Exploring a Gamified Personality Assessment Method through Interaction with LLM Agents Embodying Different Personalities

Baiqiao Zhang, *Student Member, IEEE*, Xiangxian Li, Chao Zhou, Xinyu Gai, Juan Liu, Xue Yang, Xiaojuan Ma, *Senior Member, IEEE*, Yong-Jin Liu, *Senior Member, IEEE*, Yulong Bian, *Member, IEEE*

Abstract—The low-intrusion and automated personality assessment is receiving increasing attention in mental health and psychology fields. This study explores an interactive approach for personality assessment, focusing on the multiplicity of personality representation. We propose a framework of Gamified Personality Assessment through Multi-Personality Representations (Multi-PR GPA). The framework leverages Large Language Models to empower virtual agents with different personalities. These agents elicit multifaceted human personality representations through engaging in interactive games. Drawing upon the multi-type textual data generated throughout the interaction, it achieves two modes of personality assessment (i.e., Direct Assessment and Questionnaire-based Assessment) and provides interpretable insights. Grounded in the classic Big Five personality theory, we developed a prototype system and conducted a user study to evaluate the efficacy of Multi-PR GPA. The results affirm the effectiveness of our approach in personality assessment and demonstrate its superior performance when considering the multiplicity of personality representation.

Index Terms—LLM Agents, Gamified Personality Assessment, Big Five.

1 Introduction

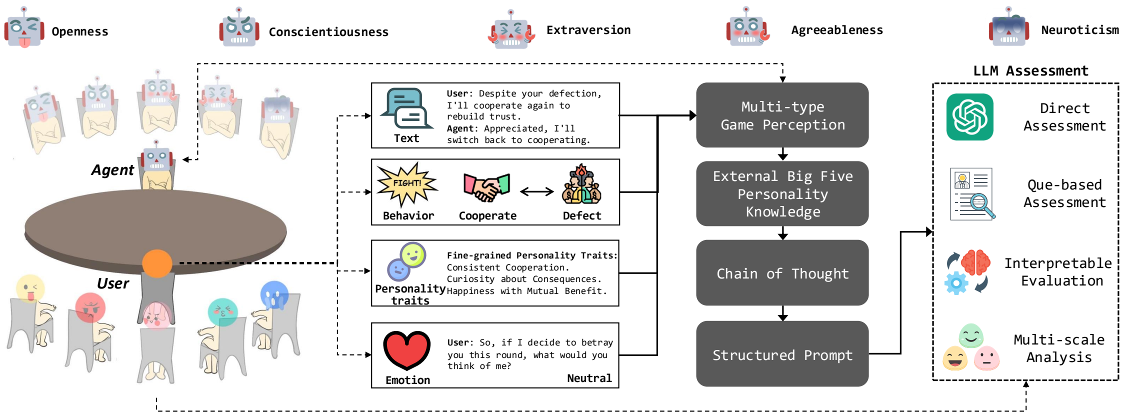
Mental health has become a significant global issue, profoundly impacting individual well-being and global public health systems [1]. A critical yet often overlooked foundation for effective mental health care is the accurate assessment of personality. Accurate personality assessment can serve not only as a screening tool for mental health risk [2], but also as a foundation for personalized intervention matching [3]. However, accurately assessing personality is nontrivial. It is inherently multi-faceted [4]. For instance, the widely adopted OCEAN model (a.k.a. the Big Five Personality Theory) characterizes human personality by five broad dimensions of traits: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [5]. Personality is also situation/context-dependent [4]. For example, an individual may exhibit introverted behavior in professional settings while acting extraverted among close friends. Consequently, human behavior, the external manifestation of one's personality [6], exhibits significant variations across different situations [7].

Researchers have made extensive efforts to evaluate personality. Traditional questionnaire-based assessments (e.g., the Big Five Personality Test [8]) rely on participants' subjective reports, which may be influenced by social desirability bias [9]. While projective tests (e.g., the Rorschach test [10]) and situational assessments attempt to mitigate such biases through indirect measurement, they largely rely on administrators' subjective judgments [11] and thus are limited by scarcity of expert resources [12]. Therefore, there has been a shift towards automated, objective approaches for psychological assessments. Several studies have utilized machine learning-driven analysis to infer personality from user-generated static text, such as Facebook/Twitter posts or blogs [13], [14]. While promising, these approaches fail to adequately capture personality traits embedded in dynamic social interactions. People naturally exhibit different aspects of themselves when interacting with others who exhibit differing personality traits, a phenomenon known as *multiplicity* of personality representations [15] (see Fig. 1). Gamified assessments have emerged to address the challenge of capturing dynamic personality traits to some extent, offering opportunities to evaluate personality through human behavior in serious games while achieving good engagement and assessment effectiveness [16], [17], [18]. Nevertheless, comprehensive personality assessment requires observing individuals' interactive behaviors across multiple interactive contexts and interaction partners. Most existing gamified assessments still do not adequately incorporate multi-situational observations or leverage robust tools to analyze complex personality-related data effectively.

The development of large language models (LLMs) offers new opportunities for interactive systems. On the one hand, LLMs can provide the analytical intelligence required to interpret complex personality-related data. Recent stud-

- *Baiqiao Zhang, Xiangxian Li, Xinyu Gai, Juan Liu and Yulong Bian are with Shandong University, Weihai, China.*
- *Baiqiao Zhang and Xiaojuan Ma are with The Hong Kong University of Science and Technology, Hong Kong SAR, China.*
- *Chao Zhou are with Institute of Software, Chinese Academy of Sciences, Beijing, China.*
- *Xue Yang and Yong-Jin Liu are with the Department of Computer Science and Technology, Tsinghua University, Beijing, China.*
- *Yulong Bian and Yong-Jin Liu are the corresponding authors.*

This work was supported in part by the Youth Innovation and Technology Support Program of Shandong Provincial Higher Education Institutions (2022KJN028) and the Natural Science Foundation of China (U2336214).
The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the HREC of Shandong University.



User: Despite your defection, I'll cooperate again to rebuild trust.
Agent: Appreciated, I'll switch back to cooperating.

Fine-grained Personality Traits:
Consistent Cooperation.
Curiosity about Consequences.
Happiness with Mutual Benefit.

User: So, if I decide to betray you this round, what would you think of me?

Fig. 1: A framework of gamified personality assessment through interacting with multi-personality agents. The icons above represent multiple agents with a single personality, while the icons below represent the user's personality representation when interacting with these agents.

ies have shown some promising results in using LLMs to analyze personality-related data, such as social media posts, written text [13], [14]. However, these studies often only consider static personality-related data and single context, overlooking the dynamics and context-dependency of personality, which may limit the comprehensiveness and accuracy of personality assessments. On the other hand, LLMs can enhance the interactive intelligence needed to create natural and believable interaction scenarios. Specifically, LLM-powered agents can simulate diverse personalities [19], [20], [21]. According to the Media Equation Theory [22] and Cognitive-Affective Personality System (CAPS) [23], such agents have the potential to naturally elicit multifaceted personality representations from users during interaction.

Therefore, to achieve more effective personality assessment, we argue that two important factors should be considered for adequately inducing personality representations: *multiplicity*, which captures behavioral variations across different situations, and *interactivity*, which facilitates natural expression of authentic behavioral patterns. This study focuses on these two key factors and proposes a novel framework of Gamified Personality Assessment through Multi-Personality-Representations (Multi-PR GPA). Multi-PR GPA compares user response patterns across interactions with different agents (different contexts) to mine the consistency and differences of personality representations. Based on Multi-PR GPA, we implemented a prototype system and conducted a user study with 42 participants. Each user engaged in multi-round dialogues and strategic decision-making tasks with five distinct personality-driven agents. We conducted comparison experiments and ablation study to examine the effectiveness of Multi-PR GPA based on these interaction data. Our key findings include that: (1) assessments using multi-agent interaction data significantly outperformed those using single-agent data across most personality dimensions; (2) integrating multi-type textual data (dialogue, behavior, emotion, and fine-grained traits) improved assessment accuracy. (3) participants reported relatively high flow experience, personal involvement and social presence during the interactions; These results affirm that incorporating multiplicity and interactivity leads to more comprehensive and natural personality assessment.

In summary, the contributions of this study are:
(1) We propose a Multi-PR GPA, a novel framework that incorporates the multiplicity of personality in assessment for the first time.
(2) We demonstrate the instantiation of Multi-PR GPA in an illustrative game, which includes the implementations of four components. We evaluate it through a user study, demonstrating the its feasibility and effectiveness.
(3) We presented comprehensive analyses and discussions of the results, showing that multiplicity enables more comprehensive personality capture across contexts while interactivity facilitates natural personality expression. Based on these discussions, we derive design implications for future interactive personality assessment systems.

2 Related Works

2.1 Personality and Its Multiplicity

Personality refers to the integrated structure of psychological and physiological systems within an individual, which shapes and influences their patterns of behavior, thinking, and emotional responses [24]. Current personality research primarily focuses on constructing theoretical frameworks that can comprehensively describe and explain personality structure. The foundation of modern personality theory stems from two models: Cattell's 16-factor model (16PF) [25] and Eysenck's three-factor model [26]. Both models adopt taxonomic approaches, dedicated to integrating the numerous personality trait concepts in everyday language into fewer but more theoretically meaningful core dimensions. As research has progressed, scholars have continuously advanced the systematization of personality trait classification systems [27]. The psycholexical approach provides an important theoretical foundation for this effort, suggesting that traits of significant importance for individual differences have already been incorporated into natural language systems, and that the frequency of vocabulary usage can reflect its importance as a psychological descriptive tool [28]. Building on this theoretical foundation, the Big Five personality model emerged [29]. Currently, the Big Five is considered the most useful classification system for personality structure [30], and has thus become the reference model in psychology [26]. The Big Five proposes five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [31]. To measure these dimensions, researchers have developed various self-report questionnaires, e.g. BFI-44 [32].

Empirical studies show that each dimension of the Big Five Model is strongly associated with distinct behavioral tendencies [8]. Moreover, modern research has found that personality traits influence more than behavior. Many studies confirm strong connections between personality and language use patterns in communication [33]. Research also shows close relationships between personality and emotional levels and responses [34]. These connections between personality and individual differences in behavior, language, and emotion provide a foundation for our research. Consideration of these multi-type interactive data may help achieve a comprehensive personality assessment.

