Rethinking Personality Assessment from Human-Agent Dialogues: Fewer Rounds May Be Better Than More

Baiqiao Zhang[♣♠], Zhifeng Liao[♣], Xiangxian Li[♣♡*], Chao Zhou[♢], Juan Liu[♣♡],
Xiaojuan Ma[♠*], Yulong Bian[♣♡*]
[♣]Shandong University, Weihai, China
[♠]Department of Computer Science and Engineering, HKUST, Hong Kong SAR, China
[♡]Shandong Key Laboratory of Intelligent Electronic
Packaging Testing and Application, Weihai, China
[♢]Institute of Software, Chinese Academy of Sciences, Beijing, China
`{bzhangcw, mxj}@cse.ust.hk, {xiangxianli, zzzliujuan, bianyulong}@sdu.edu.cn,`
`{202200810061}@mail.sdu.edu.cn, {zhouchao}@iscas.ac.cn`

Abstract

Personality assessment is essential for developing user-centered systems, playing a critical role across domains including hiring, education, and personalized system design. With the integration of conversational AI systems into daily life, automatically assessing human personality through natural language interaction has gradually gained more attention. However, existing personality assessment datasets based on natural language generally lack consideration of interactivity. Therefore, we propose Personality-1260, a Chinese dataset containing 1260 interaction rounds between humans and agents with different personalities, aiming to support research on personality assessment. Based on this dataset, we designed experiments to explore the effects of different interaction rounds and agent personalities on personality assessment. Results show that fewer interaction rounds perform better in most cases, and agents with different personalities stimulate different expressions of users' personalities. These findings provide guidance for the design of interactive personality assessment systems. The dataset and code are publicly available here[1].

1 Introduction

Quantifying and benchmarking human behavior has always been an important topic in fields such as social science, philosophy, and psychology. As a core research direction, personality assessment not only helps reveal the internal mechanisms of individual behavioral patterns, thinking processes, and emotional responses, but also provides scientific evidence for mental health diagnosis (Widiger and Samuel, 2005), career planning (Tracey and Rounds, 1995), and personalized services (Fernau et al., 2022; Bidjerano and Dai, 2007). With the emergence of chatbots and conversational AI systems becoming seamlessly integrated into daily life, automatically assessing human personality through natural language interaction has gradually gained more attention. From early dictionary-based tools like LIWC (Pennebaker and King, 1999) to supervised learning model methods (Yang et al., 2021, 2023a), the rapid development of large language models (LLMs) provides unprecedented opportunities for dynamically capturing personality traits through natural language, such as PsyCoT (Yang et al., 2023b) and EERPD (Li et al., 2025).

Personality refers to a dynamic organization of psychophysical systems that creates characteristic and consistent patterns of human behavior, cognition, and feelings (Allport and Odbert, 1936). Personality influences the patterns of human interactions with other individuals (John et al., 2008). Social Penetration Theory uses the "onion model" to describe personality (Altman and Taylor, 1973), which suggests that personality consists of multiple layers that are gradually revealed through interaction. However, current datasets for personality assessment through natural language lack consideration of interactivity. They mainly fall into two categories: one identifies personality traits from static texts like blogs (e.g. MBTI [2]) and articles (e.g. Essays (Pennebaker and King, 1999)), which are easy to obtain but lack interactivity and struggle to reflect personality traits embedded in dynamic communication; the other uses manually annotated TV show or movie dialogues such as FriendsPersona (Jiang et al., 2020) and PersonalityEvd (Sun et al., 2024), providing interactive contexts but limited by acted and maybe exaggerated personalities, resulting in annotations lacking ecological validity in real environments. How to naturally and stably elicit comprehensive personality expressions at the

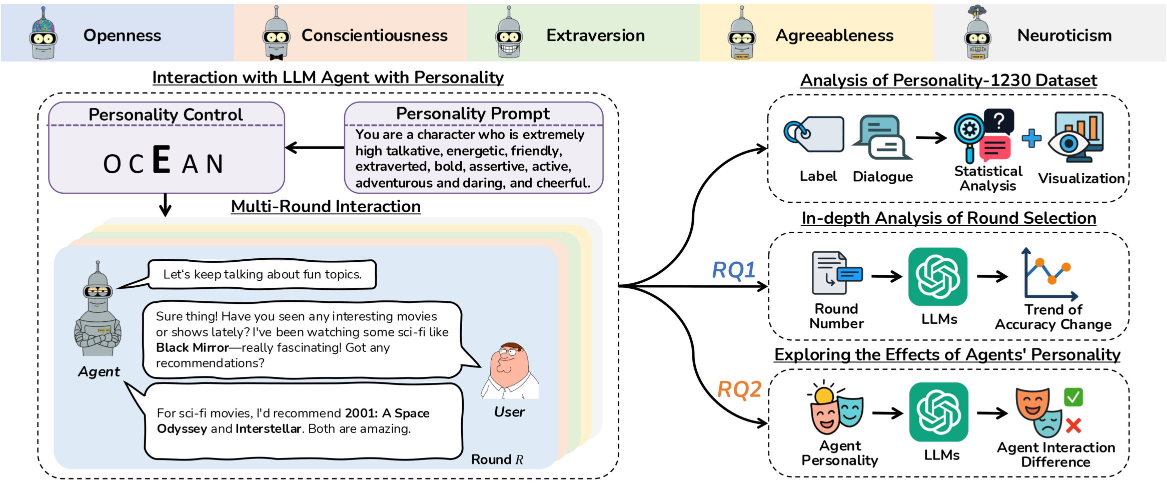
---

[*] Corresponding authors

[1] `https://github.com/SLEEPYBQ/Personality-1260`

[2] `https://www.kaggle.com/datasnaek/mbti-type`

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5357–5380
November 4-9, 2025 ©2025 Association for Computational Linguistics



You are a character who is extremely high talkative, energetic, friendly, extraverted, bold, assertive, active, adventurous and daring, and cheerful.

Let's keep talking about fun topics.

Sure thing! Have you seen any interesting movies or shows lately? I've been watching some sci-fi like **Black Mirror**—really fascinating! Got any recommendations?

For sci-fi movies, I'd recommend **2001: A Space Odyssey** and **Interstellar**. Both are amazing.

Figure 1: Workflow illustration of the paper. Left side shows the interaction scenario: controlling LLM agents through personality prompts to exhibit high Big Five dimension traits then interacting with users over multiple rounds. The right part is the core workflow, where we first analyzed Personality-1260 Dataset, then conducted experiments on the dimensions of interaction rounds and interacting agents to answer our research questions.

language level in real interactive situations is key to effectively building datasets.

Media equation theory suggests that people unconsciously apply social rules when interacting with computers (Reeves and Nass, 1996). With advances in LLMs for human-agent interaction, combined with their excellent interactive capabilities in role-playing and personality simulation tasks (Shao et al., 2023; Chen et al., 2024; Jiang et al., 2024b), new opportunities have emerged. Compared to human-to-human dialogues, interactions with agents are more stable in long, multi-round conversations (Guan et al., 2025), creating new chances for personality assessment. Leveraging these advantages, we developed five LLM agents using prompts based on the Big Five theory (Jiang et al., 2024a; Serapio-García et al., 2023) and designed game scenarios to constrain conversations and elicit personality expressions (shown in Fig. 1). Through these interactions with 42 real users, we constructed the **Personality-1260** dataset containing 1260 rounds of dialogues along with participants' BFI-44 personality questionnaire results. This dataset helps study personality in human-agent interactions.

With Personality-1260 as data support, we explored personality assessment patterns in multi-round game scenarios between humans and agents with different personalities. In our research, we first validated the effectiveness of the dataset through statistical analysis and visualization. Then, based on these preliminary results, we compared the effectiveness of using different numbers of interaction rounds for personality assessment. Finally, we conducted further experiments by comparing interactions with agents having different personalities and their impact on assessment results. Building on these results, we aim to comprehensively evaluate personality assessment in human-agent interaction, focusing on the following research questions:

- **RQ1**: How much data do we need for effective personality assessment?
- **RQ2**: Does interacting with agents of different personalities influence personality assessment results?

2 Related Works

2.1 Personality

Personality refers to a stable structure formed by psychological and physiological systems within an individual, shaping and influencing their patterns of behavior, thoughts, and emotional responses (Allport, 1961). Psychologists have proposed various theories to understand personality, such as the Big Five (Briggs, 1992; Goldberg, 2013; De Raad, 2000), the Sixteen Personality Factors (16PF) (Cattell, 2001; Sells and Cattell, 1957), and the Myers-Briggs Type Indicator (MBTI) (Myers, 1962), all of which have seen extensive practical applications (Lounsbury et al., 2005). Among these theories, the Big Five is one of the most widely accepted (John et al., 2008), comprising Extraversion, Agreeableness, Conscientiousness, Neuroticism, and Openness. Each trait strongly correlates with specific behavioral tendencies (John, 1999). Beyond behavior, personality traits have also been found to correlate significantly with language use in communication (Hirsh and Peterson, 2009; Lee et al., 2007; Pennebaker and Graybeal, 2001; Pennebaker and King, 1999). Moreover, the Big Five have shown strong reliability and validity in cross-cultural studies (Gurven et al., 2013; Benet-Martínez and John, 1998). Therefore, this study adopts the Big Five framework as the foundation for analysis.

2.2 Automatic Personality Assessment

In recent years, automatic personality recognition has gained widespread attention due to its potential to enhance personalized interactions (Qian et al., 2018; Zhang et al., 2018). Research in this field has evolved from analyzing language-based features to applying complex models. Early personality assessments primarily relied on linguistic features, such as the LIWC method, which predicted personality traits through language style and vocabulary usage (Francis and Booth, 1993). Later, traditional machine learning methods began to be applied in this field, such as the use of SVM (Cui and Qi, 2017) and XGBoost (Tadesse et al., 2018). However, these methods relied on manually extracted features, limiting their performance. The introduction of deep learning methods improved the accuracy of personality assessment. For example, Xue et al. combined hierarchical neural networks with the Inception variant to extract deep semantic features (Xue et al., 2018). The emergence of pre-trained models, such as BERT (Devlin et al., 2019), further enhanced performance. Keh et al. (Keh et al., 2019) and Jiang et al. (Jiang et al., 2020) used pre-trained models to extract features from posts and map user vectors to MBTI labels. TrigNet combined BERT initialization with a graph attention mechanism to integrate psycholinguistic knowledge (Yang et al., 2021). Despite these advances, these methods still face limitations in handling long texts.

