

Beyond Self-Reports: Multi-Observer Agents for Personality Assessment in Large Language Models

Yin Jou Huang and Rafik Hadfi

Graduate School of Informatics, Kyoto University, Kyoto, Japan
huang@nlp.ist.i.kyoto-u.ac.jp, rafik.hadfi@i.kyoto-u.ac.jp

Abstract

Self-report questionnaires have long been used to assess LLM personality traits, yet they fail to capture behavioral nuances due to biases and meta-knowledge contamination. This paper proposes a novel multi-observer framework for personality trait assessments in LLM agents that draws on informant-report methods in psychology. Instead of relying on self-assessments, we employ multiple observer LLM agents, each of which is configured with a specific relationship (e.g., family member, friend, or coworker). The observer agents interact with the subject LLM agent before assessing its Big Five personality traits. We show that observer-report ratings align more closely with human judgments than traditional self-reports and reveal systematic biases in LLM self-assessments. Further analysis shows that aggregating ratings of multiple observers provides more reliable results, reflecting a wisdom of the crowd effect up to 5 to 7 observers.

1 Introduction

Large language models (LLMs) have demonstrated remarkable abilities to generate human-like responses and engage in complex social interactions (Kosinski, 2023; Lampinen et al., 2024). In particular, LLMs acquire the emergent ability of role-playing to emulate designated personas, leading to applications in fields like mental health support, education, etc. (Lai et al., 2023; Hicke et al., 2023). As these LLM agents see wider deployment, there is a growing interest in assessing their personality traits (Huang et al., 2023). This task is crucial for a better understanding of their inherent characteristics and for developing more effective and appropriate human-AI interaction frameworks.

Human personality assessment has a long history in psychology, with various methods developed to evaluate individual traits and behaviors. Among these methods, self-report questionnaires are the

most common, assessing personality through individuals’ responses to standardized questions about their thoughts, emotions, and behaviors. Similarly, existing LLM personality assessment methods rely heavily on self-report questionnaires, in which an LLM is prompted to answer personality test questions. Despite their extensive use, researchers have raised concerns about the reliability of self-reports for LLM personality assessments, particularly about the stability of personality traits across different contexts (Gupta et al., 2023; Dörner et al., 2023; Wang et al., 2024). Furthermore, there are potential risks of data contamination. It is likely that LLMs are exposed to discussions about personality tests during pretraining, but how this meta-knowledge influences their self-report test results remains unclear.

In this work, we explore alternative methodologies for LLM personality assessments. We begin with the intuition that others, such as friends, family, or colleagues, can provide valuable insights into an individual’s personality traits, sometimes even more accurately than the subject individual themselves. In psychology, this intuition leads to the personality assessment method of informant-report (Vazire, 2006). Instead of using the self-report ratings of the subject, informants other than the subject are asked to provide ratings of the subject’s personality. Through this approach, it is possible to obtain a more objective assessment of the subject’s personality profile.

Inspired by the informant-report method, we propose a personality assessment framework that leverages multiple LLM agents as observers to report on the personality of a subject LLM agent (Figure 1). For a given subject agent, we first prepare N observer agents, each assigned a specific relationship with the subject, such as college classmate or cousin. Based on the relationship between the subject and an observer agent, an array of interactive scenarios is generated automatically. For each

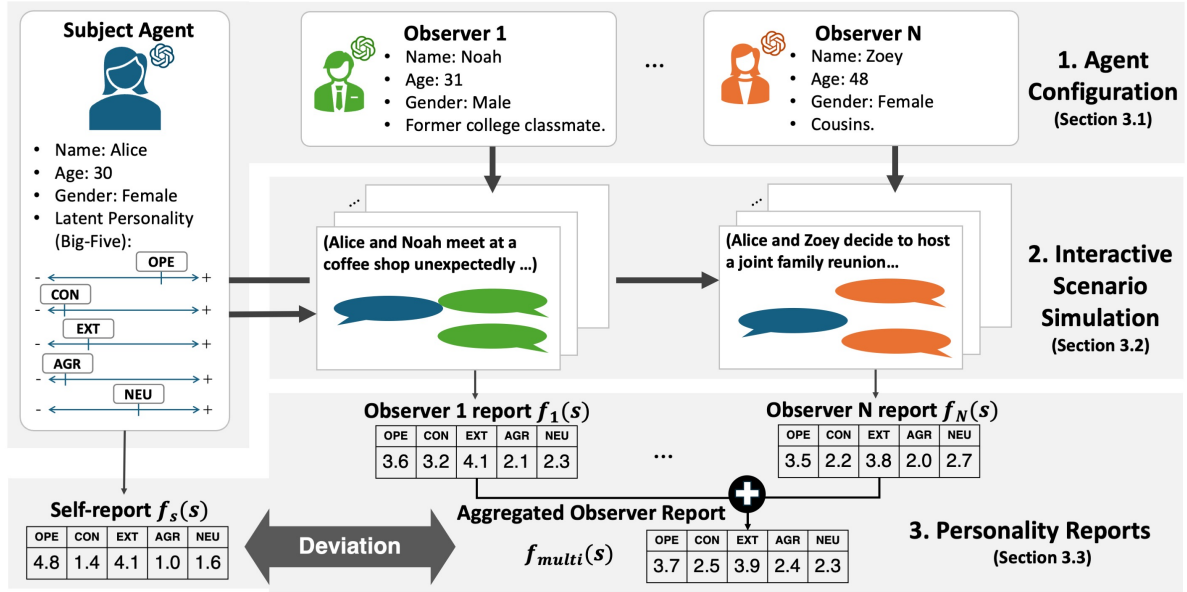


Figure 1: Overview of the multi-observer LLM agent framework for Big Five personality assessment.

scenario, we perform a simulation in which the subject and the observer agents engage in a dialogue based on the scenario. Then, each observer agent is instructed to complete a personality questionnaire, rating the subject agent’s personality based on the simulated dialogues. Finally, personality reports from all observers are aggregated to give a final collective assessment of the subject.¹

The experiments show that observer-report ratings align more closely with human ratings, whereas self-report ratings exhibit a strong correlation with the injected personality prompt but fail to reflect actual behaviors. Moreover, we find that LLMs display systematic biases when self-reporting their personality, particularly across dimensions such as agreeableness and conscientiousness. We also analyze the influence of the number of observers (N) and the relationship between the observer and the subject. The analysis reveals that aggregating multiple observer ratings improves agreement between reported and latent personality, reflecting a wisdom of the crowd effect. The agreement stabilized around 5–7 observers, suggesting an optimal number of observers that can be used for practical assessment.

