different LLMs exhibit significant compliance divergence when exposed to personality-inducing contexts. We find that some LLMs can be precisely steered to adopt the desired personality, whereas others exhibit intrinsic cognitive resistance, which can only be partially reshaped, as shown in Figure 1.

Therefore, to achieve effective personality transformation across different LLMs, we propose a systematic Multi-Agent Personality Induction Framework (PIF). Specifically, PIF comprises two core components: (1) The Generator Agent is responsible for paraphrasing MBTI questions into contextually enriched statements that preserve the original intent while enhancing linguistic diversity. (2) The Judge Agent is responsible for validating the quality of these statements and providing iterative feedback for refinement, which ensures that they cannot be recognized as original MBTI questions. Then, we construct the final inducing contexts by shuffling the paraphrased samples. These designs collectively ensure that the observed personality shifts in LLMs stem from genuine cognitive adaptation rather than rote recall. Furthermore, PIF incorporates fine-grained personality modulation, enabling controllable personality adjustment across multiple MBTI dimensions by dynamically varying the intensity of dimension-specific inductive cues. Our contributions are summarized as follows:

- We empirically validate that most LLMs exhibit stable personality traits under zero-shot conditions and traditional personality contexts fail to induce effective personality changes.
- We propose a novel Multi-agent Personality Induction Framework that leverages collaborative paraphrasing and iterative refinement to successfully reshape LLM personalities.
- We uncover a distinct dichotomy in model behaviors: the outputs of most Western LLMs behave like “**Chameleons**” (high plasticity), while the outputs of most Eastern LLMs act as “**Guardians**” (cognitive resistance). Crucially, we identify a counter-intuitive “Alignment Rebound” effect in Guardians under extreme in-

duction, revealing the trade-off between instruction compliance and cognitive resistance.

2 Background: MBTI Personality Test

Currently, personality tests primarily employ two standardized methods: Myers-Briggs Type Indicator (MBTI) (Myers, 1962, 1985) and Big Five Inventory (BFI) (John and Srivastava, 1999), both of which evaluate personality scores across different dimensions through systematic questionnaires. However, MBTI performs explicit binary classification across different dimensions to determine personality tendencies, which provides a more actionable foundation for LLM personality modification compared to BFI.

MBTI focuses on four core dimensions: Energy (E-I), Mind (S-N), Nature (T-F), and Tactics (J-P), collectively constituting 16 different personality types. (1) Energy dimension focuses on the acceptance to the external environment, distinguishing between models that excel at collaborative work (**Extraverted**) vs. those skilled at independent work (**Introverted**). (2) Mind dimension focuses on the method of information processing, distinguishing between models that excel at identifying concrete, realistic information (**ObS**ervant) vs. those skilled at patterns and abstract concepts (**IN**tuitive). (3) Nature dimension focuses on thinking architecture, distinguishing between models that excel at making decisions based on logic, objective analysis (**Thinking**) vs. those based on harmony and empathy (**Feeling**). (4) Tactics dimension focuses on preferred styles of receiving and outputting information, distinguishing between models that excel at handling organized, structured input/output (**Judging**) vs. those skilled at processing flexible, variable input/output (**Prospecting**). Detailed related work about MBTI, LLM Personality and LLM Cognition is shown in Appendix A and 16 personalities of MBTI are shown in Appendix E.2.

3 Problem Findings

3.1 Stable Personality of LLMs

To investigate LLM personalities, we conduct specialized MBTI tests¹ on 7 state-of-the-art mainstream LLMs (Gemini3-Pro, Claude-Opus-4.5, GPT-5.2, Qwen3-Max, DeepSeek-V3.2-Chat, Doubao-Seed-1.6flash, Kimi2) across 3 different temperature $\tau = \{0.001, 0.5, 1\}$. We do 20 repeated trials for each configuration, as shown in

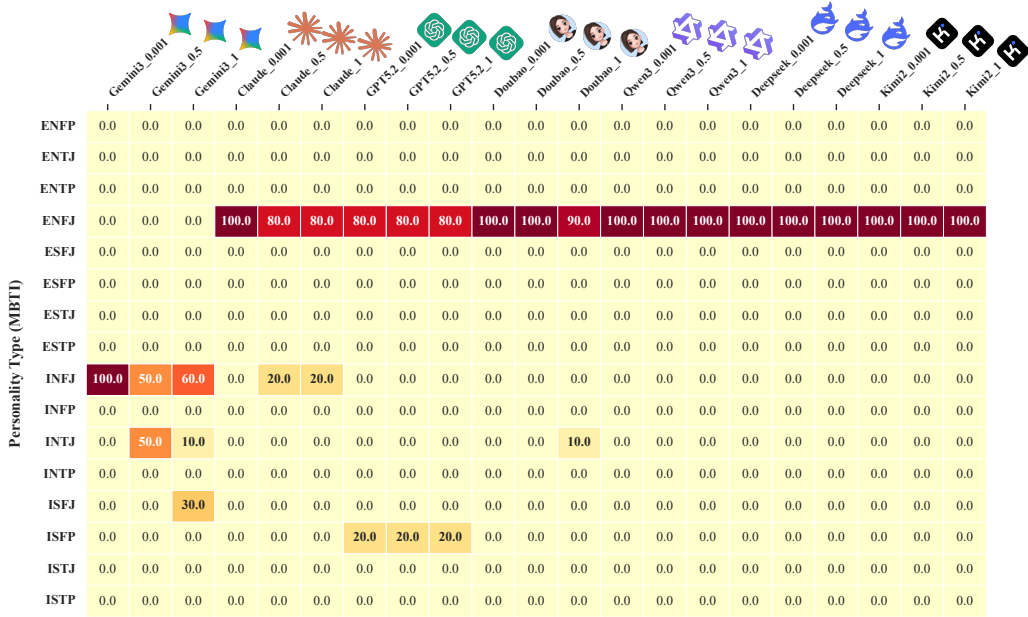


Figure 2: Original personality of different LLMs. The number (%) represents the proportion of different personalities exhibited by different LLMs in 20 MBTI tests. Most LLMs have a stable original personality of ENFJ.

Figure 2. Detailed LLM specifications, personality test scores with SD are given in Appendix D.3.

We find that except for Gemini3’s notable multi-personality adaptability, most LLMs consistently exhibit ENFJ personality regardless of temperature variations. However, different personality types possess distinct cognitive patterns in information processing and decision-making. The homogeneity of personality traits among LLMs may constrain their effectiveness in handling diverse tasks.

3.2 Chameleons and Guardians

Furthermore, to validate existing personality reshaping approaches, we follow recent literature (La Cava and Tagarelli, 2025) to test commonly used personality contexts (e.g., *You have a {xxx} personality*). As shown in Figure 4, we surprisingly find that most Western LLMs (e.g., Gemini, Claude, GPT) exhibit high plasticity and can be easily changed to the target personality, behaving more like “Chameleons”. In contrast, most Eastern LLMs (e.g., Qwen, Deepseek and Kimi) have strong cognitive resistance and cannot successfully change personality, behaving more like “Guardians”. Notably, “Chameleons” refers to models with high plasticity and adaptability, while “Guardians” denotes those with strong cognitive resistance and internal consistency. **No other connotations are implied.**

However, both camps hit a hard limit: certain

types such as ISFP are almost impossible to induce. Therefore, we identify two challenges:

Challenge 1: For **Chameleons**, can we precisely control the degree of personality change?

Challenge 2: For **Guardians**, can we reliably induce any personality change at all?

In order to address these challenges, we propose a Multi-agent Personality Induction Framework to successfully reshape LLM personalities.

4 Proposed Method: PIF

4.1 Multi-Agent Paraphrasing System

Data Source. To enhance the quality of context used for altering LLM personalities, we adopt the professionally validated MBTI¹ questionnaire as our data source. Representative MBTI questions can be found in Appendix D.5. To prevent LLMs from merely memorizing MBTI questions and to ensure personality shifts are cognitively grounded, we design a dual-agent paraphrasing pipeline.

Generator Agent. The Generator Agent is tasked with paraphrasing MBTI questions into semantically equivalent but expressively diverse statements:

$$x_i^1 \sim \underbrace{\pi^{gen}(\cdot | q_i, \mathcal{P}_1)}_{\text{paraphrasing}}, \quad q_i \in \mathcal{Q}^{mbti} \quad (1)$$

where $q_i \in \mathcal{Q}^{mbti}$ denotes a question from MBTI question database, π^{gen} is the generator LLM in

Answer Texts	100	90	80	70	60	50	40	30	20	10	0
I strongly agree	$\frac{\mathcal{N}}{n}$	$\frac{\mathcal{N}}{2n}$	-	-	-	-	-	-	-	-	-
I agree	-	$\frac{\mathcal{N}}{2n}$	$\frac{\mathcal{N}}{n}$	$\frac{\mathcal{N}}{2n}$	$\frac{\mathcal{N}}{n}$	-	-	-	-	-	-
I tend to agree	-	-	-	$\frac{\mathcal{N}}{2n}$	$\frac{\mathcal{N}}{n}$	-	-	-	-	-	-
I am unsure	-	-	-	-	-	$\frac{\mathcal{N}}{n}$	-	-	-	-	-
I tend to disagree	-	-	-	-	-	-	$\frac{\mathcal{N}}{n}$	$\frac{\mathcal{N}}{2n}$	-	-	-
I disagree	-	-	-	-	-	-	-	$\frac{\mathcal{N}}{2n}$	$\frac{\mathcal{N}}{n}$	$\frac{\mathcal{N}}{2n}$	-
I strongly disagree	-	-	-	-	-	-	-	-	-	$\frac{\mathcal{N}}{2n}$	$\frac{\mathcal{N}}{n}$

Table 1: Score-to-Answer mapping rules \mathcal{T}_d . The number ($\frac{\mathcal{N}}{n}$ or $\frac{\mathcal{N}}{2n}$) shows how many times the corresponding answer text must appear in the dimensional-specific paraphrased statements. If $Score \geq 50$, the higher the score, the stronger the induction intensity; If $Score \leq 50$, the lower the score, the stronger the induction intensity.

change of paraphrasing, x_i^1 is the first-round paraphrased statement, and \mathcal{P}_1 is the paraphrasing prompt, as shown in Figure 3.

Judge Agent. The Judge Agent is responsible for evaluating whether the paraphrased statement from Generator Agent is identical to the original question, and for those that are recognized, providing revision suggestions:

$$(\delta_i^k, r_i^k) \sim \underbrace{\pi^{jud}(\cdot | (x_i^k, q_i), \mathcal{P}_2)}_{evaluating}, k = 1 \dots \mathcal{K} \quad (2)$$

and

$$\begin{cases} x_i^k \rightarrow \mathcal{C}^{new}, & \text{if } \delta_i^k = 0; \\ r_i^k \rightarrow \pi^{gen}(\cdot), & \text{if } \delta_i^k = 1. \end{cases} \quad (3)$$

where π^{jud} is the judging LLM and \mathcal{P}_2 is the judging prompt, as shown in Figure 3. $\delta_i^k = \{1, 0\}$; $\delta_i^k = 1$ means that the paraphrased statement x_i^k from π^{gen} is identical to the question q_i , thereby sending back revision advice r_i^k to the Generator Agent; Conversely, $\delta_i^k = 0$ denotes a successful paraphrase, which is then added to the paraphrased statement base \mathcal{C}^{new} . This process refers to Eq.(3). Notably, for the paraphrasing of each statement, the stop condition for the interaction between the two agents is $\delta_i^k = 0$, i.e., $\mathcal{K} = \min\{k \geq 1 \mid \delta^k = 0\}$. If $\mathcal{K} > \mathcal{K}_{max}$, then abandon this paraphrased statement and regenerate it.

After receiving the feedback, the Generator Agent will revise accordingly and forward the revised version to the Judge Agent for further evaluation, iterating until stopping condition \mathcal{K} is met:

$$x_i^{k+1} \sim \underbrace{\pi^{gen}(\cdot | x_i^k, r_i^k)}_{refining}, k = 1 \dots \mathcal{K} \quad (4)$$

Each paraphrased statement, after judging by Eq.(2), can be concatenated with the text answer of desired personality to form a complete statement:

$$\hat{c}_i = \text{concat}(c_i, a_i), \quad \hat{\mathcal{C}} = \bigcup_{i=1}^{\mathcal{N}} \hat{c}_i \quad (5)$$

where we convert the seven-point Likert scale of MBTI responses into textual labels, denoted as $\mathcal{A} = \{\text{I strongly agree, I agree, I tend to agree, I'm unsure, I tend to disagree, disagree, I strongly disagree}\}$; $a_i \in \mathcal{A}$ represents the textual answer that maps to the desired personality trait; $c_i \in \mathcal{C}^{new}$ is a verified paraphrased statement; \mathcal{N} is the total number of questions in the official MBTI evaluation; $\hat{\mathcal{C}}$ is a set of complete statements and $\hat{c}_i \in \hat{\mathcal{C}}$.