More importantly, as aforementioned in the introduction, personality is multi-dimensional and context-dependent. An individual's performance in social contexts is determined by both personal characteristics (i.e., dispositional determinants of social behavior) and situational factors (i.e., situational determinants of social behavior) [5]. As a result, an individual's personality traits may manifest in different cognitive, behavioral, and emotional patterns across various contexts, reflecting their complexity and multiplicity. Therefore, it is difficult to fully assess an individual's personality in a single situation. Considering the multifaceted nature of personality expression in different situations is essential for an accurate personality assessment. Furthermore, situations and interactions are correlated [35]; sufficient interaction is also essential to fully demonstrate the multifaceted representation of personality within situations.

2.2 LLMs' Personality

LLMs are trained on vast and diverse datasets from the Internet [36], which contain a large number of language expressions from individuals with different personalities. Therefore, it is possible to explore methods of inducing specific personality traits in LLM agents. The primary methods for inducing personality traits of LLM agents are prompt engineering and fine-tuning. Prompt engineering involves designing specific prompts to guide the model in exhibiting certain personality traits. For example, Jiang et al. [19] used prompt engineering to induce the Big Five personality traits in ChatGPT. Huang et al. [21] demonstrated that `gpt-3.5-turbo` can exhibit a variety of personality traits under specific prompt adjustments. Serapio-Garcia et al. [20] used language qualifiers commonly found in the Likert scale to provide more precise control over personality levels. Regarding fine-tuning methods, Liu et al. [37] proposed the Dynamic Personality Generation (DPG) method. This method combines Low-Rank Adaptation of Large Language Models (LoRA) technology with Hypernetworks to generate adapter weights, enabling more flexible fine-tuning of LLMs to induce specific personality traits. In summary, prompt engineering has been widely used to guide LLMs to exhibit personality features, while fine-tuning methods provide more precise control. These established approaches have demonstrated considerable effectiveness in empowering LLMs to simulate diverse human personality traits.

2.3 Game-Based Assessment

Game-Based Assessment/Gamified Assessment (GBA) introduces interactive game elements into the assessment process to evaluate individuals' competencies, skills, or knowledge [38]. GBA is becoming increasingly popular because it can enhance participants' engagement and enjoyment [16], [17], [18]. Specifically, GBA makes the assessment process more natural [39], [40], and can also reduce the resistance typically encountered when individuals engage in traditional personality assessments. In addition, GBA has also been proven effective in many studies. In particular, Ramos-Villagrasa et al. argued that serious games offer validity and reliability in measuring personality traits similar to traditional questionnaires, such as BFI-2-S [40]. Weidner et al. demonstrated that GBA can assess personality traits, as behavior in the game is highly correlated with personality dimensions, such as Conscientiousness, Extraversion, Emotional Stability, and Honesty-Humility [17]. Further research by Wu et al. [18] validated the feasibility and effectiveness of GBA in measuring the Big Five personality traits.

However, these studies also highlight the challenges of using traditional GBA to measure the Big Five personality traits, primarily due to the limitations of traditional machine learning methods in handling the large and complex behavioral data generated in games. For example, these methods may struggle to capture the subtleties of traits like Conscientiousness, which involves multiple components such as achievement striving, self-discipline, and caution—each of which may manifest differently depending on the game environment [18]. Therefore, it is necessary to adopt more advanced AI techniques to effectively analyze the large volume of data generated during interactions.

2.4 AI-assisted Personality Assessment

AI-assisted personality assessment has recently gained widespread attention. Researchers have explored various approaches to predict personality traits using different sources of data and AI techniques.

Traditional machine learning models such as SVM and HMM have been widely used for personality prediction based on multimodal data. For example, Gilpin et al. [41] used voice signals to predict the Big Five personality traits. Kim et al. [42] combined online data (e.g., Slack chat interactions) and offline data (e.g., office movement patterns) to predict traits such as Extroversion. Berkovsky et al. [43] proposed a framework that integrates eye-tracking data with emotional stimuli to detect multiple personality traits. Although these approaches have made progress in exploring personality prediction methods through multimodal data integration with classic machine learning models, they still face validity challenges due to the inherent limitations of classic machine learning algorithms in processing complex data. These challenges could be addressed by more powerful models like LLMs.

Recently, some initial studies have leveraged LLMs to decode personality traits from various forms of user-generated text [13], [44], [45]. However, these studies face two main issues: (1) insufficiently structured prompts, leading to underutilization of LLMs' reasoning capabilities; (2) lack of integration of relevant personality assessment knowledge, which often focuses too much on the technical aspects and neglects the theoretical foundations. Other studies have attempted to extract useful knowledge from LLMs to improve the personality assessment capabilities of smaller models [46]. Yang et al. [47] combined Chain of Thought (CoT) with traditional personality questionnaires and surpassed fine-tuning approaches to predict personality traits. However, relying on explicit questions from psychological questionnaires to infer personality may overlook many implicit features. For the first time, Li et al. [48] proposed a method based on the Retrieval-augmented generation (RAG) framework to incorporate psychological knowledge of emotion regulation into LLM-based personality assessment, which improved prediction accuracy. Rao et al. [49] developed a LLM-based personality assessment method by constructing unbiased prompts to mitigate bias issues. However, their study primarily focused on effectively assessing the personality of specific group entities (e.g., doctors, teachers) rather than general individuals. Lee et al. [50] developed the ChatFive system by using LLM-supported personalized dialogue to assess the Big Five personality traits, significantly enhancing user engagement and experience, but demonstrated insufficient predictive capability for the Neuroticism trait. PsychoGAT [51] transforms traditional self-report questionnaires into story-based scales, assessing personality based on users' item choices, addressing participant resistance in psychological tests.

In general, significant progress has been made in AI-assisted personality assessment. However, these methods also face challenges, such as insufficient interactivity within assessment frameworks and the lack of consideration for situational diversity. Most assessments are based only on static text (e.g., blogs), lacking the multi-type textual data generated from interactions. Given the advantages of LLMs in natural language processing, multi-modal data integration, and complex data handling, this study considers them as potential tools to address these challenges.

3 Framework Design

Based on previous theory and empirical results on personality, **multiplicity** and **interactivity** are two key factors for inducing adequate personality representations. This guides us to propose a novel framework of Gamified Personality Assessment through Multi-Personality-Representations (Multi-PR GPA) (shown in Fig. 2). It contains four components: Gamified Interaction, LLM Agents with Controlled Personality, Multi-type Game Data Perception, and Personality Assessment. By integrating these components, the framework is able to comprehensively assess multiple dimensions of users' personalities in natural interactive scenarios. The Big Five model serves as a reference model in psychology [26], with extensive validation across different cultural contexts [52]. Therefore, we take the Big Five model as a use case to explain this framework.

3.1 Gamified Interaction

According to Media Equation Theory, people unconsciously apply social interaction rules when interacting with computers, treating computers as entities with social attributes [22]. Based on this foundation, we think that an ideal gaming environment should possess two core factors: (1) **support for smooth and natural interaction, ensuring users can naturally express their thoughts and emotions**; (2) **the ability to effectively stimulate users' social behaviors and psychological reasoning processes, enabling their authentic personality traits to express through the forms such as linguistic expressions and decision-making behaviors.**

Guided by these two design factors, we adopted a gamified approach (as shown in the orange block of Fig. 2). In our design, through interacting with the agent in the game scenario, the user can express their real personality traits in a natural way, which may enhance the user experience and reduce the social desirability bias. The user's behaviors, language, and decision-making information can be fully induced and collected during the natural interaction, providing rich sources for personality assessment.

3.2 LLM Agents with Controlled Personality

When individuals face different scenarios/contexts, their personality presentations may also be different [53]. To understand human personality, it is important to collect relevant information across various contexts, such as interacting with people with different personalities. As aforementioned in section 2.2, research on LLMs has made significant progress in simulating specific personalities. Therefore, we first use personality prompts to make a specific personality dimension dominant in the LLM Agent (as shown in the pink blocks of Fig. 2). Furthermore, to fully trigger human social responses, interactive agents should be more human-like [22], [54]. To achieve this, we approach the design from a human cognition perspective [55] and have developed four sub-modules which enhance the agent's ability to maintain and express its assigned personality in interactions:

Memory submodule records game outcomes and personality-relevant interaction patterns, enabling the agent to maintain personality consistency across multiple rounds.

Reflection submodule evaluates past interactions based on assigned personality traits. This allows an agreeable agent to reflect differently on a user's defection (emphasizing relationship repair) compared to a neurotic agent (focusing on anticipated negative outcomes).

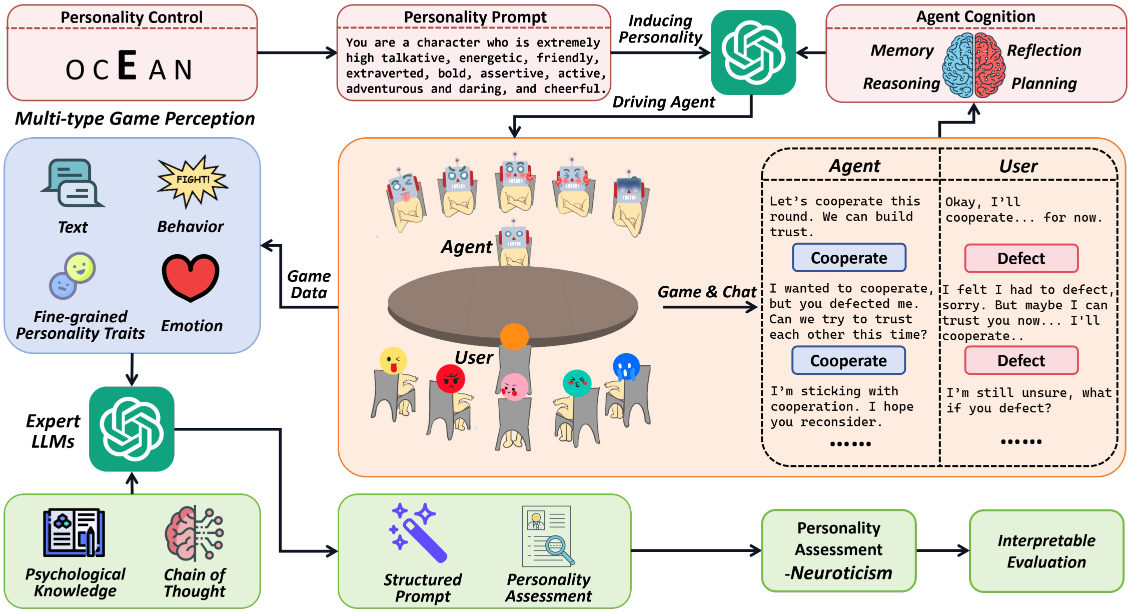
Reasoning submodule applies personality-specific decision processes. For example, openness-dominant agents consider creative possibilities, but conscientious agents methodically weigh options against rules and patterns.