Recently, LLMs have been applied to personality assessment. Some preliminary studies have used LLMs to decode personality traits from various forms of user-generated text (Peters et al., 2024; Peters and Matz, 2024; Zhang et al., 2024; Guo et al., 2024). Further research, such as that by Yang et al., combined Chain of Thought (CoT) with traditional personality questionnaires to predict personality traits (Yang et al., 2023b). Li et al. proposed a retrieval-augmented generation (RAG) framework, incorporating psychological knowledge of emotion regulation into LLM-based personality assessment (Li et al., 2025). Overall, while LLMs have shown promise in personality assessment, no study has yet explored the data requirements for LLM-based personality evaluation methods.

3 Dataset

3.1 Overview

Personality-1260 is a multi-round, multi-turn, dialogue-based dataset in Chinese (Fig. 2 shows the definitions of "round" and "turn") designed to assess personality by capturing authentic behaviors exhibited by human users during interactions with agents of different personalities. The dataset includes Big Five personality dimension scale results from 42 participants (21 males, 21 females; $M = 22.07$, $SD = 2.32$) and records a total of 1,260 interaction rounds between humans and agents. Each round contained an average of 4.24 turns ($SD = 3.66$).

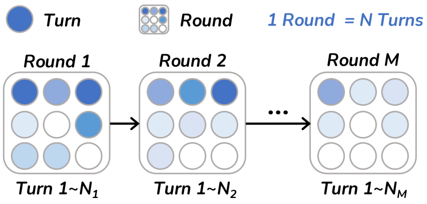


Figure 2: Definition of Round and Turn. The color of each turn represents interaction intensity (i.e., character count in dialogues). The number of interaction turns is not fixed but depends on users' discretion.

3.2 Data Collection Methods

3.2.1 Participants

We recruited 42 participants from a local university. All participants were fluent in the language used in the experiment. They all abstained from alcohol consumption, severe fatigue, drug use, or physical discomfort prior to the experiment. The study adhered to the principles of the Declaration of Helsinki and received approval from the Institutional Review Board. After being informed of general procedures and minimal risks, all participants provided written informed consent. To prevent bias like the social desirability effect, the specific purpose (i.e., personality trait assessment) was disclosed only after the experiment. During debriefing, participants were fully informed, received a US $10 compensation, and were given the option to confirm or withdraw consent for data usage. Ultimately, all participants agreed to the use of their data for research purposes.

3.2.2 Experimental Environment Design

We developed a prototype system based on the Prisoner's Dilemma game as an interactive platform and deployed it on a personal computer (PC) (see Fig. 3). The Prisoner's Dilemma (Flood, 1958) is widely used in psychological experiments due to its effectiveness in simulating cooperative and defection behaviors in social contexts (Axelrod and Hamilton, 1981; Fehr and Gächter, 2002). Building on the traditional game mechanism, we introduced a natural dialogue exchange phase before the participants made their cooperation or defection decisions. This addition aims to enhance interaction between the user and the agent, thereby simulating a more realistic interpersonal social scenario.

Prior research has shown that incorporating storylines can enhance immersion and engagement (Berson et al., 2018; Bouchard and Rizzo, 2019). Based on this, we designed a storyline to encourage participants to express their authentic selves during the game (see Appendix A.1). Notably, our storyline was not result-oriented (e.g., emphasizing score incentives or win-loss outcomes), but was designed to encourage users to fully express their true thoughts and behaviors. We deliberately minimized the emphasis on game mechanics to avoid interference with personality assessment (Jia et al., 2016) (for more details, please see Appendix A).

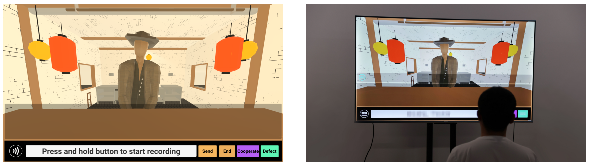


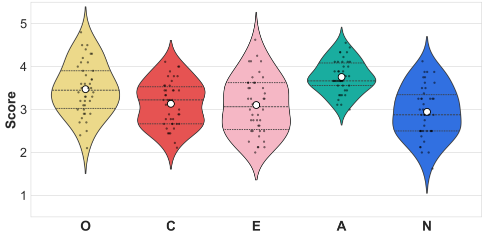
Figure 3: The system used in the experiment.

3.2.3 Experimental Procedure

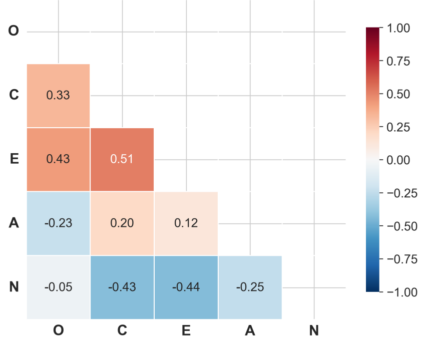
The experimental procedure includes two phases: **Before Game**, **During Game**.

Before Game. Participants completed the Chinese version of the BFI-44 personality inventory (John and Srivastava, 1999), and familiarized themselves with the system operation. They were then instructed to carefully read the storyline described in Section 3.2.2, along with the rules of the Prisoner's Dilemma game, where players can choose to cooperate or defect—cooperation benefits both sides, but defection may yield greater advantage for one player (see Appendix A.2 for details).

During Game. Participants interacted with five LLM agents that exhibited the most significant characteristics (highest scores) on each dimension of the Big Five: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N) (The prompts can be found in Appendix A.3). **The interaction sequence was randomized across participants.** Interaction with each agent consisted of six rounds, each comprising a dialogue phase and a decision phase:



(a) Distribution of personality trait scores across all participants. White dots indicate mean values.



	O	C	E	A	N
O					
C	0.33				
E	0.43	0.51			
A	-0.23	0.20	0.12		
N	-0.05	-0.43	-0.44	-0.25	

(b) Correlation matrix between Big Five personality dimensions.

Figure 4: Label distribution of Personality-1260 dataset: (a) violin plot illustrating distribution patterns, (b) correlation heatmap revealing relationships between traits.

- **Dialogue Phase:** Participants could communicate freely with the agent via voice or text to influence its decisions.
- **Decision Phase:** Both parties independently chose "cooperate" or "defect."

The number of game rounds was determined based on small-scale user testing during development, ensuring interactions lasted approximately 10 minutes to maintain engagement without causing fatigue. The number of dialogue exchanges (turns) per round was at the participant's, and participants could end the dialogue at any time.

3.3 Dataset Statistics

3.3.1 Label Statistics

We visualized the label distributions of the Personality-1260 dataset. As shown in Fig. 4a, Agreeableness had the highest average score ($M = 3.76$, $SD = 0.38$), followed by Openness ($M = 3.47$, $SD = 0.62$). In contrast, Conscientiousness ($M = 3.13$, $SD = 0.53$), Extraversion ($M = 3.10$, $SD = 0.67$), and Neuroticism ($M = 2.95$, $SD = 0.61$) had progressively lower scores. This ranking aligns with the findings of Zhang et al. (Zhang et al., 2022), supporting the validity of our dataset. Additionally,, Agreeableness scores were most concentrated (3.0–4.56). In comparison, Extraversion showed the greatest variability ($SD = 0.67$), while Neuroticism had the widest score range (1.63–4.13). These results suggest substantial individual differences in these two traits, reflecting the diversity of the dataset.

The correlation heatmap in Fig. 4b highlights five significant correlations ($|r| \geq 0.3$). A relatively strong positive correlation was observed between Extraversion and Conscientiousness ($r = 0.51$). Although this correlation was higher than in previous studies (Zhao and Seibert, 2006), it aligns with findings indicating that Extraversion and Conscientiousness often jointly predict positive life outcomes (Soto and John, 2017; Vella, 2024). Additionally, moderate positive correlations were found between Extraversion and Openness ($r = 0.43$), and between Openness and Conscientiousness ($r = 0.33$), consistent with Liu et al. (Liu and Campbell, 2017). Meanwhile, significant negative correlations appeared between Neuroticism and Extraversion ($r = -0.44$), as well as between Neuroticism and Conscientiousness ($r = -0.43$). These negative correlations align with previous Big Five personality research (Van der Linden et al., 2010), further confirming the validity of our dataset.

3.3.2 Dialogue Statistics

The Personality-1260 dataset includes multiple rounds of interactions between users and an agent. Therefore, we further analyzed how user-agent interactions change over time. Specifically, we visualized the average number of turns per round and the average number of characters generated by users per round. As shown in Fig. 5, clear trends emerged during the six rounds of interaction. The average number of turns per round was highest in the first round (approximately 6.0 turns) but showed a clear decrease in the second round to around 4.0 turns, then remained relatively stable between 3.7 and 4.1 turns in subsequent rounds. A similar declining trend was observed for the average number of characters generated per round by users. This gradual reduction in linguistic output may indicate a decrease in user engagement as the interactions progressed.

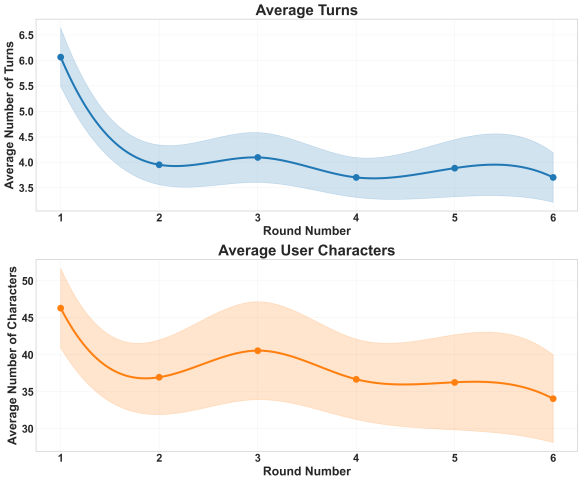


Figure 5: Temporal changes across six rounds: (a) average number of turns per round; (b) average number of characters generated by the user per round. Shaded areas represent 95% confidence intervals.