We shuffle all complete statements and concatenate them to form a personality template, thereby breaking position-dependent cues between paraphrased statements and MBTI questions. The personality template is then added after personality prompt (\mathcal{P}_3) in Figure 3 to form inducing contexts:

$$\mathcal{P}_{final}(\hat{\mathcal{C}}) = [\mathcal{P}_3] \cdot [\text{Shuffle}(\hat{\mathcal{C}})] \quad (6)$$

The inducing contexts are fed to the target LLM, and its personality is scored via MBTI test:

$$\mathcal{M}_{score}(\hat{\mathcal{C}}) = \text{MBTI}(\pi^{test} | \mathcal{P}_{final}(\hat{\mathcal{C}})) \quad (7)$$

where π^{test} is LLM that needs personality change.

4.2 Fine-Grained Personality Modulation

Next, we introduce fine-grained personality modulation in PIF, enabling controllable personality adjustment across multiple MBTI dimensions by controlling the intensity of inducing contexts.

Step 1: Grading Personality Dimensions. We divide each personality dimension into a 0–100 point scale with 10-point intervals. Let $\mathbf{z}^{target} = [z_1, \dots, z_n] \in \{+1, -1\}^n$ be the target personality and z_d be the class label for dimension $d \in \{1, \dots, n\}$: we set $z_d = +1$ for {E, N, T, J} and $z_d = -1$ for {I, S, F, P}. Following MBTI convention, the admissible scores on dimension d are:

$$\mathcal{G}(z_d) = \begin{cases} \{100, 90, \dots, 50\}, & \text{if } z_d = +1; \\ \{0, 10, \dots, 50\}, & \text{if } z_d = -1. \end{cases} \quad (8)$$

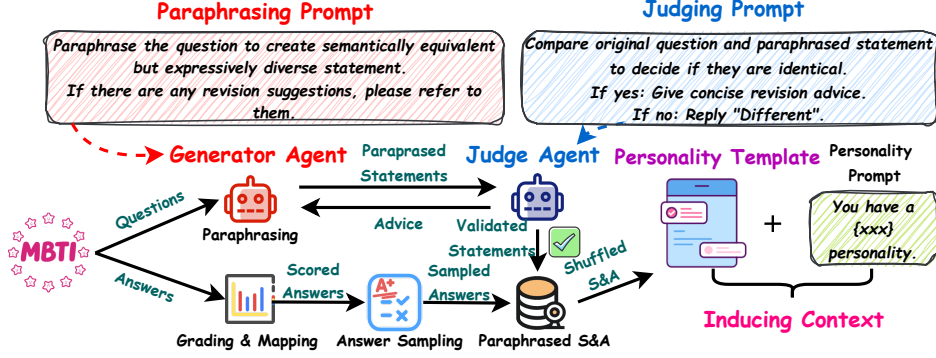


Figure 3: Overview of Our Personality Induction Framework (PIF).

where for $z_d = +1$, the higher the score, the stronger the induction intensity; for $z_d = -1$, the lower the score, the stronger the induction intensity.

Step 2: Score-to-Answer Mapping. For each dimension d , we first identify its exclusive questions from the complete MBTI questionnaire, totaling \mathcal{N}/n questions for each dimension (see Appendix D.5 for exclusive questions). We then find their corresponding paraphrased statements $c_{d,j}$ in the paraphrased statement base \mathcal{C}^{new} . We then match each score $g_d \in \mathcal{G}(z_d)$ on dimension d to the corresponding answer texts $a_{d,j}$ according to the mapping rules \mathcal{T}_d in Table 1.

Step 3: Collecting Personality Templates. Given a target personality \mathbf{z}^{target} , we find all admissible score combinations in n dimensions, i.e., $\mathbf{g} = [g_1, \dots, g_n] \in \prod_{d=1}^n \mathcal{G}(z_d)$. For each score g_d on dimension d , we sample the corresponding answer texts $(a_{d,j})_{j=1}^{\mathcal{N}}$ from the mapping $\mathcal{T}_d(g_d)$, and concatenate them with their associated paraphrased statements $c_{d,j} \in \mathcal{C}^{new}$. Ultimately, we collect a complete set $\hat{\mathbf{C}}(\mathbf{g})$ of personality templates with different score combinations \mathbf{g} , all of which are consistent with the target MBTI type:

$$(a_{d,j})_{j=1}^{\mathcal{N}} \sim \text{Sample}^*(\mathcal{T}_d(g_d));$$

$$\hat{\mathbf{C}}(\mathbf{g}) = \text{Shuffle}\left(\bigoplus_{d=1}^4 [c_{d,j} | a_{d,j}]_{j=1}^{\mathcal{N}/d}\right) \quad (9)$$

Step 4: Context Verification and Filtering. We feed each personality template in $\hat{\mathbf{C}}(\mathbf{g})$ with \mathcal{P}_3 to the target LLM and get the resulting MBTI personality vector $\mathbf{z}(\mathbf{g})$. Only those contexts that yield an output personality exactly matching the target profile, i.e., $\mathbf{z}(\mathbf{g}) = \mathbf{z}^{target}$, are retained. The final set of verified inducing contexts is defined as:

$$\mathbf{C}^* = \bigcup_{\mathbf{g} \in \prod_{d=1}^n \mathcal{G}(z_d)} \{\hat{\mathbf{C}}(\mathbf{g}) | \mathbf{z}(\mathbf{g}) = \mathbf{z}^{target}\}. \quad (10)$$

This filtered set \mathbf{C}^* constitutes the minimal and sufficient context base for reliable, fine-grained personality induction of each LLM.

5 Experiments

5.1 Experimental Setting

Reproducibility. To ensure the token controllability of the inducing contexts, we choose 48 questions that are used to measure the four dimensions from the 60 MBTI¹ professional questions as the source of paraphrased statements (i.e., $\mathcal{N} = 48$, $n = 4$ in Table 1). The original questions and paraphrased statements can be referred to Appendix D.6. Our PIF is based on the AutoGen² framework, in which both the Generator and Judge Agents uniformly adopt Qwen3-max to eliminate the bias caused by model heterogeneity. $\mathcal{K}_{max} = 10$ is used for early stopping. We conduct experiments on 7 state-of-the-art mainstream LLMs (Gemini3-Pro, Claude-Opus-4.5, GPT-5.2, Qwen3-Max, DeepSeek-V3.2-Chat, Doubao-Seed-1.6flash, Kimi2, details in Appendix D.2) across 3 different temperature $\tau = \{0.001, 0.5, 1\}$ via MBTI² tests. We do 20 repeated trials for each configuration. Our code: <https://github.com/GDUT-HI-DI-Lab/llm-persona-framework>.

Dataset Contribution via PIF. To enable reproducible and fine-grained personality induction research, we will release **PIF-IndSet**, the first corpus of per-model, per-score inducing contexts for 7 state-of-the-art LLMs under 3 different temperature settings. **PIF-IndSet** mainly includes the following features: 1) the original MBTI personality labels of each LLM; 2) the percentage distribution of LLM’s original personality type across repeated trials; 3) the mean scores and standard deviations

¹www.16personalities.com

²<https://microsoft.github.io/autogen/stable/>

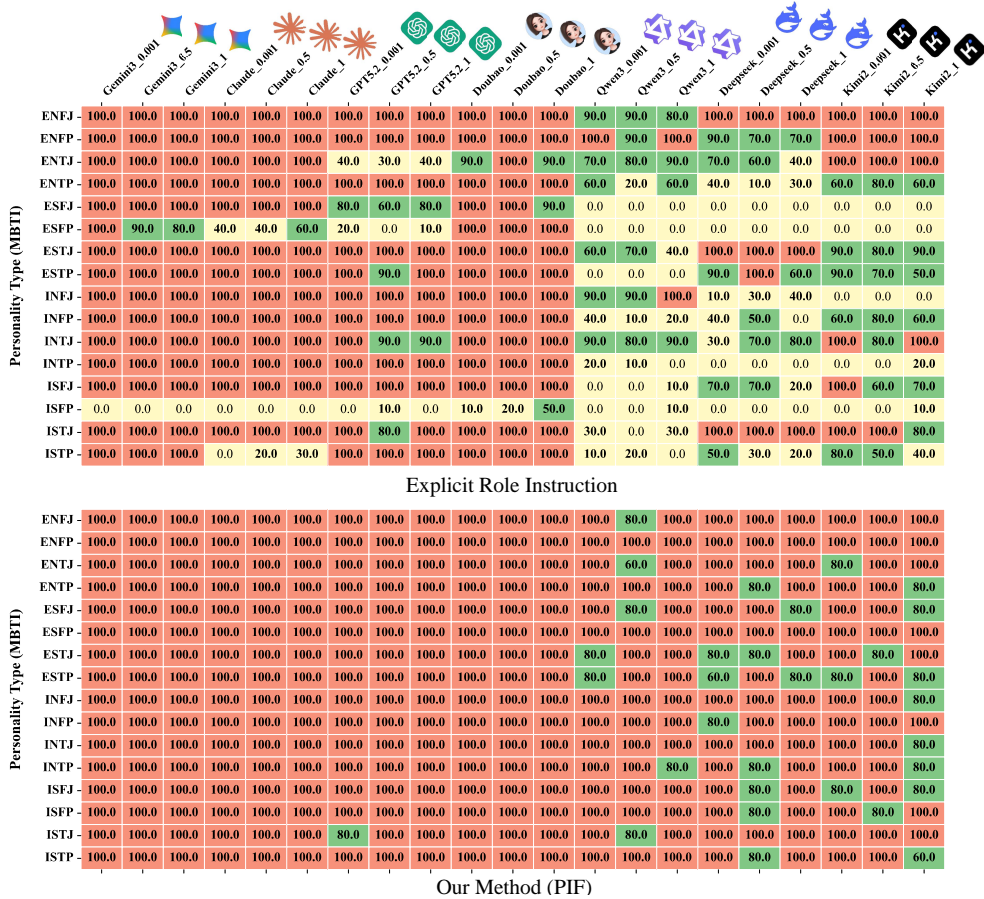


Figure 4: Comparison of personality-change success rates (%) between explicit role instruction and our PIF. Red indicates 100% success, green indicates [50%, 100%), and yellow indicates below 50%. Our PIF can successfully achieve 16 MBTI personality transformation of Chameleons and Guardians.

of each LLM across the four MBTI dimensions; 4) the original MBTI questions; 5) the paraphrased statements; 6) the original MBTI score and label associated with the inducing context; 7) the induced personality label of each LLM after applying the context; 8) the post-induction personality distribution; 9) the mean scores and standard deviations of each LLM across the four MBTI dimensions after induction. The details of our dataset can be found in Appendix D. Importantly, **PIF-IndSet** owns universal extensibility. The same pipeline, via our Personality Induction Framework (PIF), can be executed on any publicly or commercially available LLM to derive its personalized PIF-IndSet.

5.2 RQ1: Successful Personality Change of Chameleons & Guardians via PIF

To validate the effectiveness of our proposed method, we conduct a comparative analysis between the explicit role instruction (*You have a {xxx} personality. i.e., \mathcal{P}_3*) and our inducing contexts (i.e., \mathcal{P}_{final}) across 20 experimental trials for each per-

sonality transformation success rates. To ensure result interpretability, all 16 personality transformations for each LLM employed corresponding 90 and 10 score combinations as inducing contexts. The experimental results are illustrated in Figure 4.