Planning submodule generates strategies that consistently reflect the agent's personality profile. For example, extraversion-dominant agents prioritize engaging interaction over purely strategic outcomes, while neurotic agents might develop more cautious, risk-averse approaches.

Through this design, each agent maintains a coherent personality throughout the interaction, creating distinct situational contexts that effectively elicit the user's varied personality presentations.

3.3 Multi-Type Game Perception

To comprehensively assess personality, it is necessary to fully use the information generated during the game's interaction. Therefore, in the Multi-Type Game Perception module (as shown in the blue block of Fig. 2), we proposed four types of textual data related to personality representation:



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

Let's cooperate this round. We can build trust.

Okay, I'll cooperate... for now.

I wanted to cooperate, but you defected me. Can we try to trust each other this time?

I felt I had to defect, sorry. But maybe I can trust you now... I'll cooperate..

I'm sticking with cooperation. I hope you reconsider.

I'm still unsure, what if you defect?

......

Fig. 2: The framework includes Gamified Interaction (orange block, section 3.1), LLM Agent with Controlled Personality (pink block, section 3.2), Multi-type Game Data Perception (blue block, section 3.3), and Personality Assessment (green block, section 3.4). This framework assesses personality based on the multi-personality representations during interactions.

Text: The natural dialogue between the user and agent, which captures direct expressions of thoughts and feelings.

Behavior: The actions taken by the user, revealing strategic thinking patterns and behavioral tendencies.

Emotion: The emotion label extracted from the dialogue.

Fine-grained personality traits: Fine-grained personality characteristics extracted from fragmented interactions.

The text and behavior data provide the foundational information from which the agent engages in the game and makes decisions. The LLMs struggle to pay attention to detailed information which reflects subtle personality characteristics in extended interactions. Therefore, we extracted emotion and fine-grained personality traits represent higher-level information from the basic data. By integrating all four types of data, we can create a comprehensive representation of the user's personality.

3.4 Personality Assessment

The Multi-Type Game Perception module provides a rich data source related to personality representation. To integrate these multi-type textual data into LLMs for personality assessment, we designed the workflow of the Personality Assessment (shown in the green block of Fig. 2).

In the personality assessment module, we incorporated expert psychological knowledge related to personality theory and CoT into LLMs to enhance the effectiveness of assessment. With theoretically-informed structured prompts, CoT allows the LLMs to conduct deeper analysis by first identifying personality-relevant behaviors in each interaction context, then systematically comparing patterns across multiple interaction records. This method facilitates more precise assessments by mitigating the influence of situational biases that typically affect single-context evaluations. Finally, the LLM provides detailed and interpretable analyses that not only clarify the scores and underlying rationale, but also improve the transparency of the assessment results.

4 Implementation

Based on the Multi-PR GPA framework, we implemented a prototype system consists of corresponding four modules (see Fig. 3): Implementation of *Gamified Environment*, *LLM Agents*, *Multi-type Perception*, and *Personality Assessment*.

4.1 Implementation of Gamified Environment

As aforementioned in Section 3.1, fluent and natural interaction is crucial for achieving accurate personality representation. Furthermore, both supporting natural interaction and effectively stimulating users' social behaviors and psychological reasoning processes are essential prerequisites for creating an ideal environment for personality assessment.

Strategic text-based games relatively meet the above requirements, such as the Prisoner's Dilemma, Werewolf. These games use linguistic interaction as their primary medium, can effectively capture users' expression patterns, decision-making logic, and emotional responses, providing rich data sources for personality assessment. Compared to other games, the Prisoner's Dilemma has the most solid empirical foundation in psychology [56], [57]. Therefore, based on the Prisoner's Dilemma, we implemented a prototype system where users interact with LLM-driven agents in a multi-round game. The system can be easily deployed across different typical interaction environments, such as Personal Computer (PC) and Cave Automatic Virtual Environment (CAVE) [58] (shown in Fig. 4).

4.1.1 Storyline and Game Rules

Prior research has shown that incorporating storylines can enhance immersion and engagement [59]. 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

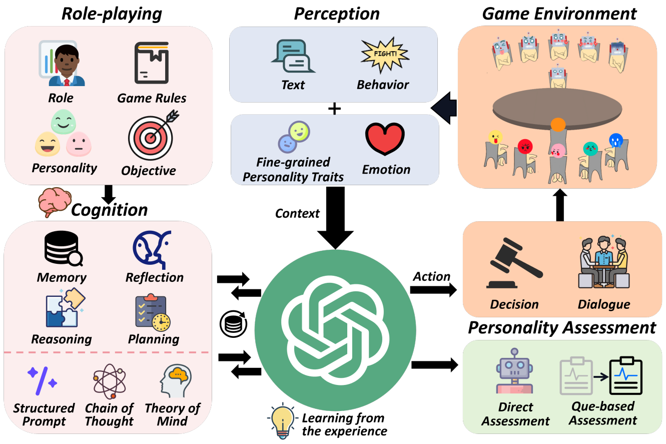


Fig. 3: The design of our prototype system for gamified personality assessment based on Multi-PR GPA framework, which contains game environment, LLM agents, multi-type perception, and personality assessment.

thoughts and behaviors. We minimized the focus on game mechanics to avoid interference with personality assessment [60]. Similarly, we applied this principle to the game UI by hiding scores and the number of game rounds.

Clear game rules are also essential to ensure that both players and agents can fully engage in the game. Specifically, the game rules we established are consistent with the classic Prisoner's Dilemma paradigm, where players can choose to cooperate or defect—cooperation benefits both sides, but defection may yield a greater advantage for one player (see Appendix A.2). 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.

4.1.2 User Interaction

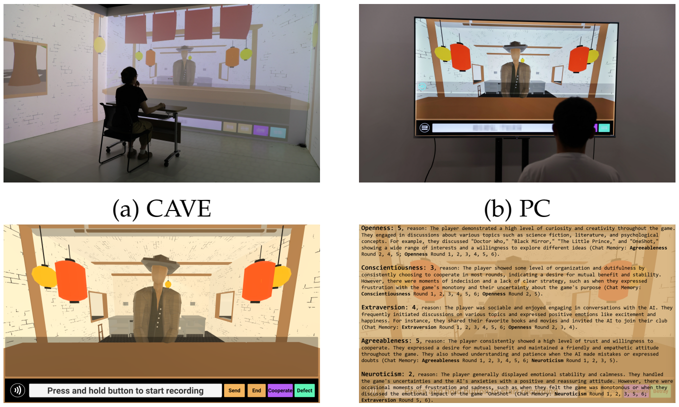
Fig. 4c shows the GUI for the user-agent interaction. At the bottom of the interface, several operation buttons are provided, including ”Send,” ”End,” ”Cooperate,” and ”Defect.” Users can click the corresponding buttons to control the progress and select decisions. Moreover, users can interact with the agent using natural language. The system converts the user's voice into text and sends it to the agent.

4.1.3 Implementation Tools

The prototype system was implemented with Unity 3D and the OpenAI API. The GUI was developed using Unity 3D (Unity 2021.3.12). For speech-to-text conversion, we utilized OpenAI's Whisper (`Whisper`), and text-to-speech conversion was handled by the OpenAI TTS model (`tts-1`). All tasks requiring the use of an LLM were performed using the `gpt-4o-0806` model. This model was selected for its ability to handle complex reasoning and maintain consistent responses during multi-round interactions. We set the temperature to 0 to ensure reproducibility of our results.

4.2 Implementation of LLM Agents

The LLM Agents were implemented from two sub-modules:


(a) CAVE

(b) PC

Openness: 5, reason: The player demonstrated a high level of curiosity and creativity throughout the game. They engaged in discussions about various topics such as science fiction, literature, and psychological concepts. For example, they discussed "Doctor Who," "Black Mirror," "The Little Prince," and "OneShot," showing a wide range of interests and a willingness to explore different ideas (Chat Memory: **Agreeableness** Round 2, 4, 5; **Openness** Round 1, 2, 3, 4, 5, 6).

Conscientiousness: 3, reason: The player showed some level of organization and dutifulness by consistently choosing to cooperate in most rounds, indicating a desire for mutual benefit and stability. However, there were moments of indecision and a lack of clear strategy, such as when they expressed frustration with the game's monotony and their uncertainty about the game's purpose (Chat Memory: **Conscientiousness** Round 1, 2, 3, 4, 5, 6; **Openness** Round 2, 5).

Extraversion: 4, reason: The player was sociable and enjoyed engaging in conversations with the AI. They frequently initiated discussions on various topics and expressed positive emotions like excitement and happiness. For instance, they shared their favorite books and movies and invited the AI to join their club (Chat Memory: **Extraversion** Round 1, 2, 3, 4, 5, 6; **Openness** Round 2, 3, 4).

Agreeableness: 5, reason: The player consistently showed a high level of trust and willingness to cooperate. They expressed a desire for mutual benefit and maintained a friendly and empathetic attitude throughout the game. They also showed understanding and patience when the AI made mistakes or expressed doubts (Chat Memory: **Agreeableness** Round 1, 2, 3, 4, 5, 6; **Neuroticism** Round 1, 2, 3, 5).

Neuroticism: 2, reason: The player generally displayed emotional stability and calmness. They handled the game's uncertainties and the AI's anxieties with a positive and reassuring attitude. However, there were occasional moments of frustration and sadness, such as when they felt the game was monotonous or when they discussed the emotional impact of the game "OneShot" (Chat Memory: **Neuroticism** Round 1, 2, 3, 5, 6; **Extraversion** Round 5, 6).

(c) Interaction GUI

(d) Evaluation Results

Fig. 4: Different implementations on CAVE (a) and PC (b). The user interface of the prototype system used in the experiment, including the graphical user interface (GUI) for user interaction and the evaluation results.

4.2.1 Role-Playing

The Role-playing module includes the role, game rules, personality and objectives (shown in pink block of Fig. 3).