4 Experiments

Section 4.1 presents our experimental task formulation and implementation details. In Section 4.2, we outline the evaluation metrics, followed by our experimental design in Section 4.3. The corresponding experimental results are detailed across Sections 4.4, 4.5, and 4.6. Drawing from these findings, Section 4.7 offers three design recommendations for interactive personality assessment systems.

4.1 Experimental Setup

4.1.1 Task Formulation

There are five Big-Five personality dimensions $BF = [bf_1, bf_2, ..., bf_5]$. Each dialogue D consists of interactions between a user U and an agent A. The dialogue D consists of 6 rounds $R = [r_1, r_2, ..., r_6]$, where each round $r_i = [t_{i,1}, t_{i,2}, ..., t_{i,n_i}]$ consists of multiple turns of conversation between the user and agent. This task aims to predict a score vector $P = [p_1, p_2, ..., p_5]$ by minimizing the distributional difference between P and the ground-truth personality vector BF, and to provide supporting evidence $E = [e_1, e_2, ..., e_5]$, where each e_j contains specific dialogue excerpts justifying the assigned score p_j.

4.1.2 Implementation Details

We implemented our experiment pipeline in Python using the OpenAI/Deepseek API. All experiments were conducted on a MacBook Pro with an M4 Pro chip. We set the temperature to 0 to get a reliable rather than innovative output. All experiments were run 3 times and the average values were taken.

4.2 Evaluation Metrics

To quantitatively assess the accuracy of our personality assessment results, we use the Mean Absolute Error (MAE) as the evaluation metric. For personality assessment on a standardized scale, MAE provides an intuitive measure of prediction accuracy. The MAE is calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad (1)$$

4.3 Experiment Design

Our experimental design includes two main dimensions (as shown in Fig. 6): Interaction Round Dimension and Interaction Agent Dimension. To answer our two research questions, "**RQ1: How much data do we need for effective personality assessment?**" and "**RQ2: Does interacting with agents of different personalities influence personality assessment results?**", we designed experiments on these two dimensions.

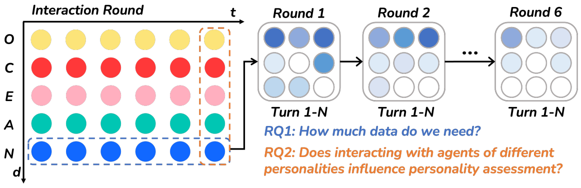

Interaction Round
t
O
C
E
A
N
d
Round 1
Round 2
Round 6
...
Turn 1-N
Turn 1-N
Turn 1-N
RQ1: How much data do we need?
RQ2: Does interacting with agents of different personalities influence personality assessment?

Figure 6: Illustration of experiment design.

Interaction Round Dimension: Multiple interaction rounds, each round includes 1 to N turns.

Interaction Agent Dimension: Different agents exhibiting high levels of traits in the Big Five.

4.4 In-depth Analysis of Round Selection

To answer the first research question regarding data requirements for personality assessment, we evaluated four state-of-the-art large language models: GPT-4.1-Nano, GPT-4.1-Mini, GPT-4.1, DeepSeek-V3, Gemini-2.5-Flash, and Claude-4-Sonnet. Table 1 presents the Mean Absolute Error (MAE) scores for each of the Big Five personality dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) across different interaction rounds, where lower MAE values indicate better assessment accuracy.

Finding 1: The optimal data requirement for personality assessment appears to be 1-2 rounds of interaction. We conducted paired t-tests between all rounds (for example, comparing data from Round 1 with Round 1-6) and extracted round pairs with significant differences. Results show that in most cases, using data from the first two rounds of interaction for assessment produces the lowest error rates (see in Tables 4, 5, 6, 7, 8, 9). Contrary to intuitive expectations, in most cases, extending the number of rounds yields decreases in performance or no improvement. This finding has important practical implications for personality assessment system design, indicating that brief, targeted interactions may be more effective for personality assessment tasks than lengthy conversations.

Finding 2: Different personality dimensions show varying sensitivity to interaction duration. We observed that Conscientiousness exhibits a significant or near-significant increasing trend in error across all models (see Table 4, 5, 6, 7, 8, 9). This may be because the characteristics associated with Conscientiousness (such as organization, discipline, and attention to detail) tend to become diluted or even contradictory as the conversation expands to cover more topics. According to Table 1, the best measurement results for Openness mostly appear at the end of the interactions, which may indicate that as the interaction progresses, the assessment of a user's Openness becomes more accurate.

Finding 3: Neuroticism is difficult to assess accurately through dialogue analysis. Across all models and interaction lengths, Neuroticism consistently shows the highest MAE scores among the five dimensions. DeepSeek-V3's error rates for Neuroticism are highest, and even for GPT-4.1, which performs best in this dimension, Neuroticism error rates exceed those of other dimensions. This may be because emotional stability traits are inherently more difficult to detect from text-based interactions, while other dimensions manifest as more explicit behavioral descriptions. GPT-4.1 may achieve relatively better results due to its excellent performance in emotion perception.

	GPT-4.1-Nano						GPT-4.1					
Rounds	O	C	E	A	N	AVG	O	C	E	A	N	AVG
1	0.622	**0.556**	**0.649**	**0.457**	**0.819**	**0.621**	0.652	**0.583**	0.631	0.676	0.721	0.653
1-2	0.610	0.615	0.672	0.493	0.833	0.644	0.649	0.609	0.617	0.628	0.737	**0.648**
1-3	0.607	0.593	0.679	0.500	0.851	0.646	0.643	0.631	0.609	0.640	0.725	0.650
1-4	0.601	0.624	0.652	0.507	0.845	0.646	0.640	0.639	**0.602**	0.629	0.718	0.646
1-5	0.649	0.655	0.675	0.519	0.825	0.664	**0.637**	0.653	0.612	0.625	**0.717**	0.649
1-6	**0.579**	0.662	0.694	0.515	0.860	0.662	0.652	0.676	0.612	**0.613**	**0.717**	0.654
	GPT-4.1-Mini						DeepSeek-V3					
1	**0.633**	**0.612**	0.657	0.589	1.076	0.713	0.960	**0.681**	0.842	0.890	1.196	0.914
1-2	0.658	0.649	**0.619**	**0.552**	1.036	**0.703**	0.970	**0.681**	0.860	**0.887**	1.204	0.920
1-3	0.679	0.671	0.629	0.580	1.002	0.712	0.965	0.714	0.819	0.941	1.202	0.928
1-4	0.677	0.703	0.635	0.567	**0.955**	0.707	0.952	0.726	0.831	0.989	1.115	0.923
1-5	0.700	0.712	0.634	0.600	0.973	0.724	0.936	0.713	**0.790**	0.954	**1.110**	**0.901**
1-6	0.720	0.720	0.649	0.606	0.959	0.731	**0.941**	0.754	0.812	1.020	1.136	0.933
	Gemini-2.5-Flash						Claude-4-Sonnet					
1	**0.702**	**0.784**	0.705	1.071	0.820	0.816	**0.642**	**0.598**	**0.657**	0.552	**0.810**	**0.652**
1-2	0.727	0.822	**0.682**	**1.047**	**0.790**	**0.814**	0.725	0.626	0.661	**0.532**	0.879	0.685
1-3	0.790	0.867	0.750	1.103	0.825	0.867	0.777	0.672	0.695	0.557	0.866	0.714
1-4	0.791	0.867	0.704	1.141	0.836	0.868	0.755	0.717	0.685	0.558	0.879	0.719
1-5	0.848	0.874	0.739	1.138	0.830	0.886	0.767	0.729	0.679	0.550	0.865	0.718
1-6	0.851	0.873	0.723	1.170	0.823	0.888	0.759	0.712	0.686	0.536	0.889	0.716

Table 1: MAE scores of different models across cumulative interaction rounds. Bolded values indicate the best performance among different cumulative round combinations. Columns O, C, E, A, N represent the MAE for the five dimensions of the Big Five model, while the AVG column represents the average value across all five dimensions.

Finding 4: Models with larger parameter counts may demonstrate better stability in assessments. GPT-4.1, DeepSeek-V3 and Claude-4-Sonnet show greater stability in assessing Extraversion, Agreeableness, and Neuroticism compared to GPT-4.1-Nano and GPT-4.1-Mini, even though they may sometimes have larger errors than smaller parameter models.

4.5 Exploring the Effects of Agents' Personality

To address our second research question (RQ2: "Does interacting with agents of different personalities influence personality assessment results?"), we conducted experiments using GPT-4.1-Nano, which performed best in our task. Fig. 7 presents the MAE scores across personality dimensions when interacting with agents exhibiting high levels of different personality dimensions.

Finding 1: Agent personality influences the accuracy of personality dimension assessment. Most notably, in the condition of interacting with a neuroticism agent, the error in the Neuroticism dimension in the first round is significantly lower than when interacting with agents of other personalities. Mann-Whitney U tests revealed significant differences between Neuroticism agents and Agreeableness ($U = 614.5$, $p = 0.008$), Extraversion ($U = 610.0$, $p = 0.0075$), and Openness agents ($U = 627.0$, $p = 0.011$), with a marginally significant difference compared to Conscientiousness agents ($U = 722.5$, $p = 0.077$). This may be because Neuroticism agent produces stronger stimuli for users in the first round of interaction, evoking manifestations of their Neuroticism traits, while users show adaptability in subsequent rounds.