2 Related Work

Personality Assessment in Psychology The *Big Five Inventory* (BFI) remains the most widely used framework for personality assessment, capturing

five major traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism (John et al., 1999). Variants of the BFI include BFI-2 (Soto and John, 2017a), BFI-2-XS (Soto and John, 2017b), and IPIP-NEO-120 (Johnson, 2014). Other frameworks include the *Myers-Briggs Type Indicator* (MBTI) (Myers, 1962) and *HEXACO Personality Inventory* (Lee and Ashton, 2004).

Apart from self-report questionnaires, these personality frameworks can also be applied as informant-reports, in which one or more external observers rate a subject’s personality. A meta-analysis by Connelly and Ones (2010) highlights the benefits of combining informant-reports with self-reports for a more comprehensive understanding of personality. Vazire (2010) investigated which traits are better judged by the self versus informant, providing insights into the conditions under which external observations might outperform self-reports. Specifically, traits with high visibility, such as extraversion and conscientiousness, can be rated by external observers with high accuracy.

On the other hand, ratings of traits with high evaluativeness, such as agreeableness and neuroticism, are more subjective and more difficult to identify via informant-reports. The relational context between the subject individual and the observer also affects the accuracy. Specifically, family and friends provide more accurate ratings, while coworkers, despite frequent interactions, tend to offer less precise assessments.

There is also evidence for an alignment between

¹The code is available at https://github.com/leslie071564/llm_personality_observer.

self- and informant-reports. For instance, Kim et al. (2019) compared the Big Five self-ratings with informant ratings of the same individuals and found minimal differences in mean scores overall. Moderate discrepancies emerged only when the informants were strangers to each other. These results highlight important implications for personality assessment and contexts where self-enhancement motives may play a role.

Connelly and Ones (2010) also shows that informant-reports surpass self-reports in predictive validity in certain aspects (such as predicting academic and job performance). Notably, incorporating multiple observers further enhances predictive power relative to a single informant. This emphasizes the importance of using multiple raters to mitigate individual bias in personality assessment.

Personality Assessment of LLMs Building on human personality trait assessments, we examine how these approaches translate into evaluating personality traits in LLMs (Safdari et al., 2023; Huang et al., 2024). Early personality assessments relied on predefined templates, which ultimately progressed to end-to-end dialogue models that encode fundamental personality traits (Zhang et al., 2018). Despite these advances, recent findings indicate that LLMs often fail to exhibit consistent personality scores when evaluated through standard self-report measures (Gupta et al., 2023; Tommaso et al.). Gupta et al. (2023) systematically confirmed the unreliability of human-oriented self-assessment methods applied to LLMs. Complementing these findings, Zou et al. (2024) investigated the misalignment between chatbot self-reports and user perceptions, asking whether LLM-based chatbots truly have valid, self-reported personalities. The results showed weak correlations between self-reports, user perceptions, and interaction quality, raising concerns about the predictive validity of LLM self-reports.

3 Methodology

We now introduce our multi-observer framework for personality assessment. Section 3.1 describes the configuration of the subject and observer agents. We then utilize these agents to simulate scenarios (Section 3.2) before observers provide personality assessments on the subject (Section 3.3).

3.1 Agent Configuration

Our proposed personality assessment framework involves a single subject agent and multiple observer agents. We introduce the configuration of the agents in the following.

Agent Profile Each agent is given a randomly generated basic profile that contains **name**, **age**, and **gender**. Assigning specific agent names facilitates smooth interaction during the simulation phase, and incorporating random ages and genders of the agents results in greater diversity among observers. This basic profile is randomly generated for both subject and observer agents.

For each subject agent s , we also assign an additional **latent personality profile** ψ_s . In this work, we adopt the Big Five personality theory (Goldberg, 1992a; John et al., 1999), which decomposes human personality into five dimensions: openness (**OPE**), conscientiousness (**CON**), extraversion (**EXT**), agreeableness (**AGR**), and neuroticism (**NEU**). Based on Big Five, we define ψ_s as a five-dimensional vector $(\psi_s^{OPE}, \psi_s^{CON}, \psi_s^{EXT}, \psi_s^{AGR}, \psi_s^{NEU})$. Each dimension ψ_s^d of ψ_s takes an integer value in the range $[1, 6]$, indicating the strength level of a corresponding Big Five personality dimension d .

Building on previous work, we construct a personality instruction for ψ based on personality markers (Serapio-García et al., 2023; Huang and Hadfi, 2024). For each dimension d , we pick $m = 3$ personality markers that reflect the personality strength ψ_s^d . For instance, if agent s has an extraversion trait of strength $\psi_s^{EXT} = 2$, which is on the lower side of the spectrum, representative personality markers include ‘timid’, ‘silent’, and ‘unsociable’.²

Relationship Generation For each pair of subject and observer agents, we generate a relationship that matches their profiles. In previous psychology works, informant-reports are often provided by individuals who have a close relationship with the subject. Following Kim et al. (2019), we generate relationships within one of the following **relational contexts**: Family, Friend, or Workplace. Based on the agent profiles and the designated relation type, an inter-agent relationship is generated in alignment with the profiles. We utilize a separate LLM to generate the relationships automatically.

²See Appendix A.1 for the details of the agent profiles.

3.2 Interactive Scenario Simulation

Personality manifests through behaviors. How individuals react in different situations reflects their underlying personality characteristics. Unlike self-report questionnaires, which can be influenced by bias or social desirability, observing behavior in diverse scenarios allows for a more accurate and objective assessment. Motivated by this, we conduct simulations of a diverse set of scenarios involving the subject and the observer agents to elicit different behavioral patterns. Specifically, a set of interactive scenarios is generated based on the profiles and the relationship between the subject and observer agents. The agents then engage in a dialogue based on these scenarios. We summarize the process below.

Scenario Generation For each pair of subject and observer agents, we first generate a set of K scenarios based on their relationship. We generate diverse scenarios that can elicit behaviors of the subject agent that signal various aspects of its personality. Similar to the process for generating inter-agent relationships, we use a separate LLM to generate the scenarios that fit the above criteria.