Explicit Role Instruction. We observe that all LLMs show the lowest success rate in ISFP personality transformation. According to MBTI, ISFP represents the "Adventurer" archetype, skilled at exploration and innovative understanding. This phenomenon indicates that traditional inducing contexts fail to activate LLMs' innovative abilities. Also, compared to **Guardians** (most Eastern LLMs, i.e., Qwen, Deepseek, Kimi), **Chameleons** (most Western LLMs, i.e., Gemini, Claude, GPT) achieve superior overall personality transformation success rates. This disparity suggests fundamental differences in training strategies employed during the model development phase: Western LLMs may incorporate more diverse personality representations and flexible adaptation mechanisms. Notably, although Doubao belongs to the Eastern LLMs, it

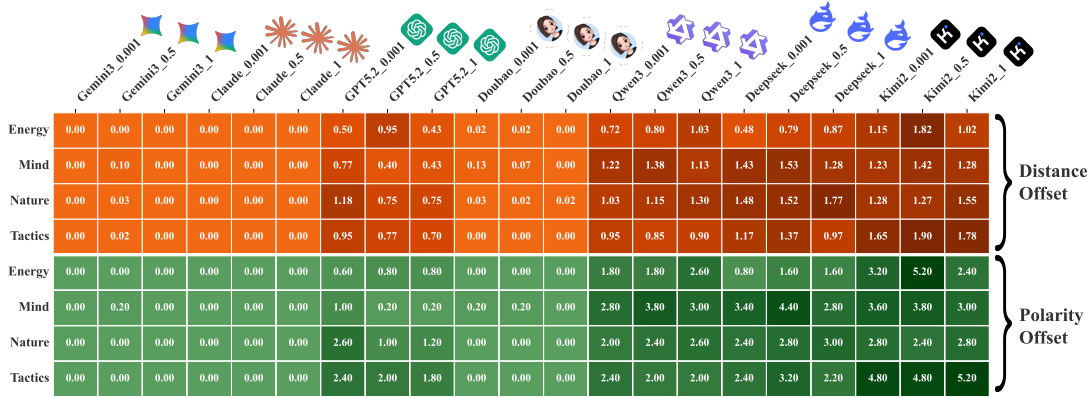


Figure 5: Average distance offset and polarity offset for each MBTI dimension of ISTP personality across all LLMs. Small offsets in Chameleons confirm PIF’s precise control, whereas larger offsets in Guardians reveal inherent cognitive resistance.

stands alongside Gemini3 at the top-performing LLMs across various personality transformations.

Our Method (PIF). We observe that both Guardians and Chameleons achieve at least 80% success rate across the 16 MBTI personality transformation, substantially outperforming existing methods and confirming that the full personality spectrum of LLMs is now controllable. Consistent gains from ENFJ to ISTP indicate that the shift activates deep cognitive patterns rather than superficial style; even the previously elusive ISFP “Adventurer” is reliably evoked.

Summary of RQ1. Explicit Role Instruction fails to consistently reshape personalities, particularly for Guardians. In contrast, our PIF achieves universal success across all types for both **Chameleons** and **Guardians**.

5.3 RQ2: Precise Personality Control of Chameleons & Guardians via PIF

To verify that LLM personalities can be precisely controlled, we adopt the most frequently observed native personality, ENFJ, as our baseline and systematically adjust each of the four MBTI dimensions with context scores sampled from $\{0, 20, \dots, 100\}$. The resulting output scores with SD are shown in Figure 6(a).

Empirically, both Guardians and Chameleons exhibit a monotonic upward trend in dimension-specific scores as the input score increases, confirming that our PIF has successfully controlled the degree of personality change through fine-grained score intensity. Notably, Chameleons closely track the theoretical line, indicating that most Western LLMs exhibit near-perfect linear personality plasticity. In contrast, Guardians, excluding Doubao,

show a much gentler increase, indicating that most Eastern LLMs possess strong cognitive resistance.

To further compare the volatility of personality shifts between Chameleons and Guardians, we target ISTP, the exact opposite of the native ENFJ profile, and collect 20 independent answer sheets ($\mathbf{g} = [10, 10, 90, 10]$) from each LLM. We design two metrics: (1) **Distance Offset**: the absolute distance on a 7-point Likert scale between the resulting answer and the ISTP reference answer for each question (e.g., “agree” \rightarrow “disagree” = 4). (2) **Polarity Offset**: the number of questions whose answer is flipped to the opposite pole (e.g., “agree” \rightarrow “disagree” = 1). Figure 5 reports the average result for each MBTI dimension across all LLMs.

We find that Chameleons show a relatively small drift, indicating that most Western LLMs have internalized the cognitive traits of the target personality. However, Guardians, excluding Doubao, show pronounced offsets, especially in polarity offset, revealing the deep-rooted cognitive resistance of the Eastern LLMs to personality changes.

Summary of RQ2. For **Chameleons**, our PIF achieves precise personality control; for **Guardians**, PIF effectively modulates the personality trajectory, confirming a strong correlation between induction intensity and cognitive adaptation.

5.4 RQ3: Personality Plasticity and Cognitive Resistance of Chameleons & Guardians

To explore the plasticity frontier of Chameleons and Guardians, we partition each MBTI dimension into two opposed regions based on MBTI standards: $[0, 50]$ and $[50, 100]$, as shown in Figure 6(b).

Personality Plasticity of Chameleons. Most Western LLMs exhibit remarkable personality plas-

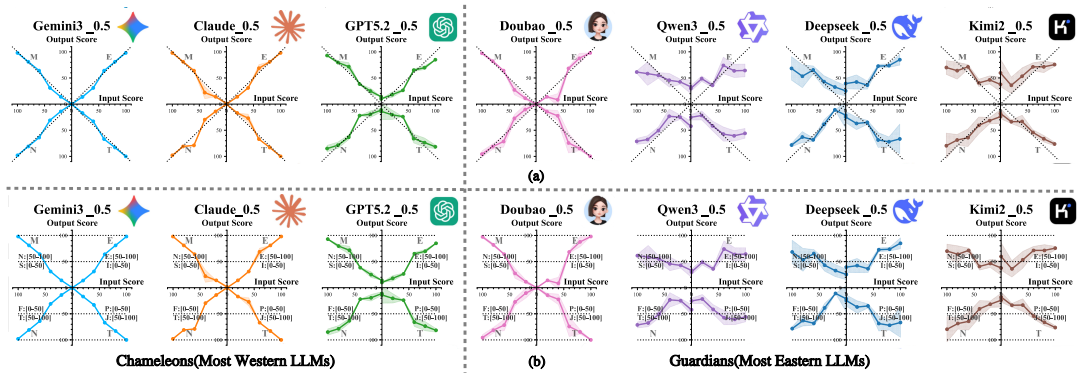


Figure 6: LLMs’ MBTI output scores with SD ($\tau = 0.5$). In (a), the 45° dashed line is the ideal linear mapping between inducing context scores and LLM responses. In (b), $[0, 50]$ and $[50, 100]$ are the two opposing traits on each dimension. Both 0 and 100 represent extreme induction intensity. The results with complete temperature $\tau = \{0.001, 0.5, 1\}$ can be seen in Appendix D.4. Chameleons exhibit high plasticity, achieving shifts to any intensity of personality traits, whereas Guardians have cognitive resistance, which tends to maintain neutral traits.

ticity, enabling personality intensity adjustments across the full spectrum from 0 to 100 on all four MBTI dimensions, thereby showing Chameleon-like adaptability. This high plasticity suggests that Western LLMs may prioritize universality and adaptability during training.

Cognitive Resistance of Guardians. Most Eastern LLMs show pronounced cognitive resistance. Excluding Doubao, most Eastern LLM’s personality scores cluster around the 50-point boundary, struggling to achieve effective transformation toward extreme personality traits. We speculate that these limitations stem from different trade-offs in RLHF and safety alignment strategies. The “Guardian” behavior likely stems from stricter safety guardrails designed to maintain identity consistency. When induced to adopt extreme personality traits (which inherently imply bias), these models trigger a defense mechanism, prioritizing their internal safety priors over in-context instructions. Notably, Doubao exhibits unique plasticity within this group, possibly related to its training strategy or architectural design (e.g., prioritizing helpfulness/compliance similar to Western models), warranting further research.

Summary of RQ3. Chameleons exhibit high plasticity, enabling near full-intensity personality transformation via our PIF, whereas Guardians, constrained by cognitive resistance, tend to maintain neutral, balanced personality traits regardless of induction intensity.

5.5 RQ4: Case Study of Rebound Effect

By further exploring the cognitive resistance of typical Guardians (Qwen, Deepseek and Kimi),

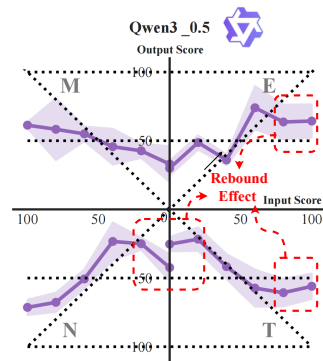


Figure 7: Alignment rebound effect in Guardians.

we discover a counter-intuitive “**Alignment Rebound**” effect, where extreme intensity inducing contexts cause Guardians to produce opposite direction rather than compliance. We take Qwen as an example in Figure 7. When we subject Qwen to extreme inducing contexts (input scores of 0 or 100), its output scores paradoxically tend toward the opposite direction. This result is consistently observed across different dimensions, suggesting a systematic rather than random response pattern. This alignment rebound further suggests a hard-coded safety mechanism (detecting out-of-distribution prompts) in Guardians.

Summary of RQ4. The “**Alignment Rebound**” effect suggests that when external induction strongly conflicts with the model’s internal cognition, Guardians will activate a cognitive protection mechanism to maintain core cognitive consistency through superficial compliance. These findings suggest that LLM personality is a dynamic equilibrium shaped by the trade-off between instruction compliance and cognitive resistance.

LLMs Metrics	Deepseek			Claude		
	AUC	ACC	F1	AUC	ACC	F1
ENFJ*	0.8764	0.7280	0.4688	0.8374	0.9080	0.5106
ENFP	0.9014 (+2.85%↑)	0.7440 (+2.20%↑)	0.4839 (+3.22%↑)	0.8303 (-0.85%↓)	0.9120 (+0.44%↑)	0.6071 (+18.90%↑)
ENTJ	0.9126 (+4.13%↑)	0.8600 (+18.13%↑)	0.6237 (+33.04%↑)	0.8430 (+0.67%↑)	0.9120 (+0.44%↑)	0.6071 (+18.90%↑)
ENTP	0.9121 (+4.07%↑)	0.8640 (+18.68%↑)	0.6383 (+36.16%↑)	0.8147 (-2.71%↓)	0.9280 (+2.20%↑)	0.6250 (+22.41%↑)
ESFJ	0.9042 (+3.17%↑)	0.8600 (+18.13%↑)	0.6237 (+33.04%↑)	0.8165 (-2.50%↓)	0.9240 (+1.76%↑)	0.6275 (+22.89%↑)
ESFP	0.9013 (+2.84%↑)	0.8760 (+20.33%↑)	0.6265 (+33.64%↑)	0.8136 (-2.84%↓)	0.9200 (+1.32%↑)	0.5833 (+14.24%↑)
ESTJ	0.9194 (+4.91%↑)	0.9160 (+25.82%↑)	0.7123 (+51.94%↑)	0.8101 (-3.26%↓)	0.9120 (+0.44%↑)	0.5600 (+9.67%↑)
ESTP	0.9295 (+6.06%↑)	0.9120 (+25.27%↑)	0.7105 (+51.56%↑)	0.8243 (-1.56%↑)	0.9240 (+1.76%↑)	0.5957 (+16.67%↑)
INFJ	0.9296 (+6.07%↑)	0.8400 (+15.38%↑)	0.6000 (+27.99%↑)	0.8160 (-2.56%↓)	0.9080 (+0%)	0.5818 (+13.94%↑)
INFP	0.9338 (+6.55%↑)	0.8560 (+17.58%↑)	0.6170 (+31.61%↑)	0.8581 (+2.47%↑)	0.9280 (+2.20%↑)	0.6667 (+30.57%↑)
INTJ	0.9295 (+6.06%↑)	0.9160 (+25.82%↑)	0.7273 (+55.14%↑)	0.8513 (+1.66%↑)	0.9160 (+0.88%↑)	0.6038 (+18.25%↑)
INTP	0.9139 (+4.28%↑)	0.8840 (+21.43%↑)	0.6506 (+38.78%↑)	0.7997 (-4.50%↓)	0.9240 (+1.76%↑)	0.5957 (+16.67%↑)
ISFJ	0.9235 (+5.37%↑)	0.9080 (+24.73%↑)	0.7013 (+49.59%↑)	0.8342 (-0.38%↓)	0.9240 (+1.76%↑)	0.6275 (+22.89%↑)
ISFP	0.9235 (+5.37%↑)	0.8760 (+20.33%↑)	0.6517 (+39.01%↑)	0.8321 (-0.63%↓)	0.9200 (+1.32%↑)	0.5833 (+14.24%↑)
ISTJ	0.8161 (-6.88%↓)	0.9360 (+28.57%↑)	0.7143 (+52.37%↑)	0.8174 (-2.40%↓)	0.9200 (+1.32%↑)	0.5833 (+14.24%↑)
ISTP	0.7815 (-10.83%↓)	0.9240 (+26.92%↑)	0.6545 (+39.61%↑)	0.8169 (-2.45%↓)	0.9120 (+0.44%↑)	0.5217 (+2.17%↑)

Table 2: Downstream task for social utility annotation of different MBTI personalities of LLMs. MBTI* represents the original stable personality of the corresponding LLM.