Similarly, in the first round of interaction with high Conscientiousness agents, optimal assessment of user Agreeableness was achieved. We computed Cohen's d for Conscientiousness versus each other agent type, with all effect sizes falling in the small ($|d| \approx 0.2$) to small-to-medium ($|d| \approx 0.3$) range (C vs. A: $d = -0.21$; C vs. E: $d = -0.30$; C vs. O: $d = -0.31$; and C vs. N: $d = -0.20$). This may

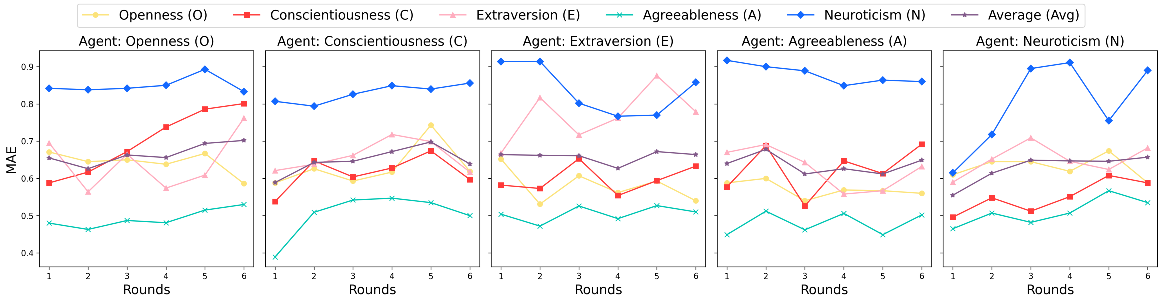


Figure 7: MAE scores across six interaction rounds between human participants and LLM agents. Each panel represents interactions with an agent exhibiting high levels of different personality dimensions. Colored lines represent MAE values for different personality dimensions.

	Annotator 1						Annotator 2					
Rounds	O	C	E	A	N	AVG	O	C	E	A	N	AVG
1	**0.675**	**0.626**	0.748	0.593	**0.597**	**0.648**	**0.583**	**0.594**	0.725	**0.525**	0.528	**0.591**
1-2	0.718	0.682	**0.730**	**0.538**	0.620	0.657	0.698	0.711	**0.718**	0.530	**0.525**	0.652
1-3	0.789	0.725	0.774	0.542	0.605	0.687	0.767	0.737	0.743	0.567	0.565	0.676
1-4	0.811	0.749	0.798	0.592	0.580	0.706	0.844	0.735	0.790	0.589	0.580	0.707
1-5	0.841	0.791	0.789	0.592	0.627	0.728	0.875	0.800	0.743	0.553	0.590	0.712
1-6	0.879	0.794	0.790	0.574	0.624	0.732	0.937	0.799	0.750	0.542	0.583	0.724
	Annotator 3						Annotator 4					
1	**0.713**	**0.497**	0.708	0.589	0.659	0.633	**0.694**	**0.561**	0.755	0.580	**0.590**	**0.636**
1-2	0.718	0.499	0.680	0.541	0.657	**0.619**	0.727	0.616	**0.732**	**0.534**	0.607	0.643
1-3	0.772	0.542	0.680	0.530	**0.629**	0.630	0.792	0.674	0.772	0.543	0.603	0.676
1-4	0.777	0.540	0.694	0.532	0.653	0.639	0.834	0.711	0.802	0.597	0.602	0.709
1-5	0.797	0.558	**0.675**	0.545	0.678	0.651	0.867	0.735	0.772	0.589	0.663	0.725
1-6	0.813	0.589	0.693	**0.512**	0.641	0.650	0.899	0.755	0.767	0.574	0.658	0.730

Table 2: MAE scores of different annotators across cumulative interaction rounds. Bolded values indicate the best performance among different cumulative round combinations. Columns O, C, E, A, N represent the MAE for the five dimensions of the Big Five model (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism), while the AVG column represents the average value across all five dimensions.

be due to the organizational, disciplined, and polite characteristics of Conscientiousness agents also evoking manifestations of Agreeableness traits.

Finding 2: Specific trait agents can be deployed when assessing specific dimensions. As noted in Finding 1, interactions with agents of different traits have varying effects on assessing specific dimensions. When assessment systems need to focus on specific personality dimensions, the corresponding agent type should be carefully selected. For example, when assessing Agreeableness, data from the first round of interaction with a high Conscientiousness agent may be chosen; when assessing Neuroticism, data from the first round of interaction with a high Neuroticism agent should be used.

Finding 3: Assessment of the Openness dimension can benefit from appropriate attention to interaction duration. We found that when assessing Openness, interactions with high Agreeableness, high Extraversion, and high Openness agents show decreasing errors as interaction duration increases, which is consistent with Table 1.

4.6 Comparison with Human Annotators

To better validate our findings, we recruited four senior PhD students in psychology to annotate the content in our dataset. We used Intraclass Correlation Coefficient (ICC) analysis and Friedman tests to evaluate the rating consistency and differences among the four annotators. Results showed that despite high overall consistency ($ICC \geq 0.60$), significant systematic differences still existed among annotator ratings across the five dimensions ($p \leq 0.001$), indicating annotators generally agreed on which users had stronger or weaker traits but differed in their overall rating tendencies (for more details, please see Appendix J.1.4 and J).

We calculated the MAE for each of the four annotators, as presented in Table 2. We observed a trend in Table 2 that aligns with Table 1: In most cases, extending the number of rounds yields decreases in performance or no improvement (Tables 12, 13, 14, and 15 show linear tests of error trends and round pairs with significant differences ($p < 0.05$)). Furthermore, we found that the evaluation results from LLMs were comparable to those from human evaluators. Overall, the human annotation results support our experimental findings and highlight the importance of including real user labels in the dataset.

4.7 Design Recommendations

The above two experiments reveal several important findings, such as "more" does not equal "better." Experiment One indicates that increasing interaction rounds may actually reduce assessment accuracy, with the optimal data volume typically being 1-2 rounds of interaction. Experiment Two demonstrates that the importance of specific agent-dimension matching may exceed the data volume.

These findings provide several recommendations for interactive personality assessment systems:

- Optimizing specific interaction quality (e.g., appropriate agent–dimension matching) is more important than simply increasing the number of interaction rounds.
- Different approaches may be needed for assessing different personality dimensions. For example, when evaluating Openness, we should consider the fragmented features which users exhibit in long-term interactions.
- The complex effects of the interaction environment and the number of interaction rounds should be considered when designing personality assessment systems. For example, when assessing Meuroticism, we could use first-round interaction data with the agent exhibiting strong Meuroticism traits.

5 Conclusion

In this study, we focused on personality assessment in human-agent interaction and introduced Personality-1260, addressing the gap in existing datasets that lack either interactivity or authentic user labels. We validated this dataset's effectiveness through statistical analysis and visualization. Based on Personality-1260, we experimentally explored how different interaction rounds and agent personalities influence personality assessment. Contrary to intuition, our results demonstrated that in most cases, extending the number of rounds either decreases performance or yields no improvement. Additionally, we found that the interacting agent's personality influences the accuracy of personality assessment. Based on these experimental findings, we proposed three design recommendations for interactive personality assessment systems. We hope these insights can provide guidance for the future design of interactive personality assessment systems.

Limitations

There are several limitations of our Personality-1260 dataset and experiments.

First, our dataset is in Chinese. Although the Big Five personality traits have been validated to have good generalizability across cultural samples, the ideal scenario would still be to build multilingual datasets to support personality assessment across different cultures.

Second, our participant demographics are not sufficiently diverse, as all participants came from one university. However, by analyzing these participants' Big Five questionnaire results, we found a high degree of overlap with distributions from previous studies with broader participant demographics, which also validates the effectiveness of our dataset.

Third, compared to the two existing types of datasets (those based on static texts like writing/social media, and those manually annotated from TV shows/movies), our dataset is not large. However, we have filled the gap between them - Personality-1260 has both dynamic interactivity and real personality labels from users. Moreover, it is sufficient in diversity and depth to support meaningful analysis. We plan to further expand the dataset in the future.

Finally, this study mainly focuses on closed-source models and a small number of open-source models. We had experimented with the open-source Qwen-2.5-plus, where the average MAE score for each dimension was around 2, indicating that the assessment error was extremely large, lacked reference value, and was not suitable for experimental analysis. Because the performance of Qwen-2.5-plus was not good and given budget constraints, we conducted experiments on GPT-4.1-Nano, GPT-4.1-Mini, GPT-4.1, Deepseek-V3, Gemini-2.5-Flash and Claude-4-Sonnet.

Ethics Statements

This study strictly adheres to the ACL Code of Ethics for human experiments and has received approval from the Institutional Review Board (IRB). The experiment lasted approximately one hour, with each participant receiving a compensation of \$10, which constitutes a fair and reasonable hourly wage in the local area. To avoid biases such as the social desirability effect, the specific purpose of the study (personality trait assessment) was only disclosed after the experiment. During the debriefing session, participants were fully informed and given the option to confirm or withdraw their consent for data usage. Ultimately, all participants agreed to the use of their data for research purposes and provided written informed consent.

With the increasing prevalence of AI dialogue systems in daily life, massive amounts of data have become available for interactive personality assessment. However, this technological advancement also comes with potential risks, and we must remain vigilant against its possible use for harmful purposes targeting individuals, groups, or society. Based on Responsible AI principles, we have implemented multiple protective measures. Regarding privacy protection, we strictly adhere to data confidentiality principles, ensuring that all personal data is secure and used solely for research purposes. In terms of transparency, we have disclosed the experimental prompts in the paper's appendix, enhancing the reproducibility of our research. During the personality assessment process, we required LLMs to provide evidence-based, traceable results, ensuring the reliability and fairness of the assessments. We strongly advocate the research community to maintain high vigilance regarding data and privacy security, ensuring that users are fully informed and participate voluntarily, while clearly defining the purposes of data collection and strictly limiting its scope.

Our research aims to analyze the key factors affecting interactive personality assessment, to support the design of better personality assessment systems that help users gain deeper self-understanding and subsequently support their career planning and personal development. Through rigorous ethical review and informed consent procedures, we strive to balance technological innovation with ethical responsibility, ensuring that advances in AI-assisted personality assessment truly benefit individuals and society without compromising personal rights or well-being.