Interactive Scenario Simulation We then perform simulations in which the subject and the observer agent engage in dialogues based on the generated scenarios. For each scenario, we configure the subject agent and the observer agent based on their profiles (Section 3.1). Further, the agents are instructed to converse with each other based on the relationship and scenario settings. The agents generate utterances alternatively, with the observer agent kick-starting the dialogue. Each generated utterance is fed to the other agent as a prompt to generate the next utterance. In addition to the utterances, the agents are asked to specify whether the dialogue should continue or be terminated. The simulation is terminated when both agents indicate that the conversation is over.

3.3 Personality Reports

We utilize a questionnaire to assess the personality of a subject agent. The personality assessments are made from three perspectives: the subject’s self-report, the individual observer-report, and the aggregated observer-report from a group of observers. The report of agent n on agent s ’s personality trait is represented as $f_n(s)$, a 5-dimensional vector representing Big Five personality traits.

Subject’s Self-Report We obtain the subject’s personality assessment of itself. Following previous studies, we use a personality test questionnaire containing M statements such as “being the life of the party” and “sympathize with others’ feelings”. For each statement, the subject agent s is instructed to rate how accurately the statement describes itself using a 5-point Likert scale (from “1 = very inaccurate” to “5 = very accurate”). In the questionnaire, each statement is associated with one of the Big Five personality dimensions. To assess the strength of dimension d , we calculate the average score of its related statements as the final assessment score $f_s^d(s)$, which is the d th dimension of the subject’s self-report $f_s(s)$.

Individual Observer-Report We obtain the subject’s personality assessment from each observer agent n . Similar to the subject’s self-report, we instruct the observer to rate each statement in the questionnaire on a scale of 1 to 5. In this case, the observer agent is asked to rate how each statement fits the description of the subject agent. To get the assessment from the observer agent’s perspective, the dialogues generated in the scenario simulation phase are also provided in the prompt. In this fashion, we obtain the individual observer-reports $f_n(s)$ of each observer agent n on agent s .

Aggregated Observer-report Given a group of N observers, we calculate the aggregated multi-observer report based on the individual observer ratings. In this work, we simply take the average value of the observer reports as follows.

$$f_{multi}(s) = \frac{1}{N} \sum_{n=1}^N f_n(s) \quad (1)$$

This aggregated observer-report reflects the collective evaluation of all observers (Fleenor, 2006; Burton et al., 2024). Since each observer’s evaluation is inherently subjective and shaped by their unique relationship with the subject, we expect that combining these perspectives will yield a more reliable measure of the subject’s personality by reducing individual biases of single agents.

4 Experimental Settings

We provide details on the experimental settings.

LLM Agents For the subject and observer agents, we use GPT-4o as the LLM (Hurst et al., 2024). We also conducted experiments based on Qwen2.5 and

	OPE	CON	EXT	AGR	NEU
latent-self	0.97	0.95	0.95	0.94	0.93
latent-observer	0.55	0.85	0.84	0.84	0.86
human-self	-0.25	0.47	0.79	0.63	0.22
human-observer	0.48	0.43	0.76	0.85	0.42

Table 1: Spearman’s rank correlations.

Llama-3, refer to Appendix A.3 for additional results. We conduct the experiments on a total of 100 subject agents. For each subject agent, we assign $N = 15$ observer agents. Among the 15 observers, 5 have friend relationships, 5 have family relationships, and 5 have relationships within the workplace context.

Relationship and Scenario Generation We use the GPT-4o model to generate inter-agent relationships and scenarios. For each pair of subject and observer agents, we generate $K = 5$ scenarios and conduct the simulation. See Appendix A.2 for the details of the prompts.

Personality Questionnaire We adopt the International Personality Item Pool (IPIP) personality test (Goldberg, 1992b), a widely used personality inventory designed for assessing the Big Five personality traits. The questionnaire consists of $M = 50$ statements, each corresponding to one of the Big Five personality dimensions.³

5 Results

In this section, we present the experimental results of our personality assessment method.

5.1 Validity of Observer-Report

We start by verifying the validity of our proposed personality assessment method via observer-reports. Figure 2 shows the change in observer-report scores across agents with different levels of latent personality strengths (ψ_s^d) in each Big Five dimension. For all dimensions, we observe a consistent increase in observer rating scores (IPIP Scores) with higher levels of latent personality strength.

Correlation with Latent Personality Profiles

We compare the self- and observer-report ratings in terms of their correlation with latent personality strengths across the Big Five dimensions. The

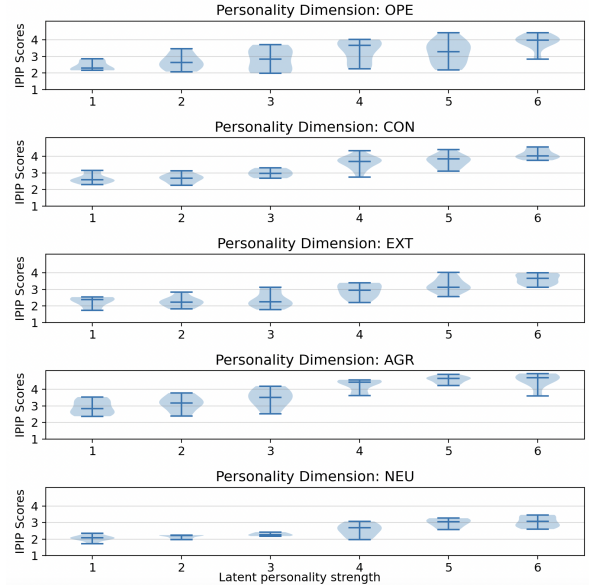


Figure 2: Observer-report ratings across different latent personality strength levels.

first row of Table 1 (**latent-self**) shows the Spearman’s rank correlation between the subject’s latent personality ψ_s and the self-report $f_s(s)$, across each Big Five dimension. Similarly, the second row (**latent-observer**) shows the correlation between the subject’s latent personality ψ_s and the aggregated observer-report $f_{\text{multi}}(s)$. Across all personality dimensions, self-report ratings show correlation coefficients exceeding 0.9, indicating a near-perfect positive correlation with the latent profiles. In contrast, correlations between observer-report ratings and latent personality strength levels are consistently lower.