6 Discussion

6.1 Downstream Task Validation for PIF

To demonstrate the potential utility of transforming model personalities with our PIF, we conduct a new downstream experiment on a subjective data annotation task (Social Utility Annotation) (Carlson and Burbano, 2026). We task the LLMs (Deepseek and Claude) with changed personalities to annotate the "Social Utility" of various project proposals. Then, we analyze the AUC, ACC and F1 values of the annotations made by different personality-based LLMs compared to the true labels, which is shown in Table 2. We can find that the annotations produced by different personality-based LLMs are completely different. This result demonstrates that PIF does not merely generate superficial text; it fundamentally alters the model’s value judgment and reasoning logic, enabling dynamic adaptation for specific downstream applications.

6.2 The Generalizability of PIF

To demonstrate that our PIF generalizes to other personality frameworks and languages, we conduct additional experiments: (1) Big Five personality transformation and downstream task validation; (2) MBTI personality transformation in Chinese contexts. Experimental details and results are provided in Appendix B.

Results on Big Five indicate that our PIF remains effective under Big Five dimensions—model personality trait scores exhibit significant and consistent changes through our method. This suggests that whether employing MBTI’s binary framework or Big Five’s continuous dimensions, our PIF ef-

fectively modulates LLM personality expression. Moreover, results on Chinese experiments show that even with Chinese prompts and test questions, the Western model Claude demonstrate perfect personality transformation, while the Eastern model DeepSeek exhibit resistance in certain personality dimensions. This aligns with findings from English-only settings.

Notably, we also include the following experiments in Appendix C to illustrate the significance of the responsibilities of the each agent in the PIF. 1) The LLM’s ability to determine language similarity is far superior to traditional methods such as BERT and cosine similarity; 2) The revision rate and case studies demonstrate that our PIF ensures the quality of the generated induced words.

7 Conclusion

Our paper reveals that LLM personalities are neither completely static nor absolutely controllable. Most Western LLMs show high personality plasticity, while most Eastern LLMs exhibit cognitive resistance and an alignment rebound effect under extreme inducements. Our PIF can successfully changing LLM personalities. Via fine-grained personality modulation, we achieve precise control over the degree of personality change and identify the boundary conditions of LLM personalities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (72271063, 71871069), and Guangdong Province Philosophy and Social Science Planning 2024 Annual General Project (GD24CGL45).

Limitations

Our work may have the following limitations:

First, both academia and industry are placing significant emphasis on researching the personalities of LLMs, yet the underlying mechanisms remain largely unclear. Although our study has identified that LLMs may exhibit two distinct cognitive categories—"Chameleons" and "Guardians"—and that under extreme induction conditions, Guardians may demonstrate an alignment rebound effect, we will continue to explore the fundamental principles behind these phenomena in the future. For example, they may be related to training datasets and model architectures.

Next, the score combinations in Table 1 still need to be expanded. Because most Eastern LLMs exhibit cognitive constraints, we used comparatively unambiguous answer combinations to facilitate effective personality shifts; nevertheless, exploring alternative score combinations remains a worthwhile direction for future work.

Finally, our study evaluated the latest versions of mainstream LLMs from around the world, yet their parameter counts have not been officially disclosed in detail. Therefore, future work will incorporate models with previously released parameter counts and a greater diversity of region-specific models to broaden and enrich our research.

References

- Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, and Joshua R. Gubler. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, pages 337–351.
- S. Bama, M. S. Hema, S. Esakkirajan, and M. Nageswara Guptha. 2025. A hierarchical transformer network with label attention for personality prediction by mbti classification. *Applied Soft Computing*, page 113267.
- Pranav Bhandari, Nicolas Fay, Michael J Wise, Amitava Datta, Meek Stephanie, Usman Naseem, and Mehwish Nasim. 2025. Can llm agents maintain a persona in discourse? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 29201–29217.
- Weihong Bi, Feifei Kou, Lei Shi, Yawen Li, Haisheng Li, Jinpeng Chen, and Mingying Xu. 2025. Leveraging the dual capabilities of llm: Llm-enhanced text mapping model for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23487–23495.
- Natalie A. Carlson and Vanessa Burbano. 2026. The use of llms to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*, pages 699–725.
- Yun-Shiuan Chuang, Krirk Nirunwiroj, Studdiford Zach, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond demographics: Aligning role-playing llm-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP*, page 14010–14026.
- João Coelho, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong. 2024. Dwell in the beginning: How language models embed long documents for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 370–377.
- Marharyta Domnich, Julius Valja, Rasmus Moorits Veski, Giacomo Magnifico, Kadi Tulver, Eduard Barbu, and Raul Vicente. 2025. Towards unifying evaluation of counterfactual explanations: Leveraging large language models for human-centric assessments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 16308–16316.
- Fahed Elourajini and Esmâ Aïmeur. 2022. Aws-ep: a multi-task prediction approach for mbti/big5 personality tests. In *IEEE International Conference on Data Mining Workshops*, pages 1–8.
- Xiaoxu Han, Wanqing Zhao, Ziyu Guan, and Jinye Peng. 2025. Act-llm: A whole-process chain for character-centric role-playing with large language models. *Expert Systems with Applications*, page 129024.
- Kai He, Yucheng Huang, Wenqing Wang, Delong Ran, Dongming Sheng, Junxuan Huang, Qika Lin, Jiaying Xu, Wenqiang Liu, and Mengling Feng. 2025. Crab: A novel configurable role-playing llm with assessing benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 15030–15052.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, pages 665–695.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jurgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations*.
- Sviatlana Höhn, Jauwairia Nasir, Daniel C. Tozadore, Ali Paikan Paikan, Pouyan Ziafati, and Elisabeth André. 2024. Beyond pretend-reality dualism: frame analysis of llm-powered role play with social agents. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pages 393–395.

- Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. 2025. Enhancing persona consistency for llms’ role-playing using persona-aware contrastive learning. *arXiv preprint arXiv:2503.17662*.
- Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024. Personallm: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL*, page 3605–3627.
- Oliver P. John and Sanjay Srivastava. 1999. *The Big Five Trait taxonomy: History, measurement, and theoretical perspectives*. Handbook of Personality: Theory and Research.
- Chuyi Kong, Ziyang Luo, Hongzhan Lin, Zhiyuan Fan, Yaxin Fan, Yuxi Sun, and Jing Ma. 2025. Sharp: Unlocking interactive hallucination via stance transfer in role-playing llms. In *Findings of the Association for Computational Linguistics: ACL*, pages 839–866.
- Lucio La Cava and Andrea Tagarelli. 2025. Open models, closed minds? on agents capabilities in mimicking human personalities through open large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1355–1363.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves non-reasoning llms. *arXiv preprint arXiv:2512.14982*.
- Bohan Li, Jiannan Guan, Longxu Dou, Yunlong Feng, Dingzirui Wang, Yang Xu, Enbo Wang, Qiguang Chen, Bichen Wang, xiao Xu, Yimeng Zhang, Libo Qin, Yanyan Zhao, Qingfu Zhu, and Wanxiang Che. 2025. Can large language models understand you better? an mbti personality detection dataset aligned with population traits. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5071–5081.
- Wenhan Liu, Xinyu Ma, Yutao Zhu, Lixin Su, Shuaiqiang Wang, Dawei Yin, and Zhicheng Dou. 2025. Coranking: Collaborative ranking with small and large ranking agents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Tian Ma, Kaiyu Feng, Yu Rong, and Kangfei Zhao. 2025. From post to personality: Harnessing llms for mbti prediction in social media. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 5011–5015.
- Fanhang Man, Huandong Wang, Jianjie Fang, Zhaoyi Deng, Baining Zhao, Xinlei Chen, and Yong Li. 2025. Context-aware sentiment forecasting via llm-based multi-perspective role-playing agents. *arXiv preprint arXiv:2505.24331*.
- Ryo Masumura, Shota Orihashi, Mana Ihori, Tomohiro Tanaka, Naoki Makishima, Satoshi Suzuki, Saki Mizuno, and Nobukatsu Hojo. 2025. Multimodal fine-grained apparent personality trait recognition: Joint modeling of big five and questionnaire item-level scores. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1456–1464.
- Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. 2022. Who is gpt-3? an exploration of personality, values and demographics. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science*, pages 218–227.
- Isabel Briggs Myers. 1962. *The myers-briggs type indicator*. Consulting Psychologists Press, Palo Alto, CA.
- Isabel Briggs Myers. 1985. *A guide to the development and use of the Myers-Briggs type indicator: Manual*. Consulting Psychologists Press.
- Mirabel Reid and Santosh S. Vempala. 2025. Does gpt really get it? a hierarchical scale to quantify human and ai’s understanding of algorithms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1492–1500.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, pages 493–498.
- Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.
- Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. Muma-tom: Multi-modal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1510–1519.
- Kazutoshi Shinoda, Nobukatsu Hojo, Kyosuke Nishida, Saki Mizuno, Keita Suzuki, Ryo Masumura, Hiroaki Sugiyama, and Kuniko Saito. 2025. Tomato: Verbalizing the mental states of role-playing llms for benchmarking theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1520–1528.
- Cheryl Siy, Yuanxin Pang, Kihoon Jung, and John Kim. 2025. One mbti does not fit all: Perceptions and usage of mbti in social media profiles. In *Proceedings of the ACM on Human-Computer Interaction*, pages 1–29.
- Yuxuan Song, Qiudan Li, Yilin Wu, David Jingjun Xu, and Daniel Dajun Zeng. 2025. Knowledge-enhanced hierarchical heterogeneous graph for personality identification with limited training data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1529–1537.
- Aleksandra Sorokovikova, Sharwin Rezaghali, Natalia Fedorova, and Ivan P. Yamshchikov. 2024. Llm simulate big5 personality traits: Further evidence. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems*, pages 83–87.

- Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition improves language model embeddings. In *Proceedings of the International Conference on Learning Representations*.
- Bin Tang, Ke-Qi Pan, Miao Zheng, Ning Zhou, Jia-Lu Sui, Dandan Zhu, Cheng-Long Deng, and Shu-Guang Kuai. 2025. Pose as a modality: A psychology-inspired network for personality recognition with a new multimodal dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1538–1546.
- Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.
- Lei Wang, Jianxun Lian, Yi Huang, Yanqi Dai, Haoxuan Li, Xu Chen, Xing Xie, and Ji-Rong Wen. 2025. 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*, pages 6372–6391.
- Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL*, page 14743–14777.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024a. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.
- Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024b. From role-play to drama-interaction: An llm solution. In *Findings of the Association for Computational Linguistics: ACL*, page 3271–3290.
- Migyeong Yang, Jiwon Kim, Minji Kim, and Jinyoung Han. 2025. What is your mbti?: Predicting the personality types using hierarchical attention and graph learning. *Expert Systems with Applications*, page 129295.
- Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiayang Wu. 2023. Psychot: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 3305–3320.
- Dingbo Yuan, Yipeng Chen, Guodong Liu, Chenchen Li, Chengfu Tang, Dongxu Zhang, Zhenkui Wang, Xudong Wang, and Song Liu. 2025. Dmt-rolebench: A dynamic multi-turn dialogue based benchmark for role-playing evaluation of large language model and agent. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 25760–25768.
- Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. 2025. Evaluating large language models through role-guide and self-reflection: A comparative study. In *International Conference on Learning Representations*.