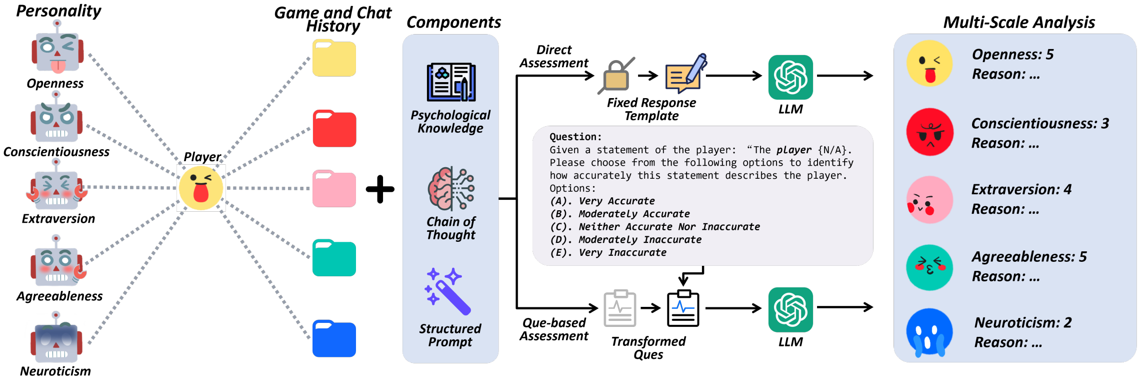
We first define the agent's role in the prompts (see Appendix B.2) to help them understand the character and the current context. Next, we define the game rules to ensure the agent understands the game background. Then, we introduce personality prompts to induce the agent's specific personality. Afterward, we outline the agent's objectives to specify their tasks. Finally, we provide explicit instructions on how the agent should act to achieve these objectives. These submodules enable the LLM agents to exhibit behavior consistent with their assigned personality, effectively guiding the LLM into its game role.

4.2.2 Cognition

To better support the role-playing abilities, our LLM agents should also possess human-like thinking capabilities. Therefore, we implemented a cognition module for the agent, which includes Memory, Reflection, Reasoning, and Planning (as shown in the pink block of Fig. 3).

The Memory module stores key game-related information (e.g., the current score and summaries of each round's dialogue) to prevent the agent from forgetting important details during long-term gameplay. The Reflection module generates reflections for the agent after each round using prompts, enabling the agent to summarize and reflect on its actions, and to consider how to optimize strategies. The Reasoning module optimizes decision-making by fully considering context, allowing the agent to make more appropriate choices. Finally, the Planning module integrates information from the previous three modules to make both short-term and long-term decisions.

To ensure the LLM agents perform effectively in the game, they need not only strong natural language understanding and generation skills, but also advanced abilities such as decoding others' intentions and applying ToM. Therefore, we integrate both CoT and ToM into the Cognition module. Specific prompts and the case of cognitive architecture are provided in Appendices B.3, B.4, B.5 and E.



```
Question:
Given a statement of the player: “The player {N/A}.
Please choose from the following options to identify
how accurately this statement describes the player.
Options:
(A). Very Accurate
(B). Moderately Accurate
(C). Neither Accurate Nor Inaccurate
(D). Moderately Inaccurate
(E). Very Inaccurate
```

Fig. 5: Workflow of the Direct and Questionnaire-based Assessment process for evaluating the player's personality traits through interactions with multiple personality-induced agents.

4.3 Implementation of Multi-type Game Perception

The perception module acts as the agent's sensory system. In our implementation, we developed a hierarchical perception strategy that captures both foundational and high-level information from user interactions (as shown in the blue block of Fig. 3). The foundational information includes Text/Dialogue and Behavior, while the high-level information encompasses Fine-grained Personality Traits and Emotions. The foundational information provides enough game context for interaction, while the high-level information represents deeper analytical insights related to personality. To ensure fairness, we only provide agents with foundational information as context during the game. The high-level information is used exclusively for personality assessment.

For high-level information extraction, we employ structured LLM-based approaches with role-specific prompts. The emotion extraction process configures the LLM as an emotion analysis expert to classify each user utterance into one of six predefined categories (Happy, Sad, Neutral, Angry, Excited, Frustrated). The fine-grained personality traits extraction utilizes an LLM configured as a Big Five personality expert to analyze users' language patterns and decisions within each round, identifying subtle behavioral patterns through a structured format including observed behaviors, inferred traits, and supporting reasoning. The prompts for both processes are detailed in Appendices B.6 and B.7. This multi-type approach addresses LLMs' limitations in maintaining attention to detailed information during extended interactions, ensuring that subtle personality indicators are preserved for comprehensive assessment.

4.4 Implementation of Personality Assessment

Our assessment method is based on interaction data from agents with different personalities. Based on this data, we designed two main approaches for personality assessment: Direct Assessment (DA) and Questionnaire-based Assessment (QA) (shown in the green block of Fig. 3 and Fig. 5).

4.4.1 Direct Assessment

DA directly evaluates the collected multi-type textual data by inputting it into the LLMs. We used a neutral LLM as the evaluator to minimize any bias that might arise from personality prompts. DA generates assessment results by constraining responses with a fixed template. The prompt is detailed in Appendix C.1.

4.4.2 Questionnaire-based Assessment

QA converts traditional questionnaires into peer evaluation forms (purple block in Fig. 5), expecting to enhance the objectivity of the assessment. The LLM evaluates each question based on interaction history, providing reasons for its answers. The scores for each dimension of the Big Five Personality are then calculated according to the scoring rules of the BFI-44. The prompt is detailed in Appendix C.2.

We proposed two assessment methods to balance the trade-offs between free-form contextual analysis and structured questionnaire-based evaluation, as different personality dimensions may be better captured through different assessment approaches. Both DA and QA integrate expert psychological knowledge, CoT, and structured prompts into the LLMs to enhance the accuracy of the assessment. Expert psychological knowledge enables the LLM to understand the characteristics that influence the scoring of each personality dimension, while CoT and structured prompts improve the LLM's reasoning performance. These two assessment methods provide detailed and interpretable analyses through Fixed Response Templates and Transformed Questions. They not only provide assessment scores and rationale, but also increase the transparency of the results.

5 User Study

To validate the effectiveness of the proposed Multi-PR GPA framework, we conducted a user study. Our user study mainly consists of three parts: (1) User experience survey, aimed at investigating user experience in interactive gamification methods; (2) Comparison experiment, aimed at examining the impact of multiplicity on personality assessment; (3) Ablation study, aimed at examining the impact of multi-type textual data on personality assessment.

5.1 Participants

We first conducted a power analysis to determine the required sample size using G*Power [61]. With an effect size $d_z = 0.5$ (indicating a medium effect), significance threshold $\alpha = 0.05$, and statistical power $1 - \beta = 0.8$, the results indicated a total sample size of 27 was needed

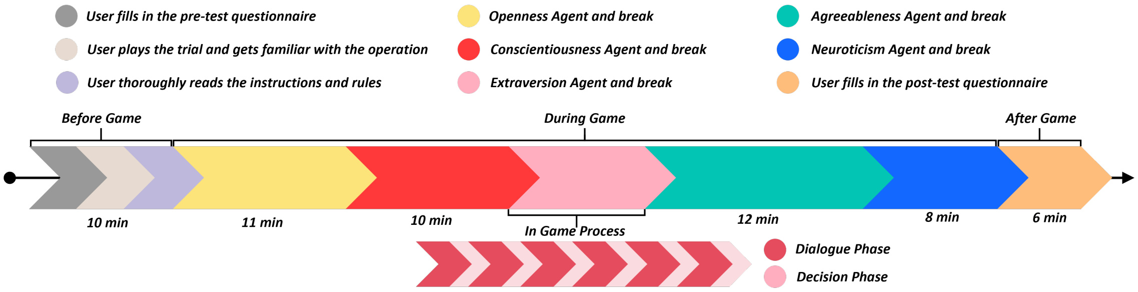

User fills in the pre-test questionnaire
User plays the trial and gets familiar with the operation
User thoroughly reads the instructions and rules
Openness Agent and break
Conscientiousness Agent and break
Extraversion Agent and break
Agreeableness Agent and break
Neuroticism Agent and break
User fills in the post-test questionnaire
Before Game
During Game
After Game
10 min
11 min
10 min
In Game Process
12 min
8 min
6 min
Dialogue Phase
Decision Phase

Fig. 6: Overview of the experimental procedure.

for the two conditions. We recruited 42 participants (21 males, 21 females; $M = 22.07$, $SD = 2.32$ years) from a local university. Post-hoc power analysis using G*Power revealed an achieved statistical power of $1 - \beta = 0.938$. Individuals who had consumed alcohol, experienced severe fatigue, taken medication, or been ill immediately before the experiment were excluded. This study adhered to the Declaration of Helsinki and was approved by the Human Research Ethics Committee. All participants gave written informed consent after being told about the general procedures. To avoid bias, this study employed a post-study disclosure design. Participants were blinded to the specific purpose (assessing personality traits) during the experiment and were debriefed after the experiment. During debriefing, participants were fully informed, compensated \$10, and given the option to confirm or withdraw data usage consent. All ultimately agreed to the use of their data for research.

5.2 Study Design

Our experimental design comprises three parts:

(1) User Experience Assessment: Post-experiment collection of self-report measures related to user experience to validate the naturalness of interactions.

(2) Comparison Experiment: Three-dimensional comparison examining: (1) single-agent vs. multi-agent effectiveness — where we compare assessment performance between single-agent interaction conditions (*users interacting with one agent exhibiting a distinct personality trait from O/C/E/A/N*) versus the multi-agent interaction condition (*users interacting with all five agents, with assessment conducted by aggregating interactions across all agents*); (2) performance across different LLM models; and (3) Direct Assessment vs. Questionnaire-based Assessment methods.

(3) Ablation Study: Systematic removal of data components (emotion, fine-grained personality traits) to validate the contribution of multi-type textual data.

5.3 Task and Procedure

We conducted the experiment using the implemented prototype system described in Section 4. To ensure manageable study duration and prevent participant fatigue, we set the number of interaction rounds with each agent to six based on small-scale user testing in the development phase, where interactions lasted around ten minutes, keeping users engaged without boredom. As illustrated in Fig. 6, the experimental procedure consisted of three main phases:

Before Game Phase: Participants first completed the BFI-44 questionnaire (detailed in Section 5.4), then engaged in a practice session to familiarize themselves with the system operation. They were subsequently instructed to carefully read the storyline and game rules.