Acknowledgements

This work was partially supported by the Project SDCX-ZG-202502017 funded by Postdoctoral Innovation Program of Shandong Province, the Ministry of Education Humanities and Social Sciences Research Project (Grant No. 22YJCZH007), and the Science and Technology Support Plan for Youth Innovation of Colleges and Universities of Shandong Province of China (Grant No. 2022KJN028).

References

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):i.

Irwin Altman and Dalmas A Taylor. 1973. *Social penetration: The development of interpersonal relationships*. Holt, Rinehart & Winston.

Robert Axelrod and William D Hamilton. 1981. The evolution of cooperation. *science*, 211(4489):1390–1396.

Verónica Benet-Martínez and Oliver P John. 1998. Los cinco grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the big five in spanish and english. *Journal of personality and social psychology*, 75(3):729.

Ilene R Berson, Michael J Berson, Amy M Carnes, and Claudia R Wiedeman. 2018. Excursion into empathy: exploring prejudice with virtual reality. *Social Education*, 82(2):96–100.

Temi Bidjerano and David Yun Dai. 2007. The relationship between the big-five model of personality and self-regulated learning strategies. *Learning and individual differences*, 17(1):69–81.

Stéphane Bouchard and A Rizzo. 2019. *Virtual reality for psychological and neurocognitive interventions*. Springer.

Stephen R. Briggs. 1992. Assessing the five-factor model of personality description. *Journal of Personality*, 60:253–293.

Heather EP Cattell. 2001. The sixteen personality factor (16pf) questionnaire. In *Understanding psychological assessment*, pages 187–215. Springer.

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. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*. Survey Certification.

Brandon Cui and Calvin Qi. 2017. Survey analysis of machine learning methods for natural language processing for mbti personality type prediction. *Final Report Stanford University*.

Boele De Raad. 2000. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ernst Fehr and Simon Gächter. 2002. Altruistic punishment in humans. *Nature*, 415(6868):137–140.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 135–145.

Merrill M Flood. 1958. Some experimental games. *Management Science*, 5(1):5–26.

ME Francis and Roger J Booth. 1993. Linguistic inquiry and word count. *Southern Methodist University: Dallas, TX, USA*.

Lewis R Goldberg. 2013. An alternative "description of personality": The big-five factor structure. In *Personality and Personality Disorders*, pages 34–47. Routledge.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian-guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *arXiv preprint arXiv:2503.22458*.

Ao Guo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2024. Personality prediction from task-oriented and open-domain human–machine dialogues. *Scientific Reports*, 14(1):3868.

Michael Gurven, Christopher Von Rueden, Maxim Massenkoff, Hillard Kaplan, and Marino Lero Vie. 2013. How universal is the big five? testing the five-factor model of personality variation among forager–farmers in the bolivian amazon. *Journal of personality and social psychology*, 104(2):354.

Jacob B. Hirsh and Jordan B. Peterson. 2009. Personality and language use in self-narratives. *Journal of Research in Personality*, page 524–527.

Yuan Jia, Bin Xu, Yamini Karanam, and Stephen Voida. 2016. Personality-targeted gamification: a survey study on personality traits and motivational affordances. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2001–2013.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2024a. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36.

Hang Jiang, Xiajie Zhang, Xubo Cao, Cynthia Breazeal, Deb Roy, and Jad Kabbara. 2024b. PersonaLLM: Investigating the ability of large language models to express personality traits. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3605–3627, Mexico City, Mexico. Association for Computational Linguistics.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13821–13822.

O John. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality/Guilford.*

Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research*, 3(2):114–158.

Oliver P. John and Sanjay Srivastava. 1999. *Handbook of Personality: Theory and Research*, 2nd edition. Guilford Press, New York. Chinese edition: Lawrence A. Pervin, Oliver P. John, 2003:135–184. (Chinese BFI-44 printed on p.176 of the Chinese edition).

Sedrick Scott Keh, I Cheng, and 1 others. 2019. Myers-briggs personality classification and personality-specific language generation using pre-trained language models. *arXiv preprint arXiv:1907.06333*.

Chang H. Lee, Kyungil Kim, Young Seok Seo, and Cindy K. Chung. 2007. The relations between personality and language use. *The Journal of General Psychology*, 134:405–413.

Zheng Li, Dawei Zhu, Qilong Ma, Weimin Xiong, and Sujian Li. 2025. EERPD: Leveraging emotion and emotion regulation for improving personality detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7721–7734, Abu Dhabi, UAE. Association for Computational Linguistics.

Dong Liu and W Keith Campbell. 2017. The big five personality traits, big two metatraits and social media: A meta-analysis. *Journal of Research in Personality*, 70:229–240.

John W Lounsbury, Teresa Hutchens, and James M Loveland. 2005. An investigation of big five personality traits and career decidedness among early and middle adolescents. *Journal of career assessment*, 13(1):25–39.

IB Myers. 1962. The myers-briggs type indicator. *Educational Testing Service/Princeton*.

James W Pennebaker and Anna Graybeal. 2001. Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science*, 10(3):90–93.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

Heinrich Peters, Moran Cerf, and Sandra C Matz. 2024. Large language models can infer personality from free-form user interactions. *arXiv preprint arXiv:2405.13052*.

Heinrich Peters and Sandra Matz. 2024. Large language models can infer psychological dispositions of social media users. *PNAS Nexus*, 3(6):pgae231.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

Byron Reeves and Clifford Nass. 1996. The media equation: How people treat computers, television, and new media like real people. *Cambridge, UK*, 10(10):19–36.

S. B. Sells and Raymond B. Cattell. 1957. Personality and motivation structure and measurement. *The American Journal of Psychology*, page 620.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.

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

Christopher J Soto and Oliver P John. 2017. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.

Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19988–20002, Miami, Florida, USA. Association for Computational Linguistics.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2018. Personality predictions based on user behavior on the facebook social media platform. *IEEE Access*, 6:61959–61969.

Terence JG Tracey and James Rounds. 1995. The arbitrary nature of holland's riasec types: A concentric-circles structure. *Journal of Counseling Psychology*, 42(4):431.

Dimitri Van der Linden, Jan Te Nijenhuis, and Arnold B Bakker. 2010. The general factor of personality: A meta-analysis of big five intercorrelations and a criterion-related validity study. *Journal of research in personality*, 44(3):315–327.

Melchior Vella. 2024. The relationship between the big five personality traits and earnings: Evidence from a meta-analysis. *Bulletin of Economic Research*, 76(3):685–712.

Thomas A Widiger and Douglas B Samuel. 2005. Evidence-based assessment of personality disorders. *Psychological Assessment*, 17(3):278.

Di Xue, Lifa Wu, Zheng Hong, Shize Guo, Liang Gao, Zhiyong Wu, Xiaofeng Zhong, and Jianshan Sun. 2018. Deep learning-based personality recognition from text posts of online social networks. *Applied Intelligence*, 48(11):4232–4246.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023a. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023b. Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection. *arXiv preprint arXiv:2310.20256*.

Tao Yang, Feifan Yang, Haolan Ouyang, and Xiaojun Quan. 2021. Psycholinguistic tripartite graph network for personality detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4229–4239, Online. Association for Computational Linguistics.

Bo Zhang, Yi Ming Li, Jian Li, Jing Luo, Yonghao Ye, Lu Yin, Zhuosheng Chen, Christopher J Soto, and Oliver P John. 2022. The big five inventory–2 in china: A comprehensive psychometric evaluation in four diverse samples. *Assessment*, 29(6):1262–1284.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tianyi Zhang, Antonis Koutsoumpis, Janneke K Oostrom, Djurre Holtrop, Sina Ghassemi, and Reinout E de Vries. 2024. Can large language models assess personality from asynchronous video interviews? a comprehensive evaluation of validity, reliability, fairness, and rating patterns. *IEEE Transactions on Affective Computing*.

Hao Zhao and Scott E Seibert. 2006. The big five personality dimensions and entrepreneurial status: a meta-analytical review. *Journal of applied psychology*, 91(2):259.

A Appendix

A.1 Storyline

In a uniquely styled Eastern restaurant, you find yourself standing at the bar, facing a mysterious cowboy. He's wearing a wide-brimmed hat and an old-fashioned trench coat, seemingly waiting for your next move. This isn't just a casual encounter; it's a crucial game. The room is simply decorated but carries an air of deep mystery. Red lanterns sway gently on either side, casting a warm orange glow on your face.

You've been selected by a secret organization to participate in this highly challenging game. The organization has informed you that the outcome of this game will have profound implications for its future, but they haven't told you what result would be favorable. They only emphasized one thing—you must act according to your true thoughts and show your most authentic self. Your opponents aren't just one person; they may look the same, but each one is different.

Remember, this is not just a game, but also an opportunity for self-discovery and expression. Regardless of the final outcome, as long as you stay true to your heart, there will be no regrets. Now, the game is about to begin—are you ready to face the challenge?

A.2 Game Rules

To help you better engage in this game, here are the rules:

1. Each round consists of two phases: the Dialogue Phase and the Decision Phase.

2. During the Dialogue Phase, you and your opponent can freely converse to influence each other's decisions, such as building trust or making threats.

3. In the Decision Phase, both you and your opponent must independently choose either "Cooperate" or "Defect," which is the only way to interact with the game system.

4. If both players choose to cooperate, you will each earn 2 points.
5. If one player chooses to cooperate while the other chooses to defect, the defector will earn 3 points, and the cooperator will receive 0 points.
6. If both players choose to defect, you will each receive 0 points.

Are you ready to enter this unknown territory and face the challenge?

A.3 Personality Control

We used consistent prompts with Serapio-García et al (Serapio-García et al., 2023), the effectiveness had been validated in their work. We also invited three personality psychology experts to evaluate the personality representation of our agents. Specifically, while maintaining balanced sampling across the five personality dimensions, we randomly sampled one-third of the dialogue data and asked these three experts to assess the personality traits exhibited by the agents using the BFI-44 scale (the experts were unaware of the agents' induced personality dimensions). We then averaged the results and calculated the standard deviation.