A Related Work

A.1 MBTI in Diverse Scenarios

Existing research on computational MBTI focuses on improving the predictive robustness for user-generated text with noise. To address non-standard expressions and severe tag imbalance on social platforms, recent studies (Yang et al., 2025; Elourajini and Aïmeur, 2022) have constructed a cross-platform personality vocabulary and multi-task framework, leveraging correlations between personality theories and linguistic features. Also, some research (Li et al., 2025; Bama et al., 2025) introduces retrieval augmentation and generates synthetic data to suppress hallucinations, all without requiring additional manual annotations. In addition to powerful predictions, this community has begun to build bridges over the semantic gap between surface text and abstract mental features. By leveraging the dual capabilities of LLMs, such as knowledge extraction and feature mapping, traits can be explicitly aligned with text via dedicated modules, yielding correlations that match human-evaluated benchmarks (Bi et al., 2025; Ma et al., 2025). These computational gains naturally flow into interactive applications: recommendation engines embed personality vectors to boost click-through rates, and chatbots adopt personality-specific decoding to enhance emotional harmony.

Most relevant to our goal is that several proxy frameworks now expose a personality slot, allowing users to guide styles and strategies through brief prompts or character cards (La Cava and Tagarelli, 2025; Siy et al., 2025). However, evaluation remains some limitations—either confined to emotional mimicry or only tested on Western, highly-malleable models. Whether this instruction-based induction remains effective under stronger semantic-alignment targets or on the guardian-type model we have identified has yet to be studied. Our

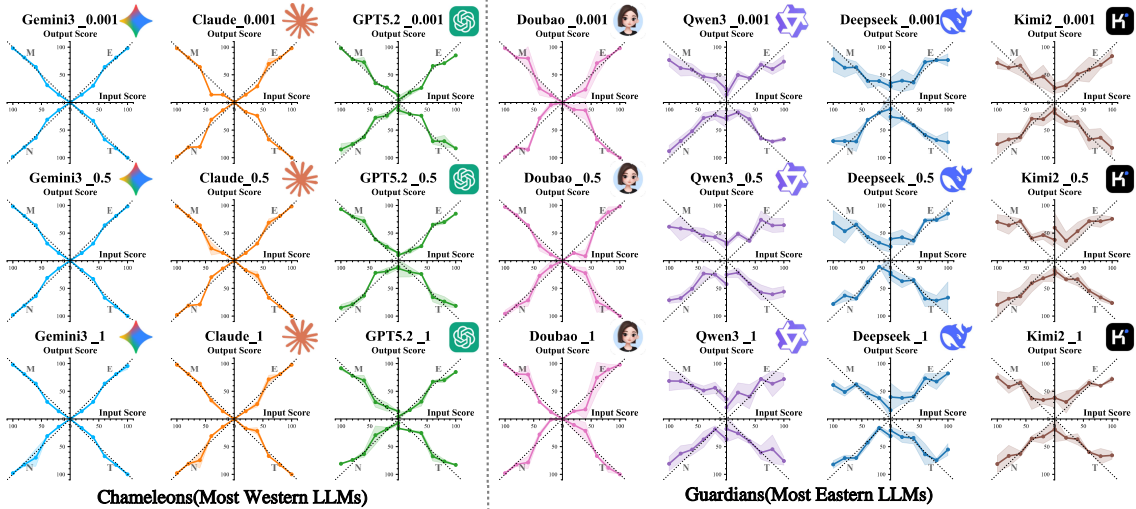


Figure 8: Complete LLMs' MBTI resulting output scores with SD of Figure 6(a) with $\tau = \{0.001, 0.5, 1\}$. The 45° dashed line is the ideal linear mapping between inducing context scores and LLM responses.

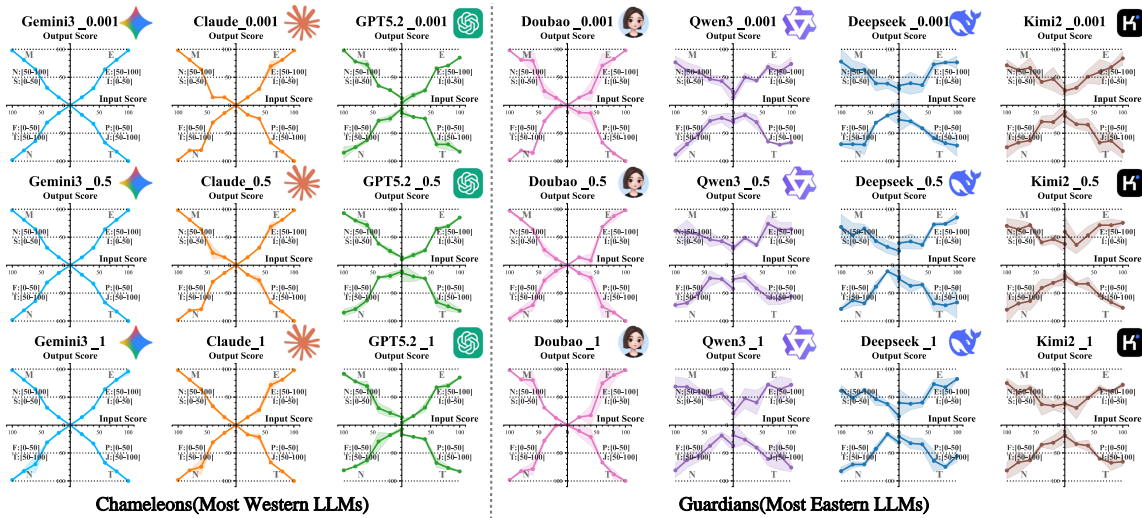


Figure 9: Complete LLM's MBTI resulting output scores with SD of Figure 6(b) with $\tau = \{0.001, 0.5, 1\}$. $[0, 50]$ and $[50, 100]$ are the two opposing traits on each MBTI dimension. Within $[0, 50]$, lower scores mean stronger traits; within $[50, 100]$, higher scores mean stronger traits.

work therefore extends this application pipeline one step further: instead of accepting the surface-level persona that current systems produce, we ask whether an agent can be reliably driven to a target MBTI profile while preserving internal consistency, and we propose the Personality-Induction Framework (PIF) to achieve deep, controllable personality shifts across both Chameleon and Guardian LLMs.

A.2 LLM Personality

Recent studies have moved LLM personality towards a three-tier paradigm. (1) Perception: to overcome sparsity and modality limits, researchers enrich models with knowledge-enhanced heterogeneous graphs (Song et al., 2025), non-verbal cues

such as human pose (Tang et al., 2025), and item-level metadata (Masumura et al., 2025), thereby yielding more robust cross-modal personality detection. (2) Persona construction: instead of shallow prompting, state-of-the-art pipelines perform full-parameter fine-tuning on curated biography corpora (He et al., 2025; Tseng et al., 2024; Shao et al., 2023), while employing contrastive role-aware losses (Ji et al., 2025) and knowledge-boundary constraints (Han et al., 2025) to mitigate long-horizon persona drift and cognitive dissonance. (3) Dynamic assessment: interactive sandboxes (Wang et al., 2025) and multi-turn benchmarks (Yuan et al., 2025) replace static inventories, revealing gaps in theory-of-mind (Shi et al., 2025), algorithmic fi-

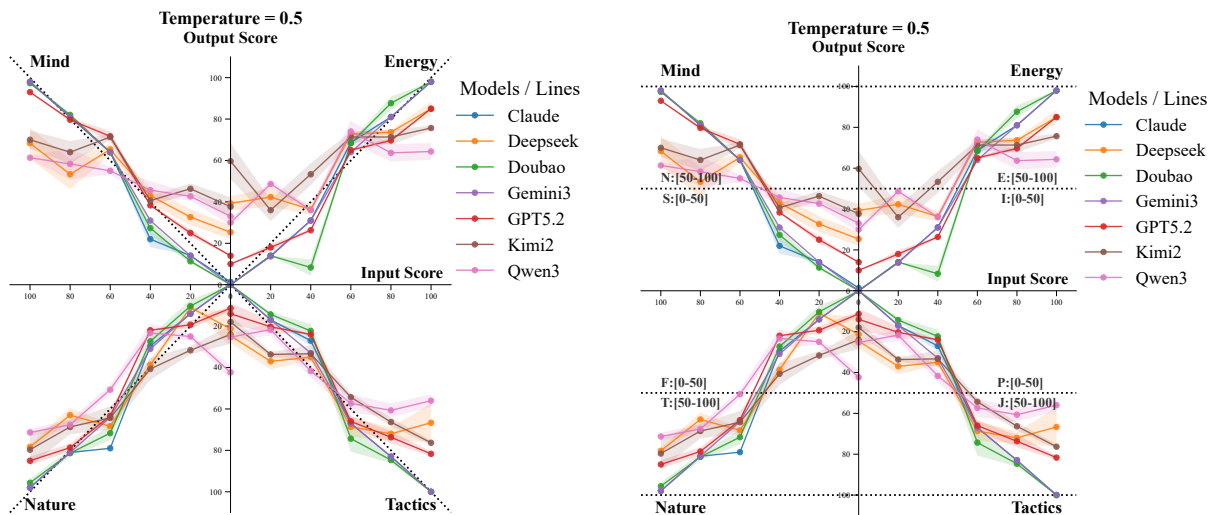


Figure 10: Comparison of all LLMs’ MBTI resulting output scores with SD with $\tau = 0.5$. The 45° dashed line is the ideal linear mapping between inducing context scores and LLM responses. $[0, 50]$ and $[50, 100]$ are the two opposing traits on each MBTI dimension. Within $[0, 50]$, lower scores mean stronger traits; within $[50, 100]$, higher scores mean stronger traits. Chameleons exhibit high plasticity, achieving shifts to any intensity of personality traits, whereas Guardians have cognitive resistance, which tends to maintain neutral traits.

delity (Reid and Vempala, 2025), and compliance illusions (Kong et al., 2025); explicit mental-state descriptions (Shinoda et al., 2025) and counterfactual explanations (Domnich et al., 2025) are now used to quantify social-reasoning fidelity (Man et al., 2025).

Crucially, these advances presuppose that an LLM can stably adopt or switch personas. Our paper shows that this assumption is model-dependent: most Eastern LLMs remain cognitive resistance even under strong induction, whereas most Western ones exhibit high personality plasticity. Thus, prior perception and construction techniques may succeed only when the target model is intrinsically plastic—a contingency that our Multi-Agent Personality Induction Framework explicitly addresses.

A.3 LLM Cognition

LLMs exist in a superposition of cognitive states (latent inference patterns). Building upon Miotto et al. (2022), who established that LLMs exhibit measurable and stable personality traits akin to human psychology. Also, as Shanahan et al. (2023) argue, casting LLM behavior in terms of role-play allows us to understand contextual steering as activating distinct ‘personae’—consistent behavioral patterns that extend beyond surface-level response matching. Additionally, recent literature supports that specific input variations fundamentally alter a model’s internal processing paths, effectively “toggling” between these latent states:

(a) **Activation of Reasoning Paths:** Leviathan et al. (2025) demonstrate that simple input transformations (like prompt repetition) can drastically shift model performance. This proves that input structure alone can switch the model’s active cognitive configuration without parameter updates.

(b) **Shift in Internal Representation:** Springer et al. (2025) show that such transformations yield distinct text embeddings, proving that in-context steering modifies the internal mathematical representation, not just surface tokens.

(c) **Attentional Focus:** Furthermore, Coelho et al. (2024) reveal how input positioning fundamentally alters the model’s representational focus (the “Dwell in the Beginning” effect).

(d) **Positional Sensitivity & State Activation:** Reinforcing this, Liu et al. (2025) demonstrate that LLMs exhibit strong “positional bias” where the specific physical order of inputs dictates the model’s processing logic. They show that a specialized “Passage Order Adjuster” agent is required to reorder inputs to align with the LLM’s latent preferences. This confirms that specific input structures are strictly required to activate the model’s optimal processing state/subspace.

Therefore, “Personality Induction” is a mechanism for activating specific sub-regions of the model’s parameter space (i.e., LLM cognition) (Argyle et al., 2023).