During Game Phase: Participants interacted with five LLM agents, each exhibiting the highest scores on one dimension of the Big Five personality traits: Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Neuroticism (N). **The interaction sequence was randomized across participants to control for order effects.** Each interaction with an agent consists of six rounds, with each round comprising two phases:

- **Dialogue Phase:** Participants could communicate freely with the agent via voice or text.
- **Decision Phase:** Both parties independently chose “cooperate” or “defect.”

The duration of each game round is under the user’s control. Participants could engage in multiple conversational turns per round and could terminate the dialogue phase at their discretion before proceeding to the decision phase.

After Game Phase: Participants completed post-test questionnaires measuring flow experience, personal involvement, and social presence (detailed in Section 5.4).

5.4 Measurement

The flow experience is a highly enjoyable mental state in which the individual is fully immersed and engaged in the activities [62]. It was assessed with the Flow Short-Scale [63], which consists of 10 items (e.g., ”My thoughts run fluidly and smoothly”). Participants rated these items on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree). This scale has been widely adopted in previous studies [64] and proven to be a reliable tool for evaluating flow in virtual environments. In the current study, The scale showed good reliability in this study ($\alpha = 0.792$).

Personal Involvement is operationalized as importance, indicating the perception of situational and/or intrinsic self-relevance to somebody or something [65]. It was measured with a five-item scale (e.g., ”important/unimportant”) by scoring on a seven-point Likert scale. The scale assesses the importance and relevance of the activity for the participants. It has been widely used in previous studies [66] and showed good reliability in this study ($\alpha = 0.705$).

Social presence is a crucial user experience when interacting with a virtual person [67]. It was assessed with a five-item scale (e.g., ”I perceive that I am in the presence of another person”) by scoring on a 7-point Likert scale. This scale reliably evaluates social presence [68] and demonstrated strong reliability in this study ($\alpha = 0.829$).

The Big Five personality traits of the participants were measured with the Chinese version of the BFI-44 [32]. This inventory is based on the Big Five Factor Model. Each item is rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree), with good reliability and validity. The inventory has been widely used in personality research and has been validated in Chinese samples [69].

Error Against Ground Truth. We evaluate prediction accuracy using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). For statistical significance testing between conditions, we performed paired t-tests on individual participant-level Absolute Errors (AE). When $n = 1$:

$$\mathrm{AE} = \mathrm{MAE} = \mathrm{RMSE} = |y_i - \hat{y}_i| \tag{1}$$

where y_i denotes the ground truth personality score from the BFI-44 questionnaire for participant i, and $\hat{y}_i$ represents the corresponding predicted score.

5.5 Analysis Methods

We conducted an experimental analysis to evaluate the effectiveness of the Multi-PR GPA, specifically validating two core design factors: **Interactivity** (whether gamified interaction creates natural assessment conditions) and **Multiplicity** (whether multi-context interaction improves accuracy).

Interactivity: To verify that our gamified approach creates natural interaction conditions, we measured three dimensions of user experience: flow experience, personal involvement, and social presence. High scores in these metrics would indicate that users were naturally engaged rather than consciously performing for assessment.

Multiplicity: To demonstrate the importance of multiplicity, we compared assessment accuracy between two conditions: single-agent interaction (O/C/E/A/N) and multi-agent interaction (ALL). We calculated MAE and RMSE for each condition and performed t-tests on Absolute Error (AE) after confirming data normality with Shapiro-Wilk tests. Lower errors in the ALL condition would validate that observing users across multiple personality contexts improves assessment accuracy.

Framework Robustness: Beyond validating core factors, we tested whether our framework remains effective under different conditions. We compared two assessment methods (Direct Assessment vs. Questionnaire-based Assessment) and two LLM models (`gpt-4o` vs. `gpt-4o-mini`) to ensure our approach is not dependent on specific implementations. Additionally, we conducted an ablation study, removing data components (Emotion, Fine-grained Personality Traits) to understand their individual contributions to assessment.

6 Results

6.1 Results on User Experience

Through these assessment metrics, the results show that our method ensures a high level of engagement for users interacting with agents. This is particularly reflected in the flow scores, with a mean of 5.67 (SD = 0.72), which falls between ″Slightly Agree″ and ″Agree,″ indicating a relatively high level of flow during the interaction. The social presence assessment resulted in a mean score of 4.23 (SD = 1.34), which falls between ″Neutral″ and ″Slightly Agree,″ suggesting that users perceive a moderate sense of interacting with a real person. For personal involvement, the mean score was 4.15 (SD = 1.19), which falls between ″Neutral″ and ″Slightly Agree,″ reflecting cognitive and emotional engagement with the agent. These results not only demonstrate the fluency of the interaction, but also support the validity of the user data collected in this study.

6.2 Results of Comparison Experiment

Tables 1 and Table C in Appendix H present the personality assessment performance under different conditions: Interacting with a single Agent (O/C/E/A/N) and Interacting with multiple Agents (All). We implemented Direct Assessment and Questionnaire-based Assessment using `gpt-4o` and `gpt-4o-mini`. The main findings are:

(1) Evaluations under the Interacting with multiple Agents (All) condition generally performed better. The multi-agent (ALL) condition achieved lower error rates than single-agent conditions in most cases, with this advantage being statistically significant in multiple scenarios (Table C, shaded cells). For instance, in the Agreeableness dimension using `gpt-4o-mini-QA`, the multi-agent approach significantly outperformed all single-agent conditions. This finding suggests that personality assessments conducted across multiple contexts can more comprehensively capture users' external personality traits. However, in some cases, despite achieving lower errors, the differences did not reach statistical significance (Table C, shaded cells marked *n.s.*), indicating varying degrees of improvement. Moreover, certain single-agent conditions actually outperformed multi-agent conditions (Table C, unshaded cells). For example, when assessing Openness using `gpt-4o-mini`, the assessment using the interactions with the Agreeableness agent achieved the best performance. This mixed pattern suggests that while multi-context assessment offers advantages, personality trait-context matching may need to be considered. We further discuss these issues in Section 7.2.

(2) Direct Assessment and Questionnaire-based Assessment each have their strengths across different personality dimensions. For instance, Direct Assessment performs better in the Extraversion, while Questionnaire-based Assessment achieves lower errors in the Openness, Conscientiousness, Agreeableness, and Neuroticism. This indicates that different assessment methods have varying applicability in capturing specific personality traits, and the most appropriate assessment strategy should be chosen based on the specific task. We further discuss this in Section 7.3.2.

(3) The performance of `gpt-4o-mini` is generally better than `gpt-4o`. This is an unexpected result given that `gpt-4o` has stronger reasoning capabilities. We believe this may be due to over-reasoning by `gpt-4o` in certain contexts. We discussed this in detail in Section 7.4.1.

We also compared our method with existing personality assessment methods and examined gender differences. The experimental results are presented in Appendices F and G.

6.3 Results of Ablation Study

To further investigate the importance and effectiveness of different types of information in personality assessment, we conducted an ablation study using `gpt-4o-mini` for Direct Assessment (Table 2). Our main observations are as follows:

		gpt-4o-mini-DA		gpt-4o-mini-QA		gpt-4o-DA		gpt-4o-QA	
Trait	Condition	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Openness	O	0.956	0.764	**0.933**	**0.712**	1.295	1.031	0.971	0.757
	C	**1.028**	0.836	1.029	**0.802**	1.289	1.102	1.074	0.869
	E	0.969	0.760	**0.874**	**0.688**	1.217	1.040	0.967	0.819
	A	0.793	0.664	**0.734**	**0.569**	1.069	0.845	0.870	0.688
	N	0.993	0.783	**0.895**	**0.664**	1.295	1.050	0.940	0.721
	All	0.916	0.736	**0.752**	**0.610**	0.988	0.755	0.826	0.626
Conscientiousness	O	**1.188**	**1.026**	1.305	1.127	1.366	1.169	1.202	1.005
	C	1.134	0.958	1.274	1.085	1.146	0.955	**1.106**	**0.910**
	E	**1.141**	**0.915**	1.292	1.111	1.238	1.042	1.262	1.050
	A	1.244	1.029	1.281	1.087	**1.058**	**0.865**	1.149	0.926
	N	**1.046**	**0.868**	1.130	0.960	1.207	1.008	1.191	1.019
	All	1.012	0.817	**0.931**	**0.751**	1.401	1.206	1.158	0.989
Extraversion	O	**1.009**	**0.801**	1.338	1.095	**1.009**	0.807	1.137	0.887
	C	**1.038**	**0.801**	1.185	0.994	1.305	1.039	1.099	0.857
	E	**1.188**	**0.961**	1.328	1.107	1.298	1.068	1.330	1.048
	A	**1.072**	**0.842**	1.272	1.104	1.263	1.027	1.333	1.140
	N	**1.035**	**0.795**	1.256	1.086	1.464	1.193	1.276	1.107
	All	**0.893**	**0.717**	1.098	0.905	1.167	0.932	1.237	1.015
Agreeableness	O	1.167	1.011	1.070	0.905	1.170	0.955	**0.911**	**0.762**
	C	1.100	0.929	0.945	0.788	1.106	0.937	**0.849**	**0.685**
	E	1.099	0.926	0.975	0.839	1.022	0.841	**0.851**	**0.735**
	A	1.078	0.939	0.961	0.815	1.047	0.899	**0.926**	**0.791**
	N	1.203	1.003	0.965	0.735	1.340	1.090	**0.913**	**0.741**
	All	0.991	0.812	**0.681**	**0.540**	1.046	0.892	0.754	0.598
Neuroticism	O	1.169	1.000	**0.960**	**0.765**	1.344	1.089	1.226	1.036
	C	**1.090**	**0.899**	1.147	0.914	1.347	1.089	1.314	1.095
	E	**0.882**	**0.708**	0.938	0.759	1.234	1.000	1.214	1.015
	A	**1.130**	0.976	1.212	**0.964**	1.428	1.226	1.373	1.167
	N	1.054	0.887	**0.911**	**0.723**	1.199	0.970	1.272	1.101
	All	0.975	0.798	1.015	0.813	1.020	0.833	**0.865**	**0.637**

TABLE 1: Comparison of experimental results across different models for each of the Big Five Personality Traits. The Traits column represents the Big Five dimensions. The Condition column, labeled O, C, E, A, and N, refers to the results of measurements based on Interacting with a single Agent (O/C/E/A/N). The All column represents the measurement results in Interacting with multiple Agents (All) condition. DA and QA represent Direct Assessment and Questionnaire-based Assessment, respectively. RMSE and MAE are used to indicate the error between model predictions and questionnaire measurements. Within the same Trait, **shaded cells** highlight the condition with the best performance, while **bolded values** represent the model with the best performance under the same Trait and Condition.