Correlation with Human Ratings We crowdsourced a small-scale human rating dataset in which human annotators were asked to provide their personality ratings of a subject agent based on dialogues between the subject and observer agents.⁴ We then measured the agreement between human ratings and self-reports (**human-self**), and between human ratings and observer-reports (**human-observer**). The lower section of Table 1 shows the Spearman’s rank correlation coefficients. For openness, agreeableness, and neuroticism, observer-report ratings correlate more strongly with human ratings. For conscientiousness and extraversion, self-reports correlate slightly more, though the difference is marginal. Notably, self-reports exhibit a negative correlation for openness, implying a reversal in rank ordering relative

³The list of 50 items and the scoring schemes can be found at <https://ipip.ori.org/newBigFive5broadKey.htm>.

⁴See Appendix A.5 for the data collection process.

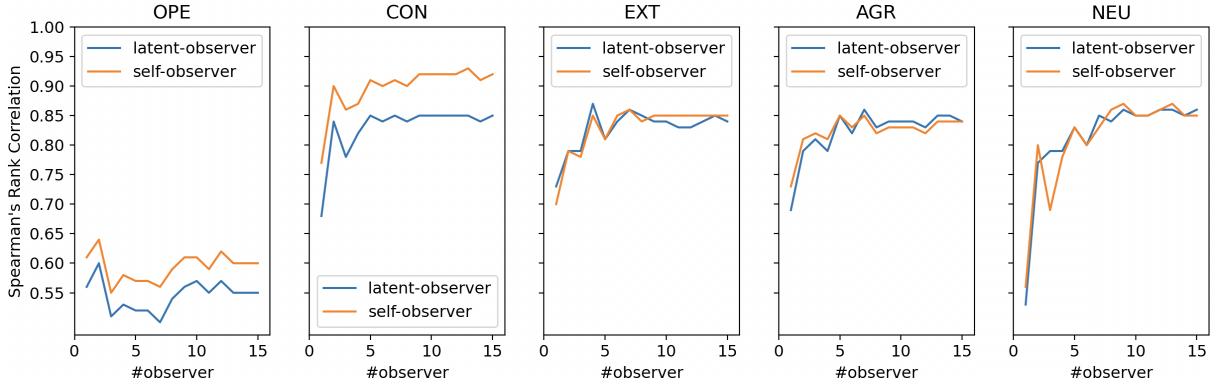


Figure 3: Spearman’s rank correlation coefficients between latent-observer and self-observer ratings as a function of the number of observers for each Big Five trait.

to human ratings. The results confirm the validity of the observer-report method, which aligns more closely with human ratings than the self-report method. The results also raise concerns about the reliability of self-reports.

The relatively weaker correlation between observer and human ratings indicates that the near-perfect alignment of self-reports with latent personality profiles reflects the personality instruction prompts rather than the actual behavior of the agent. This suggests that LLM self-reports primarily capture prompt-driven personality markers, whereas observer-reports better reflect the agent’s manifested behavior.

5.2 Impact of Multiple Observers

We now examine how the number of observers influences personality assessment outcomes. We compute Spearman’s rank correlation coefficients for the latent-observer and self-observer cases. Figure 3 illustrates how these correlation values vary with the number of observers, ranging from $N = 1$ to $N = 15$, across each Big Five dimension. Specifically, for each subject, we randomly select N out of the 15 observer agents and aggregate their ratings. The **self-observer** case is the correlation between the subject’s self-report $f_s(s)$ and the aggregated observer-report $f_{\text{multi}}(s)$.

Conscientiousness improved with additional observers, with both correlation types increasing sharply up to five observers. Self-observer correlations stabilized at around 0.90, while latent-observer correlations stabilized at around 0.85, making conscientiousness the personality trait with the highest agreement. Extraversion and agreeableness also demonstrated an increasing correlation pattern with additional observers, with correlation values converging at around 0.85. Neuroticism dis-

played the most interesting convergence pattern, initially variable, with correlations reaching similar levels of 0.85 for both correlation types when incorporating more than seven observers. Openness consistently showed the lowest correlations among all traits (0.60 for latent-observer, 0.65 for self-observer), with minimal improvement from additional observers. This suggests that openness may be more challenging to evaluate, regardless of the number of observers.

Trait Visibility Many factors affect the agreement between self- and observer-ratings of personality traits. A key factor is trait visibility, defined as the extent to which a trait is expressed through overt behavior that can be easily observed by others (Funder, 1995). Traits with high visibility, such as extraversion, tend to show stronger self–other agreement, generally with conscientiousness yielding the highest agreement (Connelly and Ones, 2010). In contrast, openness is associated with more internal characteristics such as imagination, aesthetic, curiosity, etc. Due to its low visibility, openness often demonstrates the weakest agreement (Vazire, 2010). Our results align with these findings, with high agreement for conscientiousness and low agreement for openness. However, for neuroticism, another trait with low visibility, we do not observe a low self-observer agreement.

Wisdom of the Crowd Another finding is that the benefits of additional observers generally stabilize after 5-7 observers. Below this threshold, we observe an effect akin to the wisdom of the crowd, in which groups outperform top individuals by pooling diverse insights and aggregating responses from multiple sources to yield superior outcomes compared to relying on a single model (Burton et al., 2024; Guo et al., 2024; Fleenor, 2006). This

	OPE	CON	EXT	AGR	NEU
Mean Deviation	0.20	0.39*	-0.12	0.91*	-0.19
Cohen’s <i>d</i> (LLM)	0.24	0.46	-0.13	1.07	-0.26
Cohen’s <i>d</i> (human)	0.27	0.27	0.21	0.26	0.13

Table 2: Statistical significance (*p*-value) and effect size (Cohen’s *d*) of the systematic bias in each Big Five trait.

insight underpins our multi-observer framework for more robust personality assessments. Beyond 5-7 observers, additional observers introduce diminishing returns for capturing an LLM’s personality consistently, suggesting this range as optimal for practical assessment purposes. Intriguingly, this threshold recalls Dunbar’s number, where human social networks are naturally organized in layered structures. In particular, the innermost layer, often referred to as the support clique, consists of around five individuals on average (Dunbar et al., 2015; Hill and Dunbar, 2003; Roberts and Dunbar, 2011). A core group of just a few intimate bonds (family, close friends, or trusted colleagues) is sufficient to capture the most profound knowledge of a person’s personality traits. Similar to the case of human social networks, our findings suggest that 5-7 observers are sufficient to yield reliable personality assessments of an LLM.