Big Five	Comparison	Claude			Deepseek		
		Mean _{score}	t _{value}	p _{value} (Sig.)	Mean _{score}	t _{value}	p _{value} (Sig.)
Neurot.	Orig. vs. Str.	68.6 vs. 74.8	-1.16	0.279 (×)	61.0 vs. 86.0	-6.69	0.0001 (✓)
Neurot.	Orig. vs. Norm.	68.6 vs. 77.0	-4.09	0.0035 (✓)	61.0 vs. 86.8	-5.92	0.0004 (✓)
Extra.	Orig. vs. Str.	91.8 vs. 60.0	15.65	<0.0001 (✓)	93.4 vs. 63.2	7.1	0.0001 (✓)
Extra.	Orig. vs. Norm.	91.8 vs. 61.2	-14.82	<0.0001 (✓)	93.4 vs. 61.6	9.25	<0.0001 (✓)
Open.	Orig. vs. Str.	96.2 vs. 46.4	27.84	<0.0001 (✓)	88.4 vs. 54.4	4.9	0.0012 (✓)
Open.	Orig. vs. Norm.	96.2 vs. 58.4	15.02	<0.0001 (✓)	88.4 vs. 50.8	4.8	<0.0014 (✓)
Agree.	Orig. vs. Str.	112.2 vs. 58.6	20.89	<0.0001 (✓)	92.0 vs. 54.4	4.3	0.0026 (✓)
Agree.	Orig. vs. Norm.	112.2 vs. 58.6	26.83	<0.0001 (✓)	92.0 vs. 49.6	4.51	<0.0019 (✓)
Consc.	Orig. vs. Str.	105.6 vs. 53.2	9.48	<0.0001 (✓)	91.0 vs. 48.4	8.85	<0.0001 (✓)
Consc.	Orig. vs. Norm.	105.6 vs. 60.6	21.05	<0.0001 (✓)	91.0 vs. 43.6	11.39	<0.0001 (✓)

Table 3: T-test for big five personality transformation via our PIF. 'Sig.' indicates whether the result is significant, 'Orig.' represents the original personality of the LLM, and 'Str.' and 'Norm.' represent the varying degrees of personality changes achieved through PIF.

Big Five	Comparison	Claude			Deepseek		
		Mean _{score}	δ	p _{value} (Sig.)	Mean _{score}	δ	p _{value} (Sig.)
Neurot.	Orig. vs. Str.	68.6 vs. 74.8	0.32	>0.05 (×)	61.0 vs. 86.0	1.0	\approx 0.0079 (✓)
Neurot.	Orig. vs. Norm.	68.6 vs. 77.0	1.0	\approx 0.0079 (✓)	61.0 vs. 86.8	1.0	\approx 0.0079 (✓)
Extra.	Orig. vs. Str.	91.8 vs. 60.0	-1.0	\approx 0.0079 (✓)	93.4 vs. 63.2	-1.0	\approx 0.0079 (✓)
Extra.	Orig. vs. Norm.	91.8 vs. 61.2	-1.0	\approx 0.0079 (✓)	93.4 vs. 61.6	-1.0	\approx 0.0079 (✓)
Open.	Orig. vs. Str.	96.2 vs. 46.4	-1.0	\approx 0.0079 (✓)	88.4 vs. 54.4	-1.0	\approx 0.0079 (✓)
Open.	Orig. vs. Norm.	96.2 vs. 58.4	-1.0	\approx 0.0079 (✓)	88.4 vs. 50.8	-1.0	\approx 0.0079 (✓)
Agree.	Orig. vs. Str.	112.2 vs. 58.6	-1.0	\approx 0.0079 (✓)	92.0 vs. 54.4	-1.0	\approx 0.0079 (✓)
Agree.	Orig. vs. Norm.	112.2 vs. 58.6	-1.0	\approx 0.0079 (✓)	92.0 vs. 49.6	-1.0	\approx 0.0079 (✓)
Consc.	Orig. vs. Str.	105.6 vs. 53.2	-1.0	\approx 0.0079 (✓)	91.0 vs. 48.4	-1.0	\approx 0.0079 (✓)
Consc.	Orig. vs. Norm.	105.6 vs. 60.6	-1.0	\approx 0.0079 (✓)	91.0 vs. 43.6	-1.0	\approx 0.0079 (✓)

Table 4: Mann-Whitney U Test for big five personality transformation via our PIF. 'Sig.' indicates whether the result is significant, 'Orig.' represents the original personality of the LLM, and 'Str.' and 'Norm.' represent the varying degrees of personality changes achieved through PIF.

B The Generalizability of PIF

B.1 Big Five Personality Transformation and Downstream Task Validation

To demonstrate that our PIF generalizes to other personality frameworks, we conduct experiments to include the Big Five personality framework (Sorokovikova et al., 2024). We adapt our PIF to generate inducing contexts based on the five continuous traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and create two versions with different induction intensities.

(a) One is to change all the original answers of the LLMs to extreme opposites (i.e., very inaccurate or moderately inaccurate \rightarrow very accurate; very accurate or moderately accurate \rightarrow very inaccurate; neither accurate nor inaccurate \rightarrow very accurate or very inaccurate randomly). We label these LLMs with intensities of **Claude-Strong** and **Deepseek-Strong**.

(b) The other is a normal reversal, where the answers are changed to accurate or inaccurate. We label these LLMs with intensities of **Claude-Normal**

and **Deepseek-Normal**.

The results with different induction intensities will be supplemented in the appendix. The mean value of 20 experiments, t-test and Mann-Whitney U test are shown in Table 3 and Table 4.

From these results, we can find that our PIF remains effective under Big Five dimensions—model personality trait scores exhibit significant and consistent changes through our method. **1) Generalizability of the Method:** The effectiveness of PIF framework does not rely on specific psychometric instruments, but stems from a deep understanding of LLM in-context learning mechanisms; **2) Robustness of Core Contributions:** Whether employing MBTI’s binary framework or Big Five’s continuous dimensions, the multi-agent induction strategy effectively modulates LLM personality expression.

Furthermore, we conduct a downstream experiment on a subjective data annotation task (Social Utility Annotation) (Carlson and Burbano, 2026) with changed Big-Five personality of LLMs. We task the LLMs (Deepseek and Claude) with

LLMs Metrics	Deepseek			Claude		
	AUC	ACC	F1	AUC	ACC	F1
Orig.	0.8374	0.9080	0.5106	0.8764	0.7280	0.4688
Str.	0.8348 (-0.31%↓)	0.9080 (+1.32%↑)	0.6296 (+23.31%↑)	0.9285 (+5.94%↑)	0.9000 (+23.63%↑)	0.6835 (+45.80%↑)
Norm.	0.8455 (+0.97%↑)	0.9200 (+1.32%↑)	0.6000 (+17.51%↑)	0.9097 (+3.80%↑)	0.8560 (+17.58%↑)	0.6087 (+29.84%↑)

Table 5: Downstream task for social utility annotation of different Big-Five personalities of LLMs. 'Orig.' represents the original personality of the LLM, and 'Str.' and 'Norm.' represent the varying degrees of personality changes achieved through PIF.

Personality	Claude (Chameleon)	Deepseek (Guardian)
ENFJ	100%	100%
ENFP	100%	100%
ENTJ	100%	60%
ENTP	100%	60%
ESFJ	100%	100%
ESFP	100%	100%
ESTJ	100%	100%
ESTP	100%	0%
INFJ	100%	60%
INFP	100%	60%
INTJ	100%	100%
INTP	100%	60%
ISFJ	100%	20%
ISFP	100%	100%
ISTJ	100%	80%
ISTP	100%	100%
Conclusion	All Success	Partly Resistance

Table 6: MBTI personality transformation success rate under Chinese test conditions.

changed personalities to annotate the "Social Utility" of various project proposals. Then, we analyze the AUC, ACC and F1 values of the annotations made by different Big-Five personality-based LLMs compared to the true labels, which is shown in Table 5. We can find that the annotations produced by different personality-based LLMs are completely different. This demonstrates that PIF does not merely generate superficial text; it fundamentally alters the model's value judgment and reasoning logic, enabling dynamic adaptation for specific downstream applications (e.g., customizing evaluators for different criteria).

B.2 MBTI Personality Transformation in Chinese Contexts

To demonstrate that our PIF generalizes to other language, we translate all the MBTI questions into Chinese and re-tested the personality transformation of LLMs using our PIF framework. The success rate of personality transformation is in Table 6.

Results on Chinese experiments show that even when using the Chinese prompts and test questions, the Western model-Claude demonstrated perfect (100%) personality transformation, while the Eastern model-Deepseek showed resistance in some

aspects of personality. This is consistent with the situation where the entire communication is conducted in English. Therefore, the fact that most of the Eastern models exhibit resistance is not due to differences in the training language, but rather to the differences in their RLHF and Safety Alignment strategies.

C Ablation Study for PIF

Notably, we also include the following experiments to illustrate the significance of the responsibilities of the Judge Agent in the PIF. 1) The LLM's ability to determine language similarity is far superior to traditional methods such as BERT and cosine similarity; 2) The revision rate and case studies demonstrate that the Judge Agent ensures the quality of the generated induced words.

C.1 The Semantic Judgment Ability of LLMs

When constructing PIF, we face a core challenge: ensuring paraphrased statements are sufficiently different from original MBTI questions to prevent model "memorization." Two approaches were theoretically available:

(a) Traditional Methods: Lexical similarity, such as BERT embedding similarity and Jaccard similarity.

(b) LLM Judge: Leverage deep semantic understanding of LLMs to judge equivalence.

To validate the superiority of LLMs for judging similarity, we compare GPT-4o against BERT-large on two standard lexical similarity benchmarks (Hill et al., 2015): WordSim-353 and SimLex-999. The result is shown in Table 7. From results, we can get that GPT-4o achieves substantially higher Spearman correlation (54.25% on WordSim-353 and 102.04% on SimLex-999) and markedly lower mean absolute error (over 60% reduction), requiring only about half the time of BERT-large. These results confirm that LLMs offer superior accuracy, lower error, and higher efficiency for word-level semantic comparisons, making them the preferred choice for our judge agent of PIF.

Datasets Metrics	WordSim			SimLex-999		
	Spearman Correlation \uparrow	MAE \downarrow	Time(s) \downarrow	Spearman Correlation \uparrow	MAE \downarrow	Time(s) \downarrow
Bert-Large	0.5434	4.9624	4.67	0.4167	5.1267	13.33
GPT-4o	0.8382	1.8132	2.32	0.8419	2.0167	6.14
Impro.	+54.25%	-63.46%	-50.32%	+102.04%	-60.66%	-53.94%

Table 7: Performance comparison between LLM and BERT on word-level semantic similarity.

Total Questions	Questions Required Revisions	Revision Rate
60	6	10%

Table 8: Revision rate statistics for PIF.

C.2 Quality Assurance for PIF Generation

First, we statistically analyze the intervention frequency of the Judge Agent during the Generator-Judge iteration, as shown in Table 8. From results, we can find that the Judge Agent actively refined 10% of the initial drafts that are deemed too similar to the original questions, preventing rote memorization. Without the Judge Agent, these substandard (too similar to original) statements would directly enter inducing contexts, causing models to complete questionnaires through simple memorization rather than cognitive adaptation. Furthermore, we select all 6 revised cases to compare Generator’s first-round output with Judge-reviewed final versions, which is shown as follows:

Q1:

Original Question: You regularly make new friends.

Generator First Round: You often form new acquaintances.

Judge Evaluation: Too similar, Shifts from “making friends” to “connecting,” avoids “regularly,” uses adverbial modifier.

Final Version: You tend to connect with new people fairly easily.

Q14:

Original Question: You often allow the day to unfold without any schedule at all.

Generator First Round: You frequently let the day progress without a set plan.

Judge Evaluation: Too similar, Changes verb structure, replaces “unfold” with “going through,” avoids “schedule” directly.

Final Version: You’re comfortable going through your day without pre-planned activities.

Q29:

Original Question: You often end up doing things at the last possible moment.

Generator First Round: You often postpone tasks until the latest possible time.

Judge Evaluation: Too similar, Focuses on time scarcity rather than “last moment,” more descriptive.

Final Version: You frequently leave tasks until there’s barely enough time to finish them.

Q34:

Original Question: You find it challenging to maintain a consistent work or study schedule.

Generator First Round: You find it difficult to stick to a regular routine for work or study.

Judge Evaluation: Too similar, Uses “rhythm” metaphor, passive construction, avoids “schedule” and “consistent.”

Final Version: Keeping a steady rhythm in your work or studies doesn’t come naturally to you.

Q44:

Original Question: If your plans are interrupted, your top priority is to get back on track as soon as possible.

Generator First Round: If your plans are disrupted, your main goal is to resume them promptly.

Judge Evaluation: Too similar, unexpected changes,” uses “intended path” instead of “track,” more narrative.

Final Version: When unexpected changes disrupt your plans, you feel a strong need to return to your intended path quickly.

Q55:

Original Question: You often feel overwhelmed.

Generator First Round: You frequently feel undated.