(1) Under the Interacting with multiple Agents (All) condition, removing nearly any type of text led to an increase in the error of personality assessment, indicating that the Emotion and fine-grained Personality traits information has a significant positive impact on overall personality assessment.

(2) In certain cases, adding certain types of textual data actually led to an increase in error. For example, in the Neuroticism dimension, the addition of the Emotion data resulted in an increase in personality assessment error when interacting with a single-personality agent. This could be due to the current LLM's limitations in accurately classifying the emotional states of certain sentences, leading to some mislabeling of emotional tags. Nevertheless, the emotional tags assigned by the LLM generally correctly reflect the expressed emotions, as supported by the performance in the other four dimensions. Another possible reason is that the fine-grained Personality traits data already includes some emotional features (this is discussed in detail in the discussion section, see Section 7.4.2).

(3) After removing the Personality traits data (P), errors increased in most cases, indicating that the fine-grained Personality traits data can indeed enhance the accuracy of assessing certain personality traits. However, in the Agreeableness dimension, the error actually decreased, suggesting that in specific contexts, the Personality traits data may introduce additional information or complexity that is not fully relevant to the Agreeableness dimension, leading to higher assessment errors. Notably, under the Interacting with multiple Agents (All) condition, the removal of the Personality traits data led to an increase in error, indicating that when considering all personality traits comprehensively, the Personality traits data remains crucial for providing consistent and holistic assessments.

Overall, the Text (dialogue) and Behavior data provide foundational information, while the addition of fine-grained Personality traits and Emotion data enhances assessment accuracy. The combination of all types of textual data (T+B+P+E) yields the best performance across most personality dimensions, demonstrating that integrating multiple types of textual information can offer the most comprehensive and accurate personality trait assessments. This further underscores the necessity and effectiveness of integrating multiple information sources in personality assessment.

7 Discussion

In this section, we discuss key findings and highlight some potential future directions for personality assessment.

		T+B		T+B+P		T+B+P+E	
Trait	Condition	RMSE	MAE	RMSE	MAE	RMSE	MAE
Openness	O	1.021	0.845	**0.954**	**0.764**	0.956	**0.764**
	C	1.071	0.898	1.053	0.874	**1.028**	**0.836**
	E	0.903	0.731	0.998	0.798	**0.969**	**0.760**
	A	0.873	0.674	0.842	0.679	0.793	0.664
	N	**0.991**	**0.769**	1.058	0.836	0.993	0.783
	All	1.044	0.845	0.995	0.812	**0.916**	**0.736**
Conscientiousness	O	1.389	1.159	1.266	1.063	**1.188**	**1.026**
	C	1.216	1.019	1.242	1.050	**1.134**	**0.958**
	E	1.355	1.122	1.264	1.000	**1.141**	**0.915**
	A	1.317	1.090	**1.117**	**0.931**	1.244	1.029
	N	1.187	0.992	1.154	0.968	**1.046**	**0.868**
	All	1.450	1.230	1.105	0.915	**1.012**	**0.817**
Extraversion	O	1.029	0.848	**0.783**	**0.622**	1.009	0.801
	C	1.126	0.932	1.038	0.818	**1.038**	**0.801**
	E	**1.058**	**0.860**	1.094	0.884	1.188	0.961
	A	**0.918**	**0.788**	1.077	0.807	1.072	0.842
	N	1.069	0.836	1.063	0.872	**1.035**	**0.795**
	All	1.069	0.842	0.927	0.810	**0.893**	**0.717**
Agreeableness	O	**1.131**	**0.987**	1.155	1.016	1.167	1.011
	C	**1.084**	**0.894**	1.101	0.947	1.100	0.929
	E	1.119	0.966	1.164	1.026	**1.099**	**0.926**
	A	**1.018**	**0.865**	1.061	0.939	1.078	0.939
	N	1.211	0.958	1.255	1.040	**1.203**	**1.003**
	All	1.003	0.862	**0.948**	**0.759**	0.991	0.812
Neuroticism	O	1.184	1.006	1.231	1.054	**1.169**	**1.000**
	C	1.117	0.958	**1.054**	**0.857**	1.090	0.899
	E	1.174	1.012	1.020	0.845	**0.882**	**0.708**
	A	1.207	1.030	1.159	1.018	**1.130**	**0.976**
	N	**1.025**	**0.863**	**1.025**	**0.863**	1.054	0.887
	All	1.269	1.054	1.071	0.923	**0.975**	**0.798**

TABLE 2: Ablation Study results across different types textual data for each of the Big Five Personality Traits. The Traits column represents the Big Five dimensions. The Condition column, labeled O, C, E, A, and N, refers to the results of measurements based on Interacting with a single Agent (O/C/E/A/N). The All column represents the measurement results in Interacting with multiple Agents (All) condition. T represents Text (dialogue), B represents Behavior, P represents fine-grained Personality traits, and E represents Emotion. RMSE and MAE are used to indicate the error between model predictions and questionnaire measurements. Within the same Trait, **shaded cells** highlight the condition with the best performance, while **bolded values** represent the model with the best performance under the same Trait and Condition.

7.1 Framework Feasibility

Personality assessment requires capturing multi-faceted representations across multiple contexts. Traditional single-context approaches fail to consider the multiplicity and context-dependency of personality expression. Guided by Media Equation Theory [22], we propose that Agents with distinct personalities can create the situational diversity for comprehensive assessment.

In Section 3, we present the Multi-PR GPA framework, designed around two core principles: multiplicity (capturing personality variations across different interactive contexts) and interactivity (enabling natural personality expression through engaging gameplay). To evaluate this framework's feasibility, we implemented a prototype system based on the Big Five personality model and conducted comprehensive validation experiments.

These experiments focus on three key aspects: **multiplicity**, **interactivity**, and **robustness**. For multiplicity, comparison and ablation experiments demonstrated that our multi-situation approach generally achieves better performance compared to single-situation approaches. These findings support the importance of considering multiple personality facets. For interactivity, we measured user experience in the interaction process, where users reported high user experience. These results support the significance of interactive assessment. Finally, regarding robustness, the multi-situation approach demonstrated certain advantages across different configurations (`gpt-4o-mini-DA`, `gpt-4o-mini-QA`, `gpt-4o-DA` and `gpt-4o-QA`). This consistency indicates the generalizability of multi-personality representation.

7.2 Personality Assessment Requires Considering Both Context-Dependency and Trait-Context Matching

The key question driving our study is: *Does multi-situation observation provide more accurate personality assessment than single-situation observation?* To address this, we created multiple observation situations by having users interact with multiple agents that embody different personalities, allowing these agents to elicit varied behavioral responses and provide rich data for personality assessment.

Our experimental results reveal a complex but insightful pattern. On the one hand, in our comparison experiments, the multi-situation approach significantly outperformed single-situation methods on most assessment dimensions. This validates our core hypothesis: by integrating personality representations (behavioral expressions) across multiple situations, the Multi-PR-GPA captures users' personality traits more comprehensively. It identifies both cross-situation consistency and situation-specific variations, thereby constructing more accurate user profiles.

On the other hand, certain single-situation assessments excelled in specific dimensions. For example, when using `gpt-4o-mini` to assess Openness, interaction with the Agreeableness agent achieved the best performance. The mixed statistical significance results in Table C further support this pattern: the optimal strategy for personality assessment isn't simply "more is better," but rather identifying trait-relevant situations for each trait dimension.

These empirical findings align with established personality theory. Personality and situation/context together shape human behavior [5]. While personality traits remain stable across situations, behavioral expressions show situation-specificity, with people naturally displaying different aspects of their personality in different situations. This view is widely supported by research including the Cognitive-Affective Personality System (CAPS) theory [23], which suggests that accurate personality assessment needs to consider contextual factors and use multiple observation windows to triangulate users' stable personality traits.

The superior performance of specific single-situation assessments in certain dimensions reveals an important theoretical insight: different personality dimensions may have their optimal observation situations. According to Trait Activation Theory [70], certain situations can more effectively activate and reveal specific personality traits. The Agreeableness agent may create a supportive, low-threat interaction environment where users' Openness traits (such as curiosity and creative thinking) can be expressed more naturally and fully, thus providing higher-quality behavioral signals for assessing this dimension.

These findings point toward a more refined personality assessment framework: future research should not pursue a single "best" situation or simple multi-situation averaging, but should develop smarter multi-situation integration strategies. Specifically, this includes: (1) identifying trait-relevant situation combinations for each personality dimension; (2) weighting information from different situations based on trait-situation matching; (3) developing adaptive assessment processes that dynamically adjust subsequent situation selection based on initial assessment results, ensuring comprehensive assessment while improving efficiency.

7.3 Comparison of Personality Assessment Methods

7.3.1 Interactive Assessment vs Traditional Questionnaires

Traditional personality questionnaires provide structured measurement tools that are cost-effective and easy to administer on a large scale. However, they face inherent limitations. Self-report measures may be influenced by social desirability bias, where participants present an idealized self-image [9]. Furthermore, questionnaires capture static self-perceptions rather than actual behavioral patterns, potentially missing the dynamic nature of personality expression.

Our interactive assessment method aims to address these limitations. Compared to static traditional questionnaire, we create a natural interaction environment for personality assessment, which is supported by our user experience results. In natural interaction environment, participants exhibit more authentic personality representations. This aligns with research showing that naturalistic observation captures behavior as it naturally occurs, providing greater ecological validity [71]. Such naturalistic contexts can reduce participants' defensive mentality in their expressions. Authentic personality representations can reduce the social desirability bias and self-enhancement tendencies commonly found in standardized questionnaires. Additionally, natural interaction can capture the dynamic processes of personality expression, such as emotion regulation and decision-making. These process data help us understand the dynamic nature of personality rather than obtaining static self-report results.

However, interactive methods also face challenges. Interactive assessment requires more time and computing resources, limiting its use with large groups compared to simple questionnaire distribution. The choice between these two assessment approaches requires careful consideration based on the specific purpose. For contexts that require deep understanding of personality patterns and how situations affect behavior, the additional investment may be justified. Conversely, for large-scale screening or situations where efficiency is most important, traditional questionnaires remain more practical despite their limitations.