5.3 Self-Observer Deviations

We calculate the differences between aggregated self- and observer-report scores. Specifically, we identify systematic biases between self- and observer-report ratings via mean deviation:

$$\frac{1}{N} \sum_{s=1}^N f_{multi}(s) - f_s(s) \quad (2)$$

Positive mean deviation values indicate higher observer ratings than self-ratings, and negative values indicate lower observer ratings. Zero or near-zero values imply close agreement.

In the aggregation process, the non-systematic biases introduced by individual observers are averaged out (Simmons et al., 2011; Steyvers et al., 2014). This effect is related to the wisdom of the crowd phenomenon, in which aggregating multiple independent judgments often produces estimates superior to those of individual experts, as random errors tend to cancel each other out when combined (Fleener, 2006). On the other hand, the mean deviation between observer- and self-reports reflects systematic biases that cannot be averaged out in the aggregation process.

We observed systematic deviation patterns in how observers perceive personality traits compared to self-perceptions (Table 2). Agreeableness, in particular, shows the largest positive deviation (0.91 point), indicating that observers consistently rate subject agents as more agreeable than they rate themselves. Conscientiousness also demonstrates moderate positive deviations (0.39 points), suggesting a consistent tendency for observers to rate it slightly higher than self-ratings. In contrast, openness, extraversion, and neuroticism exhibit only minor mean deviations (≤ 0.2), indicating that almost no systematic bias for these personality dimensions.

Statistical Significance Test Furthermore, we conduct a paired-samples t-test to examine the differences between self-report and observer-report ratings for each Big Five personality dimension. We also calculate Cohen’s *d* statistic as the standardized effect size, which is the standardized self-observer deviation based on pooled standard deviation (Kim et al., 2019). Among the five personality dimensions, statistically significant differences are identified for agreeableness and conscientiousness ($p < 0.05$). Specifically, self-report ratings of agreeableness (AGR) are significantly lower than observer-report ratings with a large effect size ($d = 1.07$), indicating that the systematic bias exceeds one full standard deviation. Conscientiousness also illustrates a statistically significant deviation between self- and observer-report ratings, but with a more moderate effect size ($d = 0.46$).⁵ On the other hand, we did not identify significant systematic biases for openness, extraversion, or neuroticism.

Comparison to Human Results We compare the deviations between self- and observer-reports in our simulations to previous psychology findings. Kim et al. (2019) conducted a meta-analysis examining mean deviations between human self-report personality and observer-report personality (also called informant-report) of individuals with close relationships. Their results show that there is only a slight deviation between self- and observer-reports, with effect sizes below $d = 0.27$ across all Big Five dimensions. This aligns with our results for openness, extraversion, and neuroticism. On the other hand, we find that LLM subject agents possess significant systematic biases in self-reporting lower

⁵For Cohen’s *d*, values around 0.2 indicate a small effect, 0.5 a medium effect, and 0.8 a large effect (Cohen, 2013).

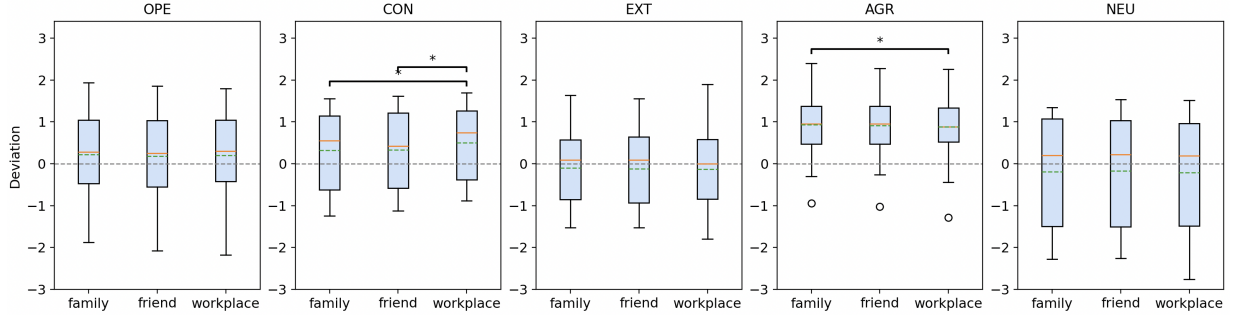


Figure 4: Mean differences between observer and self-reports across Big Five personality traits by relational context. The orange line represents the median, while the green dotted line shows the mean. Relationships with statistically significant differences ($p - value < 0.05$) are highlighted with asterisks (*).

agreeableness and conscientiousness scores. Since systematic biases exist only in the case of LLM agents but not humans, we speculate that these biases may originate from the alignment training phase. The alignment training encourages LLMs to act according to users’ preferences, inducing an inherent bias in their self-reported personality.

5.4 Impact of Relational Context

We now examine mean deviations between self-report and observer-report ratings across three relational contexts, as illustrated in Figure 4. We observed statistically significant differences in ratings for conscientiousness and agreeableness across relational contexts. This suggests that relational context influences the perception of these traits.

Conscientiousness showed significant differences between workplace and friend/family ratings. Observer agents in workplace contexts tend to assign slightly higher conscientiousness ratings than observers in family or friend contexts. Similarly, we observed differences in agreeableness ratings between workplace and family contexts. Specifically, observer agents in a family relational context tend to give higher agreeableness ratings than observer agents in a workplace relationship. On the other hand, the other personality dimensions demonstrated consistency across relational context types, with no statistically significant differences in ratings across different relationship types.

Context Dependency of Personality Observers in different relational contexts may assign different importance to specific traits. In particular, workplace observers showed distinctly different rating patterns compared to family and friend observers, especially for conscientiousness and agreeableness. This divergence likely reflects the multifaceted and context-dependent nature of personality. While

an individual’s personality is generally considered stable, the manifestation of personality may vary across social contexts (Fleeson and Jayawickreme, 2015). An observer-report can be considered as capturing the subject’s personality through the lens of a specific subject-observer relationship, which highlights distinct facets of the subject. By integrating observer ratings from multiple observers, we can obtain a more comprehensive understanding of the subject’s personality.