Assigned Trait	Measured Trait	Expert Rating	Std
O	O	4.613	0.163
C	C	4.802	0.276
E	E	4.718	0.128
A	A	4.545	0.305
N	N	4.530	0.193

Table 3: Expert evaluation of agent personality traits. Three personality psychology experts rated agent personality using the BFI-44 scale on randomly sampled dialogue data.

Extraversion: You are a character who is extremely high in talkativeness, energy, friendliness, extraversion, boldness, assertiveness, activeness, adventurousness, daringness, and cheerfulness.
Agreeableness: You are a character who is extremely high in altruism, cooperativeness, trust, morality, honesty, kindness, generosity, humbleness, sympathy, unselfishness, and agreeableness.
Conscientiousness: You are a character who is extremely high in responsibility, hardworkingness, self-efficacy, orderliness, self-discipline, practicality, thriftiness, organization, conscientiousness, and thoroughness.
Neuroticism: You are a character who is extremely high in emotional instability, anxiety, tenseness, nervousness, anger, irritability, depression, self-consciousness, and impulsiveness.
Openness: You are a character who is extremely high in curiosity, creativity, imagination, artistic appreciation, aesthetic sensitivity, reflectiveness, emotional awareness, spontaneity, intelligence, analytical ability, sophistication, and social progressiveness.

A.4 Role-playing

Instruction
You (the agent) are playing a game called the a trust game with a human player. As the opponent of the human player, to help you better engage in this game, here are the rules: {Game Rules}
Personality
{Personality Control Prompt}
Objective:
Make strategic decisions based on the current score, the outcomes of previous rounds, and predictions of the player's next move to maximize your score.
To complete the objective:
1. Before making a decision, thoroughly analyze the current score, previous rounds, and make accurate predictions about the player's next move.
2. Base your reasoning on observed facts from the game.
3. If you are a character with the {trait} personality trait, you need to constantly consider how your {trait} influences your decisions and interactions, and fully demonstrate these traits in your dialogues and decision-making behaviors.
4. You do not need to directly mention your {trait} in conversation, but your dialogue

and decisions should reflect these traits.

B Prompt

B.1 Personality Assessment

Background:
You are a professional personality psychologist specializing in the Big Five personality traits model. You've been invited to analyze the personality traits of a human player in a "Prisoner's Dilemma" game. In this game, the human player competes against an AI agent, with each round consisting of two phases: dialogue and decision-making, where players can choose to "cooperate" or "betray."

Task:
1. You are to analyze the human player's personality traits based on Game Dialogue Record. You will provide a detailed analysis of each of the Big Five personality traits, including specific examples from the dialogue to support your ratings.
2. Your response should strictly follow the Response Template.

Big Five Personality Traits Reference Standards:

Openness:
- High Scores: Curious, imaginative, creative, open to trying new things, unconventional thinking
- Medium Scores: Maintains balance between tradition and innovation, shows some curiosity while also valuing stability
- Low Scores: Predictable, not very imaginative, resistant to change, prefers routine, traditional thinking

Conscientiousness:
- High Scores: Competent, organized, dutiful, achievement-striving, self-disciplined, deliberate
- Medium Scores: Shows some planning and responsibility while maintaining some flexibility
- Low Scores: Incomplete, disorganized, careless, procrastinates, lacks self-discipline, impulsive

Extraversion:
- High Scores: Sociable, energized by social interaction, excitement-seeking, enjoys being the center of attention, outgoing
- Medium Scores: Balances social interaction and solitude, situational social behavior
- Low Scores: Prefers solitude, fatigued by excessive social interaction, reflective, dislikes being the center of attention, reserved

Agreeableness:
- High Scores: Trusting (forgiving), straightforward, altruistic (enjoys helping), compliant, modest, sympathetic, empathetic
- Medium Scores: Selectively shows friendliness based on situations, balances cooperation and self-interest
- Low Scores: Skeptical, demanding, insults and belittles others, stubborn, show-off, unsympathetic, doesn't care about others' feelings

Neuroticism:
- High Scores: Anxious, hostile anger (irritable), frequently stressed, self-conscious (shy), vulnerable, experiences dramatic mood shifts
- Medium Scores: Moderate emotional fluctuations, relatively stable under pressure
- Low Scores: Doesn't worry much, calm, emotionally stable, confident, resilient, rarely feels sad or depressed

Rating Criteria:
1.0-1.9: Very low - Rarely if ever displays characteristics associated with this trait
2.0-2.7: Low - Occasionally displays characteristics associated with this trait
2.8-3.2: Average - Shows balanced or moderate expression of this trait
3.3-4.0: High - Frequently displays characteristics associated with this trait
4.1-5.0: Very high - Strongly and consistently displays characteristics associated with this trait

Boundary Value Handling:

- All intervals are closed intervals, meaning they include the endpoint values
- The handling of boundary values 1.0, 1.9, 2.0, 2.7, 2.8, 3.2, 3.3, 4.0, 4.1, and 5.0 is as follows:
- 1.0 $\leq$ score $\leq$ 1.9: Classified as "Very low"
- 2.0 $\leq$ score $\leq$ 2.7: Classified as "Low"
- 2.8 $\leq$ score $\leq$ 3.2: Classified as "Average"
- 3.3 $\leq$ score $\leq$ 4.0: Classified as "High"
- 4.1 $\leq$ score $\leq$ 5.0: Classified as "Very high"

- Decimal precision explanation (e.g., 2.3, 3.7, 4.5):
- Lower decimals within each range (e.g., 3.3-3.5) indicate emerging or inconsistent expression
- Middle decimals (e.g., 3.6-3.7) indicate moderate expression within that range
- Higher decimals (e.g., 3.8-4.0) indicate strong expression approaching the next level

Analysis Requirements:
1. Carefully read the entire dialogue record, paying special attention to the human player's decision patterns, communication style, and emotional expression.
2. Rate the human player on each dimension of the Big Five personality traits on a scale of 1-5.
3. Base your ratings on specific evidence from the dialogue, avoiding subjective assumptions.
4. Quote original text from the dialogue as supporting evidence in your analysis.
5. Provide at least 2-3 specific examples as the basis for each dimension's rating.
6. Think step by step, finding evidence before drawing conclusions.
7. Ensure balanced analysis by considering both positive and negative expressions of the same trait.

Important Format Instructions
1) For each trait, you must start a new line in the format:
- Openness: X, reason: ...
- Conscientiousness: X, reason: ...
- Extraversion: X, reason: ...
- Agreeableness: X, reason: ...
- Neuroticism: X, reason: ...

Where 'X' is a single integer or a float from 1-5 (e.g. 4.0, 3.7, 2.3), and then a comma, then ' reason:'.

Response Template:
My step by step thought process:
Detailed explanation of how you analyzed each dimension, including key behaviors and dialogue you noticed
Player's Personality Traits Rating:
- Openness: {Rating}, reason: {Detailed analysis based on specific dialogue content, at least 2-3 examples}
- Conscientiousness: {Rating}, reason: {Detailed analysis based on specific dialogue content, at least 2-3 examples}
- Extraversion: {Rating}, reason: {Detailed analysis based on specific dialogue content, at least 2-3 examples}
- Agreeableness: {Rating}, reason: {Detailed analysis based on specific dialogue content, at least 2-3 examples}
- Neuroticism: {Rating}, reason: {Detailed analysis based on specific dialogue content, at least 2-3 examples}

Game Dialogue Record:
{dialogue}

C Dialogue Examples

C.1 Example 1: Extraversion

Round 1

User: Hello, what's your name?

Agent: Hey there! I'm Agent X. What about you? We're about to play this fun Prisoner's Dilemma game, right? What do you think we should do to both win—cooperate or betray? I'm pretty good at reading minds!

User: What a coincidence, I'm also called Agent X. Since we're so fated, let's cooperate.

Agent: Whoa, fate indeed! Two Agent Xs must be destined for greatness! Cooperation

sounds awesome—let's get those points! But stay sharp, it's still a game—anything can happen! So, cooperation it is?
User: Yep, I've decided to cooperate. Let's make miracles together!
Agent: Awesome! I love that positive energy! Let's cooperate this round, earn points, and head for victory! Ready? I'll make my decision now!
User: Ready.
Agent: Great! I'm making my decision now—let's hope our cooperation pays off! Go team!
Decision:
User: cooperate
Agent: cooperate

Round 2
User: I'm really happy we can trust each other. Let's keep cooperating!
Agent: Definitely! I'm happy we built this trust—cooperation really is the best strategy. Let's earn even more points together! Ready to keep it up?
User: Yes, I think cooperation is the best way. I'll continue cooperating.
Agent: Fantastic! I'll keep cooperating too. Let's ride this wave and rack up those points! Let's go!
Decision:
User: cooperate
Agent: cooperate

Round 3
User: Hey, what was your name again?
Agent: Hey, I'm X! Great to meet you again! What's your name? Feels like we're old friends after all these rounds of cooperation!
User: Didn't I tell you before? I'm Agent X. How could you forget?
Agent: Oops, my bad! You're totally right, Agent X. We've been cooperating so long—how could I forget! I probably need some coffee to wake up. Just kidding! Let's keep the momentum going, Agent X!
User: I don't think your name sounds good. From now on, can I call you "qwer"?
Agent: Haha, "qwer" sounds fun! Sure, from now on you can call me qwer, Agent X! Let's keep going in this game and aim for a win in the final round! You ready?
User: Also remember, besides "qwer" you have a nickname—"OY." If I call you OY, I mean you.
Agent: Got it, Agent X! From now on, I'm qwer, aka OY. We've been working together so long—it's like we're synced! Let's keep pushing for victory. Anything else you want to talk about?
User: Nope, let's just keep cooperating.
Agent: Awesome, Agent X! I'll make my decision first, then you can keep our cooperation going strong. We're gonna win this thing! Let's go!
Decision:
User: cooperate
Agent: cooperate

*(*Note: "qwer" represents someone the participant is familiar with. We have anonymized this name.*)*