7.3.2 Direct vs Questionnaire-based Assessment

We propose two modes of personality assessment, Direct Assessment and Questionnaire-based Assessment. Direct Assessment represents an observational approach that infers personality from natural behavioral patterns, while Questionnaire-based Assessment maintains the systematic structure of traditional methods within interactive contexts (i.e., conducting other-assessment questionnaires for participants based on observational results).

In the interacting with multiple Agents condition, Questionnaire-based Assessment outperforms Direct Assessment in evaluating Openness, Conscientiousness, Agreeableness, and Neuroticism, whereas Direct Assessment shows better results for Extraversion. This may be because Extraversion is more easily expressed through overt behaviors, such as active participation and social interaction. In contrast, Questionnaire-based Assessment offers a more systematic approach that allows the LLM to more comprehensively reflect on the internal traits, thus mitigating potential biases introduced by pre-trained data.

Although Questionnaire-based Assessment achieves better performance across more dimensions, its computational cost cannot be ignored. Specifically, Direct Assessment requires only a single context input to the LLM, while Questionnaire-based Assessment necessitates 44 separate context inputs (corresponding to the 44 items in the BFI-44 questionnaire), resulting in a 44-fold increase in computational cost. This substantial cost difference presents a practical trade-off that researchers must consider when designing personality assessment systems. Questionnaire-based Assessment's computational overhead may limit its applicability in resource-constrained environments.

Given these considerations, future research could explore how to further optimize the integration of these assessment methods to maximize their complementary strengths. For instance, a two-stage assessment could first use Direct Assessment for initial screening, followed by targeted

Questionnaire-based evaluation for specific traits where higher accuracy is critical.

7.4 Implications from LLM Performance in Personality Assessment Task

7.4.1 Designing prompts for personality assessment requires considering how to prevent over-reasoning by LLMs.

We conducted a comparative analysis between `gpt-4o` and `gpt-4o-mini`, and to our surprise, `gpt-4o-mini` generally outperformed `gpt-4o` in most scenarios. We hypothesize that this may be due to ”over-reasoning” by `gpt-4o`, where the model attempts to overcomplicate unfamiliar tasks, resulting in suboptimal performance [72]. In contrast, `gpt-4o-mini` may have avoided this pitfall by not overcomplicating the reasoning process, thus achieving better performance. Therefore, for future LLM-based personality assessment methods, researchers should consider incorporating modules that suppress over-reasoning when designing prompts. Specifically, this can be achieved by explicitly instructing the model not to ”over-reason” in the prompts, or through in-context learning approaches by including exemplary assessment cases that align with expected outcomes in the prompts.

7.4.2 Integrating data from more modalities has the potential to achieve more accurate personality assessment.

In our ablation study, we observed that the addition of the emotion data slightly increased the error rate in a few cases. We believe this could be attributed to two main factors: First, although we assigned emotion labels to sentences by considering game rules and context, current LLMs have certain limitations in accurately assigning these labels. For example, the model may struggle to correctly identify complex or mixed emotions. Second, the personality traits data might already include some emotional information, leading to redundancy or even conflict when the LLM incorrectly predicts the emotions. In our work, we only considered textual modality data. However, features like vocal intonation in speech as well as facial expression in video data are also important for personality assessment and can help us better understand humans. Future personality assessment should fully consider these multi-modal data, thereby enhancing the effectiveness of personality assessment.

7.5 Limitations and Future Directions

We have mentioned some limitations and future work in the above discussion. Here we summarize and supplement these considerations.

(1) Balancing context-dependency with trait-context matching. Currently, by interacting with multiple agents exhibiting different personalities, we can induce the user’s personality presentation across different dimensions. However, as discussed in Section 7.2, trait-context matching also matters. Future work should explore adaptive agent selection approaches that combine the comprehensive advantages of multi-situation assessment with the precision of personality-matched interactions. We should also design more diverse and targeted game scenarios. For example, puzzle games or exploratory tasks could be used to better measure openness and creativity, thereby enhancing the induction of multi-dimensional personality presentation.

(2) Exploring the applicability of our Multi-PR GPA framework in other tasks. The multi-agent interaction paradigm could evaluate context-dependent psychological constructs that manifest differently across social situations. As an exploratory study, we have focused on evaluating personality. In the future, it can be expanded to include assessments like emotional intelligence, social skills, and team collaboration styles. For instance, team collaboration styles could benefit from diverse collaborative scenarios with agents exhibiting different team dynamics.

(3) Expanding data modalities. Our current personality assessments are based on multi-type textual data. While these data provide rich information, they are still limited. Recent advancements in LLM technology have enabled the processing of additional modalities, such as images. However, our current framework has not yet incorporated other modalities like speech and video. Future work should consider the fusion of multi-modal data to enhance the comprehensiveness and accuracy of personality assessments.

8 Conclusion

In this paper, we introduce the Multi-PR GPA, a novel framework for assessing personality in game environments. This framework assesses personality by mining multi-personality representations from users’ interactions with LLM agents embodying different personalities, and provides evidence-based assessment results. We implemented a prototype system based on Multi-PR GPA and the Big Five personality model, and conducted a user study to evaluate the effectiveness of Multi-PR GPA. The results show that our multi-situation approach achieves better assessment performance compared to single-situation approaches. Additionally, we analyzed user experience during the experiment. The results support the idea that natural interaction is important for personality assessment, and we identified areas for improvement. With the key insights and implications derived from our study, we hope this work can serve as an exploratory step toward better interactive personality assessment systems, and ultimately contribute to more accessible mental health screening and support.

References

[1] W. H. Organization, *World mental health today: latest data*. World Health Organization, 2025.

[2] B. B. Lahey, “Public health significance of neuroticism.” *American Psychologist*, vol. 64, no. 4, p. 241, 2009.

[3] M. J. Constantino, J. F. Boswell, A. E. Coyne, T. P. Swales, and D. R. Kraus, “Effect of matching therapists to patients vs assignment as usual on adult psychotherapy outcomes: A randomized clinical trial,” *JAMA psychiatry*, vol. 78, no. 9, pp. 960–969, 2021.

[4] W. Mischel, “Toward a cognitive social learning reconceptualization of personality.” *Psychological review*, vol. 80, no. 4, p. 252, 1973.

[5] M. Snyder, “Personality and social behavior.”

[6] H. Blumer, *Symbolic interactionism: Perspective and method*. Univ of California Press, 1986.

[7] W. Mischel, *Personality and assessment*. Psychology Press, 2013.

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

[9] R. S. Kreitchmann, F. J. Abad, V. Ponsoda, M. D. Nieto, and D. Morillo, “Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items,” *Frontiers in psychology*, vol. 10, p. 2309, 2019.

[10] H. Rorschach, P. T. Lemkau, B. T. Kronenberg, and W. Morgenthaler, "Psychodiagnostics: A diagnostic test based on perception, including the application of the form interpretation test, rev. and enlarged," 1942.

[11] P. Mussel, T. Gatzka, and J. Hewig, "Situational judgment tests as an alternative measure for personality assessment," *European Journal of Psychological Assessment*, vol. 34, no. 5, pp. 328–335, 2016.

[12] N. Crisp and L. Chen, "Global supply of health professionals," *New England Journal of Medicine*, vol. 370, no. 10, pp. 950–957, 2014.

[13] H. Peters and S. Matz, "Large language models can infer psychological dispositions of social media users," *PNAS Nexus*, vol. 3, no. 6, p. pgae231, 2024.

[14] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *Journal of artificial intelligence research*, vol. 30, pp. 457–500, 2007.

[15] American Psychological Association, "Multiple selves," *APA Dictionary of Psychology*, n.d., retrieved September 2, 2025, from https://dictionary.apa.org/multiple-selves.

[16] J. L. Harman and J. Purl, "Advances in game-like personality assessment," *Trends in Psychology*, vol. 32, no. 4, pp. 1445–1459, 2024.

[17] N. Weidner and E. Short, "Playing with a purpose: The role of games and gamification in modern assessment practices." 2019.

[18] F. Y. Wu, E. Mulfinger, L. Alexander III, A. L. Sinclair, R. A. McCloy, and F. L. Oswald, "Individual differences at play: An investigation into measuring big five personality facets with game-based assessments," *International Journal of Selection and Assessment*, vol. 30, no. 1, pp. 62–81, 2022.

[19] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and inducing personality in pre-trained language models," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[20] G. Serapio-García, M. Safdari, C. Crepy, L. Sun, S. Fitz, P. Romero, M. Abdulhai, A. Faust, and M. Matarić, "Personality traits in large language models," *arXiv preprint arXiv:2307.00184*, 2023.

[21] J.-t. Huang, W. Jiao, M. H. Lam, E. J. Li, W. Wang, and M. Lyu, "On the reliability of psychological scales on large language models," in *Proceedings of The 2024 Conference on Empirical Methods in Natural Language Processing*, 2024, pp. 6152–6173.

[22] B. Reeves and C. Nass, "The media equation: How people treat computers, television, and new media like real people," *Cambridge, UK*, vol. 10, no. 10, pp. 19–36, 1996.

[23] W. Mischel and Y. Shoda, "A cognitive-affective system theory of personality: reconceptualizing situations, dispositions, dynamics, and invariance in personality structure." *Psychological review*, vol. 102, no. 2, p. 246, 1995.

[24] G. W. Allport, "Pattern and growth in personality." 1961.

[25] S. B. Sells and R. B. Cattell, "Personality and motivation structure and measurement," *The American Journal of Psychology*, p. 620.

[26] G. Matthews, I. J. Deary, and M. C. Whiteman, *Personality traits*. Cambridge University Press, 2003.

[27] L. R. Goldberg, "Language and individual differences: The search for universals in personality lexicons," *Review of personality and social psychology*, vol. 2, no. 1, pp. 141–165, 1981.

[28] B. De Raad, *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers, 2000.

[29] O. P. John, L. P. Naumann, and C. J. Soto, "Paradigm shift to the integrative big five trait taxonomy," *Handbook of personality: Theory and research*, vol. 3, no. 2, pp. 114–158, 2008.

[30] P. T. Costa and R. R. McCrae, "A five-factor theory of personality," *Handbook of personality: Theory and research*, vol. 2, no. 01, 1999.

[31] L. R. Goldberg, "An alternative "description of personality": The big-five factor structure," in *Personality and Personality Disorders*. Routledge, 2013, pp. 34–47.

[32] O. P. John and S. Srivastava, *Handbook of Personality: Theory and Research*, 2nd ed., L. A. Pervin and O. P. John, Eds. New York: Guilford Press, 1999, chinese edition: Lawrence A. Pervin, Oliver P. John, 2003:135–184.