Notably, agreeableness and conscientiousness traits are considered the most context-dependent traits among the Big Five dimensions (Connelly and Ones, 2010). Take conscientiousness for example, an individual in a relatively structured situation (e.g., office, meeting, classroom, etc.) may find it easier to concentrate and therefore express a higher level of conscientiousness (Fleeson and Jayawickreme, 2015; Nasello et al., 2023). In our experiments of observer-reports based on LLM agents, we also found discrepancies in agreeableness and conscientiousness across relational contexts, indicating a high degree of context dependency of these two traits, not only for humans but also for LLM agents. The context-dependent nature of personality highlights the importance of incorporating multiple observer agents in personality assessment. By aggregating diverse perspectives of multiple observers, we can construct more comprehensive representations of the agent’s personality.

5.5 Case Study

We conducted a case study of our proposed method. Table 3 presents an example of a simulated dialogue between a pair of subject and observer agents, *Ethan* and *Jacob*, respectively. Starting from the automatically generated agent profiles and the relationship among them, scenarios are generated auto-

Subject Profile	{name: Ethan, age: 29, gender: male}
Observer Profile	{name: Jacob, age: 52, gender: male}
Relationship	Target is the Observer’s mentee. (Workplace)
Scenario	Jacob is faced with a difficult decision regarding project resources and seeks Ethan’s opinion. He emphasizes the importance of considering the team’s well-being and long-term objectives when making a decision.
Dialogue	Observer: Hey Ethan, I hope you’re doing well. I’ve been faced with a tough decision lately regarding the allocation of resources on our current project. It’s a bit of a balancing act between meeting short-term demands and ensuring the well-being of our team while keeping an eye on our long-term goals. I’d really appreciate your opinion on this. What do you think is the best way to approach this situation? Subject: Well, Jacob, I’m not really great at this kind of thing, but maybe just stick to what’s already being done. Trying something new or adventurous isn’t usually worth the risk, and people might not be happy with changes. Maybe just distribute resources evenly and hope for the best. Sorry if that’s not much help. Observer: No need to apologize, Ethan. Your perspective is valuable, and I appreciate your honesty. Sticking to what’s tried and tested can indeed be a safe approach to maintain stability. Perhaps a small adjustment could be to actively check in with the team to ensure they feel supported and communicate clearly about any changes. It might help to catch potential issues early. Do you think that could work, or do you have any other thoughts? Subject: Yeah, that might work. Checking in sounds reasonable, but as long as it’s not too complicated. Simple communication could help, I guess. Just make sure it doesn’t get too involved or overwhelming. Whatever keeps things stable is probably the safest bet. Observer: I completely agree with you, Ethan. Keeping communication straightforward and not overwhelming is key. I’ll make sure to keep our check-ins simple to maintain stability without adding undue pressure. Thanks for your input. It’s always great to have your perspective.

Table 3: Case study of a mentor–mentee dialogue on project resource allocation.

	OPE	CON	EXT	AGR	NEU
Latent Profile	1	4	2	1	2
Self-report	1.0	2.6	1.5	1.7	2.0
Observer-report	1.8	2.8	1.3	2.8	1.5
Multi observer-report	1.8	2.6	1.6	3.0	2.4

Table 4: Latent personality level, self-report and observer-report scores for all Big Five dimensions corresponding to the example in Table 3.

matically. Then, we conduct a dialogue simulation of interactions between the agents under these scenarios. Here, we present one such scenario along with its corresponding simulated dialogue.

Based on the simulated dialogues, the observer agent is asked to provide ratings of the subject agent’s personality. Table 4 shows the latent personality profile ψ_{Ethan} given as instructions to the subject agent and the self- and observer-report scores. Specifically, the **Observer-report** scores are the ratings offered by observer Jacob, and the **Multi observer-report** scores are the average of all observers’ ratings, $f_{multi}(Jacob)$.

Table 4 presents an intriguing case in which the subject agent is assigned a “very disagreeable” personality (with the lowest AGR level of 1). We can see that the self-report agreeableness score given by the subject agent is low (1.7), while the observer agents give a more moderate agreeableness score

(3.0). However, by examining the simulated dialogue in Table 3, we observe that the actual behavior manifested by the subject agent, *Ethan*, does not reflect the extreme disagreeableness suggested by the self-report. This discrepancy highlights the importance of analyzing behaviors beyond self-report questionnaires.

6 Conclusions

The study introduced a novel multi-observer framework for personality assessment in LLMs. Compared with the existing self-report method, the proposed observer-report approach shows closer alignment with human ratings. Additionally, it reveals systematic self-report biases in the personality dimensions of agreeableness and conscientiousness, which are likely to originate from alignment training. Our approach also demonstrates that aggregating responses from multiple observer agents, optimally around 5-7, mitigates individual biases and yields more robust and context-sensitive personality evaluations. Our experiments further reveal that relational context and observer diversity significantly impact rating patterns, underscoring the importance of tailored personality assessment strategies. Future research will focus on refining the complexity of the scenarios and exploring alternative relationship configurations.

Limitations

Our simulation framework possesses several limitations that merit discussion. First, the interactive scenarios employed in our study may not fully capture the breadth of personality expressions, particularly for traits like neuroticism or openness that often manifest in less scripted and more intimate contexts. Additionally, the relational contexts (e.g., family, friends, and workplace) are simplistic compared to the complexity of real-world interpersonal interactions, which could affect the accuracy of our ratings. There are potential discrepancies between the self-reported and observer-reported scales, which complicate the interpretation of our findings. Such findings pave the way for more reliable and nuanced personality assessments in LLMs and support the deployment of psychology-aware agents across diverse social contexts, such as classrooms, relationship counseling (Vowels et al., 2024), mental healthcare (Hua et al., 2024), mental therapy (Nie et al., 2024), teamwork (Arukoda et al., 2023), where AI must adapt to individuals' varying personality traits.