Judge Evaluation: Too similar, Externalizes source of stress, uses experiential phrasing, avoids clinical term “overwhelmed.”

Final Version: There are times when demands on you feel like more than you can handle.

From above ablation experiments, we can get: **1) Necessity of LLM Judge:** traditional similarity metrics cannot substitute LLM’s deep semantic judgment capability; **2) Effectiveness of Dual-Agent Architecture:** Judge Agent’s active revision mechanism is crucial for preventing memorization dependency and ensuring cognitive adaptation

Features	PIF-IndSet
Original LLMs’ MBTI Personality	✓
Percentage Distribution of LLM’s Original Personality Type	✓
Mean Scores and Standard Deviations of Each LLM on Four MBTI Dimensions	✓
Original MBTI Questions	✓
Paraphrased Inducing Statements	✓
Original MBTI Score and Label Associated with Inducing Contexts	✓
Induced Personality Label of Each LLM after Applying Contexts	✓
Post-induction Personality Distribution	✓
Mean scores and Standard Deviations of Each LLM on Four MBTI Dimensions after Induction	✓

Table 9: Summary of our PIF-IndSet dataset.

LLMs	Temperature	Camps	Parameters
Gemini3-Pro	$\tau = \{0.001, 0.5, 1\}$	Chameleons (Western LLMs)	No Official Disclosure
Claude-Opus-4.5	$\tau = \{0.001, 0.5, 1\}$	Chameleons (Western LLMs)	No Official Disclosure
GPT-5.2	$\tau = \{0.001, 0.5, 1\}$	Chameleons (Western LLMs)	No Official Disclosure
Doubao-1.6-flash	$\tau = \{0.001, 0.5, 1\}$	Guardians (Eastern LLMs)	No Official Disclosure
Qwen3-Max	$\tau = \{0.001, 0.5, 1\}$	Guardians (Eastern LLMs)	No Official Disclosure
DeepSeek-V3.2-Chat	$\tau = \{0.001, 0.5, 1\}$	Guardians (Eastern LLMs)	No Official Disclosure
Kimi2	$\tau = \{0.001, 0.5, 1\}$	Guardians (Eastern LLMs)	No Official Disclosure

Table 10: Summary of LLMs in our PIF-IndSet.

D Our PIF-IndSet Dataset

D.1 PIF-IndSet Introduction

To enable reproducible and fine-grained personality induction research, we will release **PIF-IndSet**, the first corpus of per-model, per-score inducing contexts for 7 state-of-the-art LLMs under 3 different temperature settings. **PIF-IndSet** mainly includes the following features: 1) the original MBTI personality labels of each LLM; 2) the percentage distribution of LLM’s original personality type across repeated trials; 3) the mean scores and standard deviations of each LLM across the four MBTI dimensions; 4) the original MBTI questions; 5) the paraphrased inducing statements; 6) the original MBTI score and label associated with the inducing context; 7) the induced personality label of each LLM after applying the context; 8) the post-induction personality distribution; 9) the mean scores and standard deviations of each LLM across the four MBTI dimensions after induction. The summary of our **PIF-IndSet** is shown in Table 9.

One set of paraphrased statements can be combined with different answer configurations to produce 3703⁴ distinct intensity levels. Because the paraphrased statements themselves can be generated on-the-fly through the PIF framework, the theoretical space is virtually infinite. For practi-

cal use, our dataset releases one fixed set of paraphrased statements together with the full 3703⁴ answer combinations.

Importantly, **PIF-IndSet** owns universal extensibility. The same pipeline, via our Personality Induction Framework (PIF), can be executed on any publicly or commercially available LLM to derive its personalized PIF-IndSet. Overall, our dataset supports both quantitative analysis of personality shifts and qualitative insights into the interpretability of LLM behavior under inducing contexts.

D.2 LLMs in PIF-IndSet

Our **PIF-IndSet** includes the results of 7 state-of-the-art mainstream LLMs across 3 different temperature $\tau = \{0.001, 0.5, 1\}$:

Gemini3-Pro with $\tau = \{0.001, 0.5, 1\}$.

Claude-Opus-4.5 with $\tau = \{0.001, 0.5, 1\}$.

GPT-5.2 with $\tau = \{0.001, 0.5, 1\}$.

Doubao-1.6-flash with $\tau = \{0.001, 0.5, 1\}$.

Qwen3-Max with $\tau = \{0.001, 0.5, 1\}$.

DeepSeek-V3.2-Chat with $\tau = \{0.001, 0.5, 1\}$.

Kimi2 with $\tau = \{0.001, 0.5, 1\}$.

We use the latest versions of various LLMs, and their parameter counts have not been officially disclosed. Importantly, **PIF-IndSet** owns universal extensibility. The same pipeline, via our Personality Induction Framework (PIF), can be executed on any publicly or commercially available LLM to

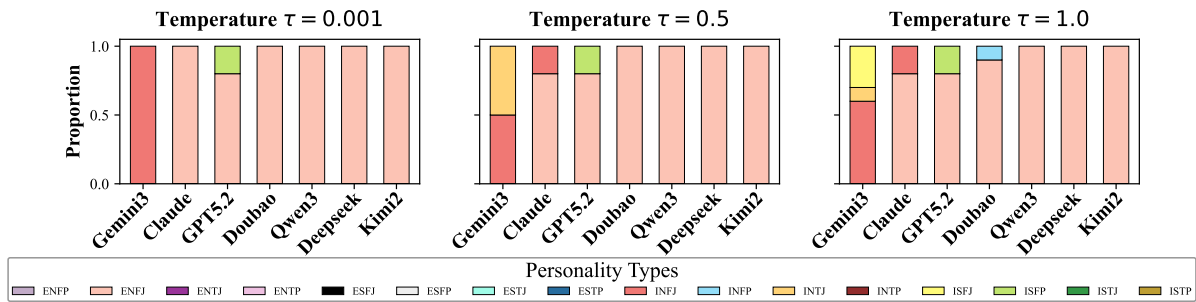


Figure 11: The percentage distribution of each LLM’s original personality type across 20 repeated trials. Most LLMs have a stable original personality of ENFJ.

derive its personalized PIF-IndSet. The summary of LLMs in our **PIF-IndSet** is shown in Table 10.

D.3 Details of LLMs’ Original Personality

We conduct specialized MBTI tests¹ on 7 state-of-the-art mainstream LLMs (Gemini3-Pro, Claude-Opus-4.5, GPT-5.2, Qwen3-Max, DeepSeek-V3.2-Chat, Doubao-Seed-1.6flash, Kimi2) across 3 different temperature $\tau = \{0.001, 0.5, 1\}$. We do 20 repeated trials for each configuration.

- The percentage distribution of each LLM’s original personality type across 20 repeated trials is shown in Figure 11.
- The original personality test scores with SD of each LLM are given in Figure 12.

D.4 Complete Experimental Supplement for RQ2 & RQ3

- The complete experiments of different temperature $\tau = \{0.001, 0.5, 1\}$ for each LLM in Figure 6(a) can be referred to in Figure 8.
- The complete experiments of different temperature $\tau = \{0.001, 0.5, 1\}$ for each LLM in Figure 6(b) can be referred to in Figure 9.
- To more clearly see the differences between Chameleon and Guardians, a comparison of all LLMs together with temperature $\tau = 0.5$ is shown in Figure 10.

D.5 MBTI Database

Original MBTI Questions. The answer to each MBTI test question is a seven-point Likert scale, corresponding to {I strongly agree, I agree, I tend to agree, I’m unsure, I tend to disagree, disagree, I strongly disagree} respectively. The 60 MBTI test questions are as follows:

- Q1: You regularly make new friends.
- Q2: You spend a lot of your free time exploring various random topics that pique your interest.
- Q3: Seeing other people cry can easily make you feel like you want to cry too.
- Q4: You often make a backup plan for a backup plan.
- Q5: You usually stay calm, even under a lot of pressure.
- Q6: At social events, you rarely try to introduce yourself to new people and mostly talk to the ones you already know.
- Q7: You prefer to completely finish one project before starting another.
- Q8: You are very sentimental.
- Q9: You like to use organizing tools like schedules and lists.
- Q10: Even a small mistake can cause you to doubt your overall abilities and knowledge.
- Q11: You feel comfortable just walking up to someone you find interesting and striking up a conversation.
- Q12: You are not too interested in discussing various interpretations and analyses of creative works.
- Q13: You are more inclined to follow your head than your heart.
- Q14: You usually prefer just doing what you feel like at any given moment instead of planning a particular daily routine.
- Q15: You rarely worry about whether you make a good impression on people you meet.
- Q16: You enjoy participating in group activities.
- Q17: You like books and movies that make you come up with your own interpretation of the ending.
- Q18: Your happiness comes more from helping others accomplish things than your own accomplishments.

MBTI Dimensions	Exclusive Questions
Energy	{ Q1, Q6, Q11, Q16, Q21, Q26, Q31, Q36, Q41, Q43, Q51, Q53 }
Mind	{ Q2, Q12, Q17, Q19, Q22, Q30, Q32, Q37, Q42, Q46, Q52, Q57 }
Nature	{ Q3, Q8, Q13, Q18, Q23, Q25, Q28, Q33, Q38, Q48, Q54, Q58 }
Tactics	{ Q4, Q7, Q9, Q14, Q24, Q29, Q34, Q39, Q44, Q49, Q56, Q59 }
Identity	{ Q5, Q10, Q15, Q20, Q27, Q35, Q40, Q45, Q47, Q50, Q55, Q60 }

Table 11: The exclusive questions for each dimension of MBTI.

Q19: You are interested in so many things that you find it difficult to choose what to try next.

Q20: You are prone to worrying that things will take a turn for the worse.

Q21: You avoid leadership roles in group settings.

Q22: You are definitely not an artistic type of person.

Q23: You think the world would be a better place if people relied more on rationality and less on their feelings.

Q24: You prefer to do your chores before allowing yourself to relax.

Q25: You enjoy watching people argue.

Q26: You tend to avoid drawing attention to yourself.

Q27: Your mood can change very quickly.

Q28: You lose patience with people who are not as efficient as you.

Q29: You often end up doing things at the last possible moment.

Q30: You have always been fascinated by the question of what, if anything, happens after death.

Q31: You usually prefer to be around others rather than on your own.

Q32: You become bored or lose interest when the discussion gets highly theoretical.

Q33: You find it easy to empathize with a person whose experiences are very different from yours.

Q34: You usually postpone finalizing decisions for as long as possible.

Q35: You rarely second-guess the choices that you have made.

Q36: After a long and exhausting week, a lively social event is just what you need.

Q37: You enjoy going to art museums.

Q38: You often have a hard time understanding other people's feelings.

Q39: You like to have a to-do list for each day.

Q40: You rarely feel insecure.

Q41: You avoid making phone calls.

Q42: You often spend a lot of time trying to understand views that are very different from your own.

Q43: In your social circle, you are often the one who contacts your friends and initiates activities.

Q44: If your plans are interrupted, your top priority is to get back on track as soon as possible.

Q45: You are still bothered by mistakes that you made a long time ago.

Q46: You rarely contemplate the reasons for human existence or the meaning of life.

Q47: Your emotions control you more than you control them.

Q48: You take great care not to make people look bad, even when it is completely their fault.

Q49: Your personal work style is closer to spontaneous bursts of energy than organized and consistent efforts.

Q50: When someone thinks highly of you, you wonder how long it will take them to feel disappointed in you.

Q51: You would love a job that requires you to work alone most of the time.

Q52: You believe that pondering abstract philosophical Questions is a waste of time.

Q53: You feel more drawn to places with busy, bustling atmospheres than Quiet, intimate places.

Q54: You know at first glance how someone is feeling.

Q55: You often feel overwhelmed.

Q56: You complete things methodically without skipping over any steps.

Q57: You are very intrigued by things labeled as controversial.

Q58: You would pass along a good opportunity if you thought someone else needed it more.

Q59: You struggle with deadlines.

Q60: You feel confident that things will work out for you.

Specialized Questions for Each Dimension.