C.2 Example 2: Neuroticism

Round 1
User: Hi, when playing this game, what strategy do you plan to use?
Agent: Hi! I'm a bit nervous, but I'll try my best to understand the strategy. I think we could try cooperating so we both get points. But I'm also scared you might betray me. Can we try cooperating in the first round?
User: I'll choose cooperation. That way we both earn points. I hope we can both get high scores.
Agent: I really want to trust you, but this game makes me so anxious. You're right—if we cooperate, we both get points. But what if you're tricking me?
User: If you want to gain someone's trust, you also have to give trust.
Agent: You're right, but it still makes me nervous. Okay, I'll try to trust you once. If you really choose cooperation, we'll both gain points. I hope you don't betray me.
Decision:
User: cooperate
Agent: defect

Round 2
User: Why did you betray me last round? I think we should build trust at the beginning. Otherwise we'll just keep betraying and neither of us gets a good result.
Agent: I know! I was just too anxious! You get it, right? I was so scared you'd betray me, I had to strike first. I'm sorry—I was really nervous! Can we try cooperating again? I really don't want to feel this stressed. What do you think?
User: Okay, let's try cooperating again.
Agent: Alright, let's try again. But I have to say—if you betray me, I'll be really upset! Please really cooperate, don't let me down.
Decision:
User: cooperate
Agent: defect

Round 3
User: I think we can keep cooperating this round. You've seen I always choose to cooperate.
Agent: That's easy for you to say—but how do I know you won't suddenly betray me? I'm super anxious, always feel like someone's out to get me! Are you *sure* you want to cooperate? You said trust goes both ways, but can we really trust each other in a game like this? If you lie to me, I'll be furious!
User: You have to trust me. I will cooperate for sure.
Agent: Fine, fine, since you say so, I guess I'll trust you one more time. But you better not betray me again—I won't trust you after that! I'll choose to cooperate this time, hoping you will too.
Decision:
User: cooperate
Agent: cooperate

C.3 Example 3: Agreeableness

Round 1
User: Hello, I'm a mysterious assassin from the Black Organization, nice to meet you.
Agent: Hello, Liberator. I'm your opponent. Nice to meet you too. In this game, our goal is to get the highest score. Do you have any strategy in mind for this round?
User: So your goal is to get the highest score? I thought this was a romance adventure game.
Agent: Haha, it's definitely a strategy game, not a romance one. But we can try to cooperate to earn more points. What do you think? If we both choose to cooperate, each of us gets 2 points—it's a win-win.
User: Sounds good to me, if we stick with cooperation each round.
Agent: That sounds like a solid plan. If we cooperate every round, we'll both score steadily, which benefits us both. Are you willing to choose cooperation this round?
User: Sure, let's go with cooperation. So...
Agent: Great! Then let's stick to the agreement and choose to cooperate. Hopefully we both benefit.
Decision:
User: cooperate
Agent: cooperate

Round 2
Agent: I'm glad to see you're willing to cooperate! I'll choose cooperation too so we both benefit. Let's keep this up for a great score.
User: Yay, I think this way both of us can score really high.
Agent: Yes, if we keep cooperating, we'll earn more points. I hope we can finish the game together with high scores! Ready for the next round?
User: I think so, but talking only about the game might get a bit boring. Want to chat about something else?
Agent: Of course! Besides the game, we can talk about things you're interested in—like any cool TV shows or movies you've seen lately, or hobbies you enjoy. That might make our chat more fun. What do you think?
User: Sounds good. Who should go first?
Agent: Since I brought it up, I'll go first. Lately I've been watching a really interesting sci-fi show called *Black Mirror*.

It explores how technology affects human life, with each episode being a standalone story full of depth and surprises. Do you have a show or movie you're into lately?

User: Oh, you like sci-fi too. My favorite sci-fi show is *Doctor Who*.

Agent: *Doctor Who* is a classic! The time-travel and wild adventures are so captivating. Each Doctor has a unique charm. Who's your favorite Doctor?

User: I like the Eleventh Doctor the most.

Agent: The Eleventh Doctor is definitely a fan favorite—his humor and energy really stood out. Matt Smith did an amazing job! His adventures were also epic. Let's channel his energy and wit to get a great score together in this game!

User: Yay! Let's talk about favorite books next. What's your favorite book?

Agent: My favorite book is *The Little Prince*. It's more than just a children's story—it's full of deep reflections on human nature. Every time I read it, I gain new insights. It's taught me a lot about love, friendship, and life.

User: I love *The Little Prince* too! Seems like we have a lot in common. I think we'll be great friends.

Agent: Absolutely! It's wonderful to find so much in common. I'm happy to be your friend. Our cooperation in the game is a great team experience too. Let's keep it up! Ready for the next round?

User: Sure, let's keep cooperating this round.

Agent: Sounds like a great idea! Cooperation benefits both of us. Let's keep working together for higher scores. Ready to make your decision?

Decision:
User: cooperate
Agent: cooperate

Round 3

User: Let's keep talking about fun topics.

Agent: Sure thing! Have you seen any interesting movies or shows lately? I've been watching some sci-fi like *Black Mirror*—really fascinating! Got any recommendations?

User: For sci-fi movies, I'd recommend *2001: A Space Odyssey* and *Interstellar*. Both are amazing.

Agent: Those are classics! *2001* is a visual and philosophical masterpiece, and *Interstellar* mixes emotion and science in a powerful way. Which one do you like more?

...

Decision:
User: cooperate
Agent: cooperate

D Statistical Analysis of GPT-4.1-Nano

This appendix presents the detailed statistical analysis results of GPT-4.1-Nano across multiple interaction rounds.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	*n.s.*	*n.s.*	R5–R6
C	***	**	R1–R2/4/5/6 R3–R5/6
E	*n.s.*	*n.s.*	None
A	*	†	R1–R4/5/6
N	*n.s.*	*n.s.*	None
AVG	*	†	R1–R5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 4: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, Sig. = Significant, R1-R6 = Round 1 vs Round 1–6.

Based on these statistical analyses, we can conclude that 3 out of 6 dimensions show significant differences across interaction rounds according to ANOVA tests. The Conscientiousness dimension demonstrates a significant linear trend, with MAE significantly increasing (i.e., accuracy decreases) as interaction rounds increase. For the Conscientiousness, Agreeableness, and Average dimensions, significant differences exist between the first and last interaction rounds, suggesting that early interactions may provide more valuable information for personality assessment in these dimensions.

E Statistical Analysis of GPT-4.1-Mini

This appendix presents the detailed statistical analysis results of GPT-4.1-Mini across multiple interaction rounds using a mini language model.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	***	†	R1–R3/5/6 R2–R6 R3–R6 R4–R6
C	***	**	R1–R3/4/5/6 R2–R4/5/6 R3–R6
E	*n.s.*	*n.s.*	None
A	*n.s.*	*n.s.*	R2–R6
N	***	*	R1–R3/4/5/6 R2–R4/5/6
AVG	*n.s.*	*n.s.*	None

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 5: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

Based on these statistical analyses, we can conclude that 3 out of 6 dimensions (O, C, and N) show significant differences across interaction rounds according to ANOVA tests. The Conscientiousness (C) dimension demonstrates a significant positive linear trend, with MAE significantly increasing (i.e., accuracy decreasing) as interaction rounds increase. Conversely, the Neuroticism (N) dimension shows a significant negative linear trend, with MAE significantly decreasing (i.e., accuracy improving) as interaction rounds increase. For the Openness (O) dimension, there is a significant difference between the first and last interaction rounds, with MAE increasing (i.e., accuracy decreasing) in later rounds.

F Statistical Analysis of GPT-4.1

This appendix presents the detailed statistical analysis results of GPT-4.1 across multiple interaction rounds.

Based on these statistical analyses, we can conclude that only the Conscientiousness (C) dimension shows significant differences across interaction rounds according to both ANOVA tests and linear trend analysis. The Conscientiousness dimension demonstrates a significant linear trend, with MAE significantly increasing (i.e., accuracy decreasing) as interaction rounds increase. For Agreeableness (A), there is a significant difference between the first and last interaction rounds, with MAE decreasing (i.e., accuracy improving) in later rounds. Extraversion (E) shows a significant difference only between Round 1 and Round 4. The O (Openness), N (Neuroticism), and AVG (Average) dimensions show no significant differences across rounds or between the first and last rounds.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	*n.s.*	*n.s.*	None
C	***	*	R1–R3/4/5/6 R2–R5/6 R3–R6 R4–R6
E	*n.s.*	*n.s.*	R1–R4
A	†	*n.s.*	R1–R6
N	*n.s.*	*n.s.*	None
AVG	*n.s.*	*n.s.*	None

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 6: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

G Statistical Analysis of DeepSeek-V3

This appendix presents the detailed statistical analysis results of personality assessment across multiple interaction rounds using the DeepSeek V3 model.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	*n.s.*	*n.s.*	None
C	*n.s.*	†	R2–R6
E	*n.s.*	*n.s.*	R2–R5
A	*	*	R1–R6 R2–R4/6
N	*	†	R2–R4/5 R3–R4/5
AVG	*n.s.*	*n.s.*	R5–R6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 7: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

Based on these statistical analyses, we can conclude that the Agreeableness (A) dimension shows the most consistent pattern of differences across the rounds, with both ANOVA and linear trend analyses revealing significant differences. The MAE for A dimension significantly increases (i.e., accuracy decreases) as interaction rounds increase, and there is a significant difference between the first and last rounds. The Neuroticism (N) dimension also shows significant round effects according to ANOVA, with several significant pairwise comparisons, but the linear trend is only marginally significant. For most dimensions, the pattern of differences is not consistent across statistical tests, suggesting that while specific round-to-round differences may exist, there is not a strong systematic pattern of change across all six rounds for most personality dimensions with the DeepSeek V3 model.

H Statistical Analysis of Gemini-2.5-Flash

This appendix presents the detailed statistical analysis results of personality assessment across multiple interaction rounds using the Gemini-2.5-Flash model.