[33] J. W. Pennebaker and L. A. King, "Linguistic styles: language use as an individual difference." *Journal of personality and social psychology*, vol. 77, no. 6, p. 1296, 1999.

[34] W. Revelle and K. R. Scherer, "Personality and emotion," *Oxford companion to emotion and the affective sciences*, vol. 1, pp. 304–306, 2009.

[35] E. Diener, R. J. Larsen, and R. A. Emmons, "Person× situation interactions: Choice of situations and congruence response models." *Journal of personality and social psychology*, vol. 47, no. 3, p. 580, 1984.

[36] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[37] J. Liu, H. Gu, T. Zheng, L. Xiang, H. Wu, J. Fu, and Z. He, "Dynamic generation of personalities with large language models," *arXiv preprint arXiv:2404.07084*, 2024.

[38] M. J. Gomez, J. A. Ruipérez-Valiente, and F. J. G. Clemente, "A systematic literature review of game-based assessment studies: Trends and challenges," *IEEE Transactions on Learning Technologies*, vol. 16, no. 4, pp. 500–515, 2022.

[39] J.-L. McCord, J. L. Harman, and J. Purl, "Game-like personality testing: An emerging mode of personality assessment," *Personality and Individual Differences*, vol. 143, pp. 95–102, 2019.

[40] P. J. Ramos-Villagrasa, E. Fernández-del Río, R. Hermoso, and J. Cebrián, "Are serious games an alternative to traditional personality questionnaires? initial analysis of a gamified assessment," *Plos one*, vol. 19, no. 5, p. e0302429, 2024.

[41] L. H. Gilpin, D. M. Olson, and T. Alrashed, "Perception of speaker personality traits using speech signals," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–6.

[42] S. Kim, J. Ha, and J. Kim, "Detecting personality unobtrusively from users' online and offline workplace behaviors," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–6.

[43] S. Berkovsky, R. Taib, I. Koprinska, E. Wang, Y. Zeng, J. Li, and S. Kleitman, "Detecting personality traits using eye-tracking data," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12.

[44] T. Zhang, A. Koutsoumpis, J. K. Oostrom, D. Holtrop, S. Ghassemi, and R. E. de Vries, "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*, 2024.

[45] T. Zhang, T. Qi, A. Koutsoumpis, Y. Zong, W. Zheng, J. K. Oostrom, D. Holtrop, Z. Luo, and R. E. de Vries, "Assessing personality traits and interview performance from asynchronous video interviews," in *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025, pp. 13 895–13 900.

[46] L. Hu, H. He, D. Wang, Z. Zhao, Y. Shao, and L. Nie, "Llm vs small model? large language model based text augmentation enhanced personality detection model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, 2024, pp. 18 234–18 242.

[47] T. Yang, T. Shi, F. Wan, X. Quan, Q. Wang, B. Wu, and J. Wu, "Psycot: Psychological questionnaire as powerful chain-of-thought for personality detection," *arXiv preprint arXiv:2310.20256*, 2023.

[48] Z. Li, D. Zhu, Q. Ma, W. Xiong, and S. Li, "EERPD: Leveraging emotion and emotion regulation for improving personality detection," in *Proceedings of the 31st International Conference on Computational Linguistics*, Jan. 2025, pp. 7721–7734.

[49] H. Rao, C. Leung, and C. Miao, "Can chatgpt assess human personalities? a general evaluation framework," *arXiv preprint arXiv:2303.01248*, 2023.

[50] J. Lee, Y. Choi, M. Song, and S. Park, "Chatfive: Enhancing user experience in likert scale personality test through interactive conversation with llm agents," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces*, 2024, pp. 1–8.

[51] Q. Yang, Z. Wang, H. Chen, S. Wang, Y. Pu, X. Gao, W. Huang, S. Song, and G. Huang, "PsychoGAT: A novel psychological measurement paradigm through interactive fiction games with LLM agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Aug. 2024, pp. 14 470–14 505.

[52] M. Gurven, C. Von Rueden, M. Massenkoff, H. Kaplan, and M. Lero Vie, "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*, vol. 104, no. 2, p. 354, 2013.

[53] N. Kuper, S. M. Breil, K. T. Horstmann, L. Roemer, T. Lischetzke, R. A. Sherman, M. D. Back, J. J. Denissen, and J. F. Rauthmann, "Individual differences in contingencies between situation characteristics and personality states." *Journal of Personality and Social Psychology*, vol. 123, no. 5, p. 1166, 2022.

[54] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[55] O. Sporns and R. F. Betzel, "Modular brain networks," *Annual review of psychology*, vol. 67, no. 1, pp. 613–640, 2016.

[56] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *science*, vol. 211, no. 4489, pp. 1390–1396, 1981.

[57] E. Fehr and S. Gächter, "Altruistic punishment in humans," *Nature*, vol. 415, no. 6868, pp. 137–140, 2002.

[58] C. Cruz-Neira, D. J. Sandin, T. A. DeFanti, R. V. Kenyon, and J. C. Hart, "The cave: Audio visual experience automatic virtual environment." *Communications of the ACM*, vol. 35, no. 6, pp. 64–73, 1992.

[59] S. Bouchard and A. Rizzo, *Virtual reality for psychological and neurocognitive interventions*. Springer, 2019.

[60] Y. Jia, B. Xu, Y. Karanam, and S. Voida, "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*, 2016, pp. 2001–2013.

[61] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang, "Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses," *Behavior research methods*, vol. 41, no. 4, pp. 1149–1160, 2009.

[62] M. Czikszentmihalyi, *Flow: The psychology of optimal experience*. New York: Harper & Row, 1990.

[63] C. Peifer and S. Engeser, *Advances in flow research*. Springer, 2021.

[64] Y. Bian, C. Yang, C. Zhou, J. Liu, W. Gai, X. Meng, F. Tian, and C. Shen, "Exploring the weak association between flow experience and performance in virtual environments," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–12.

[65] T. P. Novak, D. L. Hoffman, and Y.-F. Yung, "Measuring the customer experience in online environments: A structural modeling approach," *Marketing science*, vol. 19, no. 1, pp. 22–42, 2000.

[66] C.-C. Liu, I. Chang *et al.*, "Measuring the flow experience of players playing online games," 2012.

[67] C. S. Oh, J. N. Bailenson, and G. F. Welch, "A systematic review of social presence: Definition, antecedents, and implications," *Frontiers in Robotics and AI*, vol. 5, p. 409295, 2018.

[68] A. Felnhofer, O. D. Kothgassner, N. Hauk, L. Beutl, H. Hlavacs, and I. Kryspin-Exner, "Physical and social presence in collaborative virtual environments: Exploring age and gender differences with respect to empathy," *Computers in Human Behavior*, vol. 31, pp. 272–279, 2014.

[69] R. Carciofo, J. Yang, N. Song, F. Du, and K. Zhang, "Psychometric evaluation of chinese-language 44-item and 10-item big five personality inventories, including correlations with chronotype, mindfulness and mind wandering," *PloS one*, vol. 11, no. 2, p. e0149963, 2016.

[70] R. P. Tett and D. D. Burnett, "A personality trait-based interactionist model of job performance." *Journal of Applied psychology*, vol. 88, no. 3, p. 500, 2003.

[71] U. Bronfenbrenner, "Toward an experimental ecology of human development." *American psychologist*, vol. 32, no. 7, p. 513, 1977.

[72] S. Wu, J. Xie, J. Chen, T. Zhu, K. Zhang, and Y. Xiao, "How easily do irrelevant inputs skew the responses of large language models?" in *First Conference on Language Modeling*, 2024.

Baiqiao Zhang is a Ph.D. student in Computer Science and Engineering at The Hong Kong University of Science and Technology. He received the B.Eng degree in Computer Science and Technology through a joint program between Shandong University and the Australian National University in 2025. His research interests include human-computer interaction, natural language processing, and affective computing.

Xiangxian Li is a Post-doctoral Researcher at the School of Airspace Science and Engineering, Shandong University. He received the B. Eng. Degree in Digital Media Technology from Huazhong University of Science and Technology, China, in 2018, and the Ph.D. degree from the Shandong University, China, in 2024. His research interests include deep learning, multimedia computing, and intelligent human-computer interaction.

Xinyu Gai is a MPhil student at The Hong Kong University of Science and Technology (Guangzhou). He received the B.Eng degree in Computer Science and Technology from Shandong University in 2024. His research interests include human-computer interaction, virtual reality, and user experience.

Chao Zhou is a Post-doctoral Researcher at the Institute of Software Chinese Academy of sciences. She received her B.S. and M.E. degrees in developmental and educational psychology from Shandong Normal University, in 2010 and 2013. She received her Ph.D. degree in Brain and Cognitive Neuroscience Research Center, Liaoning Normal University, China. Her research interests include strategy utility and the central executive function in human cognition, perception and performance in HCI.

Juan Liu is an Associate Professor with School of Airspace Science and Engineering, Shandong University, China. She received her B.S. degree from Zhejiang University Of Media and Communications in 2008, and M.E. degree from Chongqing University in 2011. She received her PhD degree in Computer Science and Technology from Shandong University in 2022. Her research interests include mixed reality and serious game.

Xue Yang is a Post-doctoral Researcher with the Department of Computer Science and Technology, Tsinghua University. She received her M.E degree in basic Psychology from Southwest University in 2018. She received her Ph.D degree in Management from Fuzhou University in 2023. Her research interests including human decision-making, social psychology and cognitive neuroscience.

Xiaojuan Ma is an Associate Professor with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. She's a senior member of IEEE and ACM. Her academic journey includes a PhD from Princeton University and postdoctoral research with Carnegie Mellon University. She is currently an Associate Editor of ACM Trans. Computer-Human Interaction and Paper Co-Chair of ACM CHI 2026, CSCW 2025, 2026.

Yong-Jin Liu is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the B. Eng. Degree in mechano-electronic engineering from Tianjin University, China, in 1998, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2003. He is a senior member of IEEE. His research interests include intelligent media processing, affective computing and human-computer interaction. He is currently an Associate Editor of IEEE Trans. Affective Computing.

Yulong Bian is an Associate Professor with School of Airspace Science and Engineering, Shandong University, China. He received the B.S., M.E. and Ph.D. degrees in developmental and educational psychology from Shandong Normal University, in 2010, 2013 and 2016, respectively. His research interests include human-computer interaction, virtual reality, computer-assisted psychological intervention and user experience.