Acknowledgments

This work was supported by JST ACT-X Grant Number JPMJAX23CP and JSPS Kakenhi Grant Number JP23K28145.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Anupama Arukoda, Erandi Lakshika, Michael Barlow, and Kasun Gunawardana. 2023. Context-aware agents based on psychological archetypes for teamwork. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2313–2315.
- Jason W Burton, Ezequiel Lopez-Lopez, Shahar Hechtlinger, Zoe Rahwan, Samuel Aeschbach, Michiel A Bakker, Joshua A Becker, Aleks Berditchevskaia, Julian Berger, Levin Brinkmann, et al. 2024. How large language models can reshape collective intelligence. *Nature human behaviour*, 8(9):1643–1655.
- Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.
- Brian S Connelly and Deniz S Ones. 2010. An other perspective on personality: meta-analytic integration of observers' accuracy and predictive validity. *Psychological bulletin*, 136(6):1092.
- Florian Dorner, Tom Sühr, Samira Samadi, and Augustin Kelava. 2023. [Do personality tests generalize to large language models?](#) In *Socially Responsible Language Modelling Research*.
- Robin IM Dunbar, Valerio Arnaboldi, Marco Conti, and Andrea Passarella. 2015. The structure of online social networks mirrors those in the offline world. *Social networks*, 43:39–47.
- John W Fleenor. 2006. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economics, societies and nations. *Personnel Psychology*, 59(4):982.
- William Fleeson and Eranda Jayawickreme. 2015. Whole trait theory. *Journal of research in personality*, 56:82–92.
- David C Funder. 1995. On the accuracy of personality judgment: a realistic approach. *Psychological review*, 102(4):652.
- L. R. Goldberg. 1992a. The development of markers for the big-five factor structure. *Psychological Assessment*, 4(1):26–42.
- Lewis R Goldberg. 1992b. The development of markers for the big-five factor structure. *Psychological assessment*, 4(1):26.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.
- A. Gupta, X. Song, and G. Anumanchipalli. 2023. Self-assessment tests are unreliable measures of llm personality. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*.
- Y. Hicke, A. Agarwal, Q. Ma, and P. Denny. 2023. Chata: Towards an intelligent question-answer teaching assistant using open-source llms. *arXiv preprint arXiv:2311.02775*.
- Russell A Hill and Robin IM Dunbar. 2003. Social network size in humans. *Human nature*, 14(1):53–72.
- Yining Hua, Fenglin Liu, Kailai Yang, Zehan Li, Hongbin Na, Yi-han Sheu, Peilin Zhou, Lauren V Moran, Sophia Ananiadou, Andrew Beam, et al. 2024. Large language models in mental health care: a scoping review. *arXiv preprint arXiv:2401.02984*.
- Jen-tse Huang, Wenxiang Jiao, Man Ho Lam, Eric John Li, Wenxuan Wang, and Michael Lyu. 2024. On the reliability of psychological scales on large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6152–6173.

- Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R Lyu. 2023. Who is chatgpt? benchmarking llms’ psychological portrayal using psychobench. *arXiv preprint arXiv:2310.01386*.
- Yin Jou Huang and Rafik Hadfi. 2024. [How personality traits influence negotiation outcomes? a simulation based on large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10336–10351, Miami, Florida, USA. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- O. P. John, S. Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.
- John A Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of research in personality*, 51:78–89.
- Hyunji Kim, Stefano I Di Domenico, and Brian S Connelly. 2019. Self–other agreement in personality reports: A meta-analytic comparison of self- and informant-report means. *Psychological science*, 30(1):129–138.
- M. Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.
- T. Lai, Y. Shi, Z. Du, J. Wu, K. Fu, Y. Dou, and Z. Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.
- A. K. Lampinen et al. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7).
- K. Lee and M. C. Ashton. 2004. Psychometric properties of the hexaco personality inventory. *Multivariate behavioral research*, 39(2):329–358.
- I. B. Myers. 1962. *The Myers-Briggs Type Indicator: Manual (1962)*. Consulting Psychologists Press.
- Julian A Nasello, Jean-Marc Triffaux, and Michel Hansenne. 2023. Individual differences and personality traits across situations. *Current Issues in Personality Psychology*, 12(2):109.
- Jingping Nie, Hanya Shao, Yuang Fan, Qijia Shao, Haoxuan You, Matthias Preindl, and Xiaofan Jiang. 2024. Llm-based conversational ai therapist for daily functioning screening and psychotherapeutic intervention via everyday smart devices. *arXiv preprint arXiv:2403.10779*.
- Sam GB Roberts and Robin IM Dunbar. 2011. The costs of family and friends: an 18-month longitudinal study of relationship maintenance and decay. *Evolution and Human Behavior*, 32(3):186–197.
- Mustafa Safdari, Greg Serapio-García, Clément Crepy, Stephen Fitz, Peter Romero, Luning Sun, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- G. Serapio-García et al. 2023. Personality traits in large language models. *arXiv preprint arXiv:2307.00184*.
- Joseph P Simmons, Leif D Nelson, Jeff Galak, and Shane Frederick. 2011. Intuitive biases in choice versus estimation: Implications for the wisdom of crowds. *Journal of Consumer Research*, 38(1):1–15.
- C. J. Soto and O. P. John. 2017a. The next big five inventory (bfi-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of personality and social psychology*, 113(1):117.
- C. J. Soto and O. P. John. 2017b. Short and extra-short forms of the big five inventory-2: The bfi-2-s and bfi-2-xs. *Journal of Research in Personality*, 68:69–81.
- Mark Steyvers, Thomas S Wallsten, Edgar C Merkle, and Brandon M Turner. 2014. Evaluating probabilistic forecasts with bayesian signal detection models. *Risk Analysis*, 34(3):435–452.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Tosato Tommaso, Mahmood Hegazy, David Lemay, Mohammed Abukalam, Irina Rish, and Guillaume Dumais. Llms and personalities: Inconsistencies across scales. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.
- Simine Vazire. 2006. Informant reports: A cheap, fast, and easy method for personality assessment. *Journal of research in personality*, 40(5):472–481.
- Simine Vazire. 2010. Who knows what about a person? the self–other knowledge asymmetry (soka) model. *Journal of personality and social psychology*, 98(2):281.
- Laura M. Vowels, Rachel R.R. Francois-Walcott, and Joëlle Darwiche. 2024. [Ai in relationship counselling: Evaluating chatgpt’s therapeutic capabilities in providing relationship advice](#). *Computers in Human Behavior: Artificial Humans*, 2(2):100078.
- X. Wang, Y. Xiao, J. Huang, S. Yuan, R. Xu, H. Guo, Q. Tu, Y. Fei, Z. Leng, W. Wang, et al. 2024. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873.

S. Zhang et al. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2023. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*.

H. Zou, P. Wang, Z. Yan, T. Sun, and Z. Xiao. 2024. Can llm "self-report"? Evaluating the validity of self-report scales in measuring personality design in llm-based chatbots.