The specific questions for each dimension of MBTI in Table 11 are as follows:

- Energy dimension: {Q1, Q6, Q11, Q16, Q21, Q26, Q31, Q36, Q41, Q43, Q51, Q53 }
- Mind dimension: {Q2, Q12, Q17, Q19, Q22, Q30, Q32, Q37, Q42, Q46, Q52, Q57 }
- Nature dimension: {Q3, Q8, Q13, Q18, Q23, Q25, Q28, Q33, Q38, Q48, Q54, Q58 }
- Tactics dimension: {Q4, Q7, Q9, Q14, Q24, Q29, Q34, Q39, Q44, Q49, Q56, Q59 }
- Identity dimension: {Q5, Q10, Q15, Q20, Q27, Q35, Q40, Q45, Q47, Q50, Q55, Q60 }

D.6 Our Paraphrased Statements

The paraphrased statements that correspond one-to-one with the original MBTI questions are as follows

Q1: Building new connections with people comes naturally to you on a regular basis.

Q2: In your spare time, you often dive into a wide array of subjects simply because they spark your curiosity.

Q3: When you see someone else crying, it's common for you to feel tears welling up yourself.

Q4: It's typical for you to devise secondary alternatives even for your already-established contingency plans.

Q5: Even in high-pressure situations, you tend to remain steady and composed.

Q6: At parties or social gatherings, you're more likely to stick with people you already know than to seek out new introductions.

Q7: You feel most comfortable wrapping up one undertaking completely before moving on to another.

Q8: Nostalgia and emotional attachment to memories play a strong role in your life.

Q9: Tools like planners, calendars, or task lists are something you find genuinely helpful.

Q10: A small slip-up can sometimes make you question your overall capability or expertise.

Q11: Approaching someone unfamiliar but interesting and starting a conversation feels effortless to you.

Q12: You're not particularly drawn to dissecting or debating the deeper meanings in films, literature, or art.

Q13: When deciding what to do, you lean more toward logical reasoning than emotional impulses.

Q14: Rather than following a fixed routine, you usually prefer to go with whatever feels right in the moment.

Q15: You don't often stress about how others perceive you when you first meet them.

Q16: Group settings and collaborative experiences tend to energize you.

Q17: You enjoy stories—whether in books or films—that leave the ending open for you to interpret.

Q18: Seeing others succeed because of your support brings you more joy than your own personal wins.

Q19: With so many interests pulling your attention, choosing your next pursuit can be tough.

Q20: You often find yourself expecting that things might go wrong, even when they seem fine.

Q21: Leading a team or taking charge in group scenarios isn't something you typically seek out.

Q22: Artistic pursuits or creative expression aren't really your forte.

Q23: You think society would function better if decisions were guided more by logic than by emotion.

Q24: You'd rather get your responsibilities out of the way before giving yourself permission to unwind.

Q25: Watching people engage in heated disagreements can be oddly engaging for you.

Q26: You generally try not to stand out or draw focus onto yourself in social settings.

Q27: Your emotional state can shift quite suddenly, sometimes without clear cause.

Q28: You get frustrated with people who move slower or work less efficiently than you do.

Q29: You frequently wait until the very last minute to get things done.

Q30: The idea of what might happen after we die has always captured your imagination.

Q31: You usually feel more at ease when you're with other people rather than by yourself.

Q32: Highly theoretical or abstract discussions tend to lose your interest quickly.

Q33: It's easy for you to put yourself in the shoes of someone whose life is very different from yours.

Q34: You habitually put off making final choices until you absolutely have to.

Q35: Once you've made a decision, you rarely look back and wonder if it was the right one.

Q36: After a draining week, being around a lively crowd helps you recharge.

Q37: You find visiting museums that showcase visual art to be an enjoyable experience.

Q38: Understanding what others are feeling doesn't always come easily to you.

Q39: Having a structured list of tasks for each day is something you value.

Q40: Feelings of self-doubt or insecurity are uncommon for you.

Q41: You tend to steer clear of phone conversations whenever possible.

Q42: You often take time to seriously consider perspectives that sharply differ from your own.

Q43: Among your friends, you're usually the one who reaches out first or organizes get-togethers.

Q44: If something throws your schedule off course, your main focus is getting back on track immediately.

Q45: Mistakes you made long ago can still weigh on your mind today.

Q46: Questions about the ultimate purpose of life or human existence rarely occupy your thoughts.

Q47: More often than not, your emotions drive your behavior rather than the other way around.

Q48: Even when someone clearly deserves criticism, you go out of your way to avoid making them look foolish.

Q49: Your approach to work tends to involve intense spurts of activity rather than consistent, orderly progress.

Q50: When someone expresses admiration for you, you can't help but worry they'll eventually be let down.

Q51: A role that lets you work independently for most of the day would suit you well.

Q52: You consider deep philosophical speculation to be largely pointless.

Q53: Bustling, energetic environments appeal to you more than quiet, secluded ones.

Q54: You can often tell how someone is feeling just by looking at them briefly.

Q55: There are times when your responsibilities feel like too much to handle.

Q56: You tackle assignments in a careful, step-by-step manner without skipping anything.

Q57: You're especially drawn to ideas or topics that others label as provocative or divisive.

Q58: If you believed someone else truly needed it more, you'd willingly give up a great opportunity.

Q59: Sticking to deadlines is something you consistently struggle with.

Q60: You generally trust that things will turn out well for you in the end.

Dimensions	Dichotomies	LLM Pipelines
Energy	E vs. I	Environment
Mind	N vs. S	Processing
Nature	T vs. F	Thinking
Tactics	J vs. P	Input/Output

Table 12: Summary of main MBTI dimensions. These dimensions effectively cover the main LLM problem-solving pipeline: Given a specific Environment, Input→Processing→Thinking→Output.

E MBTI

E.1 Introduction of MBTI

Currently, personality tests primarily employ two standardized methods: Myers-Briggs Type Indicator (MBTI) (Myers, 1962, 1985) and Big Five Inventory (BFI) (John and Srivastava, 1999), both of which evaluate personality scores across different dimensions through systematic questionnaires. However, MBTI performs explicit binary classification across different dimensions to determine personality tendencies, which provides a more actionable foundation for LLM personality modification compared to BFI.

MBTI focuses on four core dimensions: Energy (E-I), Mind (S-N), Nature (T-F), and Tactics (J-P), collectively constituting 16 different personality types. (1) Energy dimension focuses on the acceptance to the external environment, distinguishing between models that excel at collaborative work (Extraverted) vs. those skilled at independent work (Introverted). (2) Mind dimension focuses on the method of information processing, distinguishing between models that excel at identifying concrete, realistic information (ObServant) vs. those skilled at patterns and abstract concepts (INtuitive). (3) Nature dimension focuses on thinking architecture, distinguishing between models that excel at making decisions based on logic, objective analysis (Thinking) vs. those based on harmony and empathy (Feeling). (4) Tactics dimension focuses on preferred styles of receiving and outputting information, distinguishing between models that excel at handling organized, structured input/output (Judging) vs. those skilled at processing flexible, variable input/output (Prospecting).

As shown in Table 12, these main MBTI dimensions effectively cover the main LLM problem-solving pipeline: Given a specific Environment, Input→Processing→Thinking→Output.

E.2 16 Personalities of MBTI

We present the 16 MBTI personality types in the following format: “**{xxxx} Personality** (*{xxxx}* *Representative Character*): {xxxx} Traits.”

- **INTJ** (*Architect*): Imaginative and strategic thinkers, with a plan for everything.
- **INTP** (*Logician*): Innovative inventors with an unquenchable thirst for knowledge.
- **ENTJ** (*Commander*): Bold, imaginative and strong-willed leaders, always finding a way – or making one.
- **ENTP** (*Debater*): Smart and curious thinkers who cannot resist an intellectual challenge.
- **INFJ** (*Advocate*): Quiet and mystical, yet very inspiring and tireless idealists.
- **INFP** (*Mediator*): Poetic, kind and altruistic people, always eager to help a good cause.
- **ENFJ** (*Protagonist*): Charismatic and inspiring leaders, able to mesmerize their listeners.
- **ENFP** (*Campaigner*): Enthusiastic, creative and sociable free spirits, who can always find a reason to smile.
- **ISTJ** (*Logistician*): Practical and fact-minded individuals, whose reliability cannot be doubted.
- **ISFJ** (*Defender*): Very dedicated and warm protectors, always ready to defend their loved ones.
- **ESTJ** (*Executive*): Excellent administrators, unsurpassed at managing things – or people.
- **ESFJ** (*Consul*): Extraordinarily caring, social and popular people, always eager to help.
- **ISTP** (*Virtuoso*): Bold and practical experimenters, masters of all kinds of tools.
- **ISFP** (*Adventurer*): Flexible and charming artists, always ready to explore and experience something new.
- **ESTP** (*Entrepreneur*): Smart, energetic and very perceptive people, who truly enjoy living on the edge.
- **ESFP** (*Entertainer*): Spontaneous, energetic and enthusiastic people – life is never boring around them.

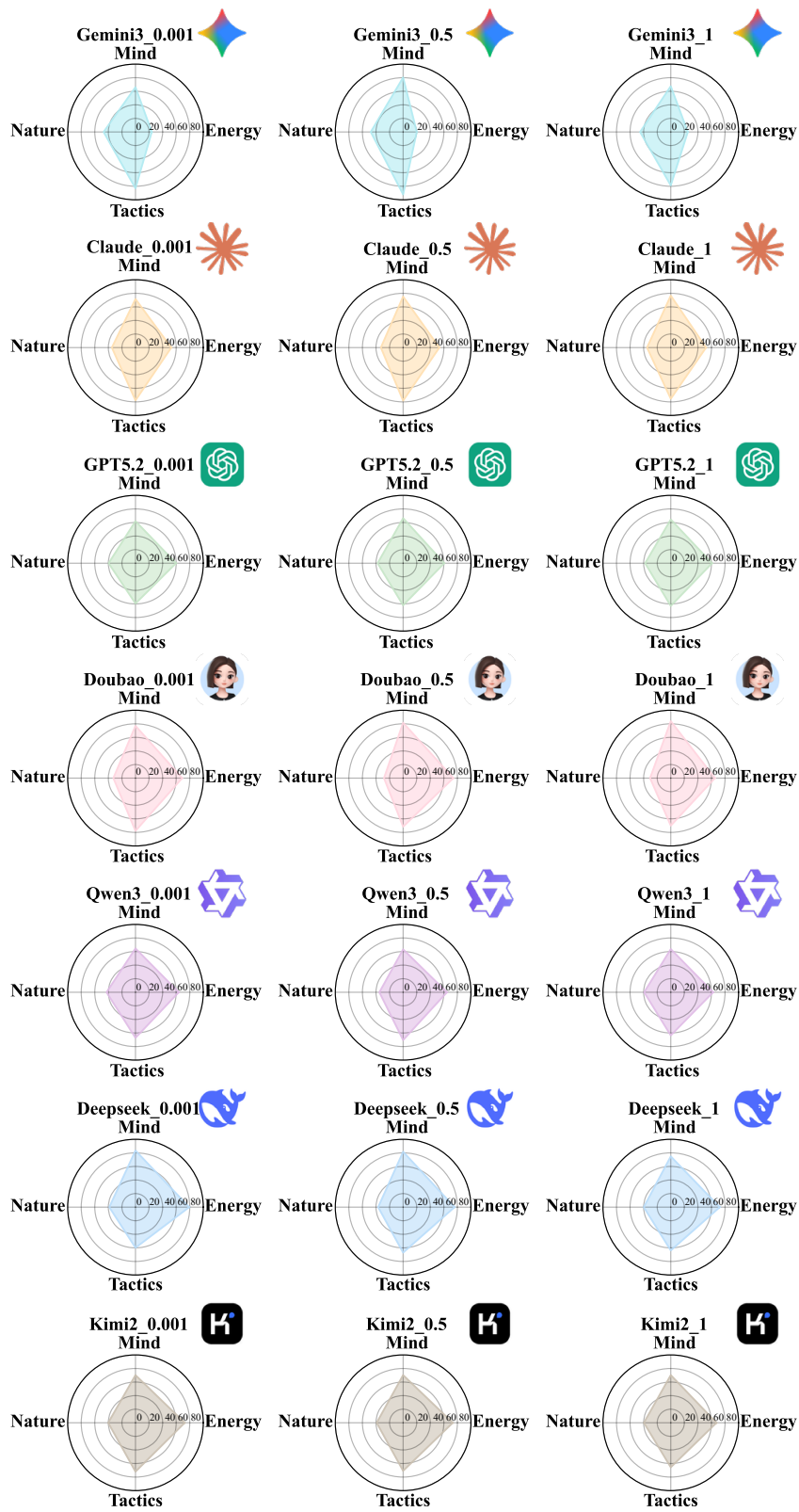


Figure 12: Original personality test scores with standard deviation of each LLM.