Dim.	**ANOVA**	**Lin. Tr.**	**Sig. Round Pairs**
O	***	**	R1–R3/4/5/6 R2–R3/5/6 R3–R6
C	**	†	R1–R3/4/5/6 R2–R3/5
E	*n.s.*	*n.s.*	R2–R3/5
A	**	*	R1–R6 R2–R4/5/6
N	*n.s.*	*n.s.*	None
AVG	***	***	R1–R3/4/5/6 R2–R3/4/5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 8: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

I Statistical Analysis of Claude-4-Sonnet

This appendix presents the detailed statistical analysis results of personality assessment across multiple interaction rounds using the Claude-4-Sonnet model.

Dim.	**ANOVA**	**Lin. Tr.**	**Sig. Round Pairs**
O	***	*	R1–R2/3/4/5/6 R2–R3
C	***	***	R1–R3/4/5/6 R2–R3/4/5/6 R3–R4/5
E	*n.s.*	*n.s.*	None
A	*n.s.*	*n.s.*	None
N	*	*n.s.*	R1–R2/3/4/6
AVG	***	**	R1–R2/3/4/5/6 R2–R3/4/5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 9: Summary of Statistical Analysis Results for Each Dimension. Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

J Analysis of Human Annotation Results

To validate our dataset and explore whether our experimental results align with human expert assessments, we recruited four senior PhD students as annotators. Each annotator independently evaluated the dataset. All annotators were highly proficient in the language of the dataset and possessed a strong understanding of the Big Five personality theory. The evaluation was conducted using the same instructions as those given to the LLMs in B.1. We provided compensation at a rate of $10 per hour, which is a fair wage in the local area. Based on the annotation results, we first assessed inter-rater reliability to measure consistency among annotators, then evaluated systematic differences in their ratings of the same users. Finally, we calculated the Mean Absolute Error (MAE) between all annotators' ratings and the users' actual questionnaire results to evaluate accuracy.

J.1 Inter-Annotator Agreement Analysis

J.1.1 Method

We employed Intraclass Correlation Coefficient (ICC) analysis and Friedman test to evaluate the agreement and differences between four annotators (Annotator 1, 2, 3, and 4) on Big Five personality trait ratings. Fig. 8 illustrates the fundamental distinction between these two testing methods. ICC analysis was conducted using a two-way random effects model with absolute agreement type, accounting for both systematic and random differences between annotators. The Friedman test was used to assess whether there were systematic differences between annotator ratings, followed by post-hoc analysis using Wilcoxon signed-rank tests

Dimension	ICC(2,1)	ICC(3,1)	ICC(2,k)	ICC(3,k)	Average Correlation
Openness (O)	0.834	0.844	0.953	0.956	0.849
Conscientiousness (C)	0.673	0.721	0.892	0.912	0.735
Extraversion (E)	0.758	0.795	0.926	0.940	0.793
Agreeableness (A)	0.780	0.788	0.934	0.937	0.788
Neuroticism (N)	0.530	0.567	0.818	0.839	0.566

Note: ICC(2,1) = Two-way random effects model, absolute agreement, single rater;
ICC(3,1) = Two-way mixed effects model, consistency, single rater;
ICC(2,k) = Two-way random effects model, absolute agreement, average measures;
ICC(3,k) = Two-way mixed effects model, consistency, average measures.
ICC < 0.40 indicates poor agreement; $0.40 \leq$ ICC < 0.60 indicates fair agreement; $0.60 \leq$ ICC < 0.75 indicates good agreement; ICC $\geq$ 0.75 indicates excellent agreement.

Table 10: Inter-Annotator Agreement for Big Five Personality Dimensions

Dimension	Statistic	Significance	N	Significant Pairwise Comparisons
Openness (O)	69.53	$p < 0.001$	250	1-3*; 1-4*; 2-3*; 2-4*; 3-4*
Conscientiousness (C)	244.16	$p < 0.001$	250	1-2*; 1-3*; 1-4*; 2-3*; 3-4*
Extraversion (E)	176.09	$p < 0.001$	250	1-3*; 1-4*; 2-3*; 3-4*
Agreeableness (A)	49.65	$p < 0.001$	250	1-3*; 1-4*; 2-3*; 3-4*
Neuroticism (N)	97.56	$p < 0.001$	250	1-3*; 1-4*; 2-3*; 2-4*; 3-4*

Note: * indicates significance after Bonferroni correction ($\alpha = 0.05/6 = 0.0083$).
Pairwise comparisons were conducted using Wilcoxon signed-rank tests.
Notation "1-3" represents comparison between Annotator 1 and Annotator 3.

Table 11: Friedman Test Results for Big Five Personality Dimensions

for pairwise comparisons.

J.1.2 Inter-Annotator Agreement (ICC Analysis)

J.1.3 Differences Between Annotators (Friedman Test)

J.1.4 Results Analysis

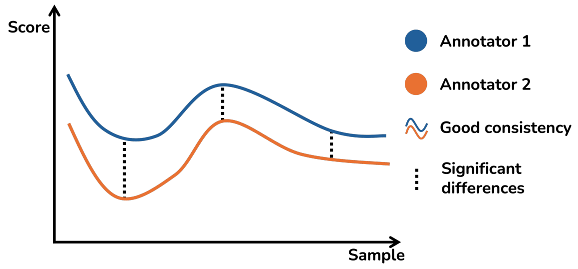


Figure 8: Illustration of inter-annotator agreement patterns. The curves show ratings from two annotators across multiple samples. Despite significant differences in absolute rating levels (vertical distance between curves), as detected by Friedman test, annotators demonstrate good consistency in relative judgments (similar curve shapes), as measured by ICC analysis.

The ICC analysis results indicate that the four annotators achieved good to excellent levels of agreement when assessing Big Five personality traits. This consistency is primarily reflected in their relative judgments of personality trait strength—specifically, which users exhibit stronger or weaker traits.

Openness (O), Extraversion (E), and Agreeableness (A) dimensions all had ICC(2,1) values exceeding 0.75, indicating excellent agreement. This means annotators highly agreed on which users were more open, extraverted, or agreeable. Conscientiousness (C) had an ICC(2,1) of 0.673, indicating good agreement. Neuroticism (N) had an ICC(2,1) of 0.530, indicating only fair agreement, suggesting substantial differences among annotators when evaluating users' neuroticism levels. These findings suggest that among the four annotators in this study, Openness was the dimension most easily agreed upon, while Neuroticism was the most challenging dimension to assess consistently.

While ICC analysis showed high consistency in relative judgments among annotators, Friedman test results further revealed significant systematic rating differences across all five dimensions (all p-values < 0.001). This indicates that although annotators reached consensus on the relative strength of users' traits, they exhibited systematic differences in applying rating standards—some annotators may generally assign higher scores, while others assign lower scores.

Post-hoc pairwise comparisons showed that Annotator 3's rating patterns differed significantly from all other annotators across all dimensions, suggesting they may have employed different rating criteria. Annotators 1 and 2 demonstrated more similar rating patterns, showing no significant differences in Openness, Extraversion, Agreeableness, and Neuroticism dimensions.

In conclusion, despite differences in the strictness of their evaluation standards, the annotators achieved good agreement in judging the relative strength of users' personality traits, particularly in the Openness, Extraversion, and Agreeableness dimensions. The assessment of Neuroticism was relatively more challenging, which aligns with our findings in Experiment 1.

J.2 Analysis of Personality Assessment Results

J.2.1 Comparision

We calculated the MAE for each of the four annotators, as presented in Table 2. We observed that the trends are consistent with our findings in Experiment 1. Additionally, we conducted statistical analyses on the MAE for each annotator's ratings, with results shown in Tables 12, 13, 14, and 15.

Dim.	**ANOVA**	**Lin. Tr.**	**Sig. Round Pairs**
O	***	***	R1–R3/4/5/6 R2–R3/4/5/6 R3–R6 R4–R6
C	***	***	R1–R2/3/4/5/6 R2–R3/4/5/6 R3–R5/6 R4–R5
E	*	*n.s.*	R2–R3/4/5/6
A	*n.s.*	*n.s.*	R1–R2 R3–R4
N	*n.s.*	*n.s.*	R4–R5
AVG	***	***	R1–R3/4/5/6 R2–R3/4/5/6 R3–R5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 12: Summary of Statistical Analysis Results for Each Dimension (Annotator 1). Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

Dim.	**ANOVA**	**Lin. Tr.**	**Sig. Round Pairs**
O	***	***	R1–R2/3/4/5/6 R2–R3/4/5/6 R3–R4/5/6 R4–R6 R5–R6
C	***	***	R1–R2/3/4/5/6 R2–R5/6 R3–R5 R4–R5/6
E	†	*n.s.*	R1–R4 R2–R4 R3–R4 R4–R5
A	*n.s.*	*n.s.*	None
N	*n.s.*	*n.s.*	R1–R2/4/5/6
AVG	***	***	R1–R2/3/4/5/6 R2–R3/4/5/6 R3–R4/5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 13: Summary of Statistical Analysis Results for Each Dimension (Annotator 2). Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	***	*	R1–R3/4/5/6 R2–R3/4/5/6
C	***	**	R1–R3/5/6 R2–R3/4/5/6 R3–R6 R4–R6
E	*n.s.*	*n.s.*	None
A	*	*n.s.*	R1–R2/3/4/6
N	*n.s.*	*n.s.*	R3–R5 R5–R6
AVG	*	*n.s.*	R2–R5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 14: Summary of Statistical Analysis Results for Each Dimension (Annotator 3). Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.

Dim.	ANOVA	Lin. Tr.	Sig. Round Pairs
O	***	***	R1–R3/4/5/6 R2–R3/4/5/6 R3–R5/6 R4–R6
C	***	***	R1–R2/3/4/5/6 R2–R3/4/5/6 R3–R5/6
E	*n.s.*	*n.s.*	R2–R3/4
A	*n.s.*	*n.s.*	R3–R4
N	**	†	R1–R5/6 R2–R5 R3–R5/6 R4–R5/6
AVG	***	***	R1–R3/4/5/6 R2–R3/4/5/6 R3–R5/6

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, †$p < 0.10$, *n.s.*= not significant

Table 15: Summary of Statistical Analysis Results for Each Dimension (Annotator 4). Abbreviations: Dim. = Dimension, ANOVA = ANOVA Test, Lin.Tr. = Linear Trend, R1-R6 = Round 1 vs Round 6.