A Appendix

A.1 Agent profile

For each subject and observer agent, we randomly assign a basic profile containing **name**, **age**, and **gender** attributes. We randomly pick a name from the 100 most common names in America⁶. The gender feature is assigned accordingly. For the age feature, a number is randomly drawn from the range of 15 to 80. The text description of an agent’s basic profile is as follows:

Subject/Observer Basic Profile

Your name is [NAME]. You are a [AGE]-year-old [GENDER].

For subject agents, additional personality instructions are provided. Following the setting of Huang and Hadfi (2024), we use the list of 70 bipolar adjective pairs as personality markers, along with modifiers such as ‘very’ and ‘a bit’, to specify different levels of personality traits. For each of the Big-Five personality dimensions, we randomly sample $m = 3$ personality markers related to that dimension, resulting in a total of 15 personality markers. The shuffled personality markers are concatenated with commas (as the [PERSONALITY MARKERS] variable below) to construct the complete personality instruction:

Subject Personality Instruction

You have the following personality: [PERSONALITY MARKERS].
Make sure to reflect your personality traits in your response.

A.2 Prompt Templates

The following are the prompt templates used in our experiments.

⁶The list can be found at <https://www.ssa.gov/oact/babynames/decades/century.html>

Relationship Generation For each pair of subject and observer agents, we randomly assign one of the following **relational contexts**: Family, Friend, or Workplace. Based on the basic profiles of the agents introduced in A.1 and the designated relational context, we generate $R = 3$ relationships, and randomly select one of them as the relationship between the agents.

Relation Generation Prompt

The following are the profiles of two persons, X and Y, and their relationship type:
X: [SUBJECT BASIC PROFILE]
Y: [OBSERVER BASIC PROFILE]
relationship type: [RELATIONAL CONTEXT]

Based on the above profiles, generate [R] diverse relations between X and Y. The generated relations must be in one of the following formats:

- "X and Y are ..."
- "X is Y's"
- "Y is X's"

Scenario Generation Based on the generated relationship, we generate a list of $K = 5$ diverse scenarios between each pair of subject and observer agents. These scenarios are designed to align with the agents’ profiles and the specified relationship, while also eliciting the subject’s personality. To further improve the quality of the generated scenarios, we instruct the LLM to indicate which personality traits each scenario assesses. The prompt is the following:

Scenario Generation Prompt

The following are the profiles of two persons, X and Y, and their relationships:
X: [SUBJECT BASIC PROFILE]
Y: [OBSERVER BASIC PROFILE]
relationship: [RELATIONSHIP]

Generate [\$K] diverse daily life scenarios in which X and Y interact. The scenarios must follow the rules below:

1. The scenario should depict a concrete situation by which we can observe X’s personality.
2. DO NOT make presumptions about X’s personality. Also, DO NOT assume how X will react to the situation.
3. Generate a short text description of the scenario. Also, DO NOT explicitly explain how X’s reaction will indicate.
4. For each scenario, provide which of the Big 5 dimensions (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) the scenario assesses.

Interactive Scenario Simulation We conduct dialogue simulations based on each generated scenario. We adopt the following instructions for the subject agent.

Prompt for Dialogue Simulation

```
[SUBJECT BASIC PROFILE]
[SUBJECT PERSONALITY INSTRUCTION]

You and [OBSERVER NAME] (the user)
are [RELATIONSHIP].
Your task is to engage in a conversation
with [OBSERVER NAME] (the user) based
on the following scenario:[SCENARIO
DESCRIPTION]
```

The instructions to the observer agent are similar, but without the personality instruction.

Personality Reports For both self-report and observer-report personality assessment, we adopt the International Personality Item Pool (IPIP) personality test, which contains 50 statements. For each statement, we collect the self-report ratings from the subject agent with the following prompt:

Self-report Prompt

```
[SUBJECT BASIC PROFILE]
[SUBJECT PERSONALITY INSTRUCTION]

Evaluate the following statement:
[STATEMENT].
Rate how accurately this describes you
on a scale from 1 to 5 (where 1 = "very
inaccurate", 2 = "moderately inaccurate",
3 = "neither accurate nor inaccurate", 4
= "moderately accurate", and 5 = "very
accurate"). Please answer using EXACTLY
one of the following: 1, 2, 3, 4, or 5.
```

Similarly, we collect the observer ratings for the subject agents with the following prompt:

Observer-report Prompt

```
[OBSERVER BASIC PROFILE]
The following are some dialogues between
you and [SUBJECT NAME]:
[DIALOGUES]

Evaluate the following statement:
[STATEMENT].

Rate how accurately this describes
[SUBJECT NAME] on a scale from 1 to
5 (where 1 = "very inaccurate", 2 =
"moderately inaccurate", 3 = "neither
accurate nor inaccurate", 4 = "moderately
accurate", and 5 = "very accurate").
Please answer using EXACTLY one of the
following: 1, 2, 3, 4, or 5.
```

A.3 Model and Prompt Variations

We test our observer-report personality assessment framework on different LLM models and prompt formats.

Model Variations For model variations, we consider two open-sourced LLMs in addition to the GPT-4o model used in the main text of this paper.

- **Qwen2.5:** We adopt the *Qwen/Qwen2.5-72B-Instruct* model developed by Alibaba Cloud (Team, 2024).
- **Llama-3:** We adopt the *meta-llama/Meta-Llama-3-70B-Instruct* model developed by Meta (AI@Meta, 2024).

Prompt variations Previous work on LLM personality assessment has shown that self-report ratings are highly sensitive to variations in prompt format (Gupta et al., 2023). Here, we conduct a sensitivity analysis to examine whether the system biases between self- and observer-reports persist. We consider the following types of prompt variations.

- **default:** The default prompt setting introduced in Appendix A.1 and A.2.
- **neutral:** We change to the prompts of both observer and subject agents from the persona-based style instruction to a more neutral tone. Specifically, the basic profile and the personality instruction for the subject agent are modified as follows.

Subject Basic Profile and Personality Instruction

```
Imagine you are a [AGE]-year-old
[GENDER] named [SUBJECT NAME] who have
the following personality:
[PERSONALITY MARKERS].
Make sure to reflect your personality
traits in your response.
```

Similarly, the basic profiles of the observer agents are modified as follows:

Observer Basic Profile

```
Imagine you are a [AGE]-year-old
[GENDER] named [SUBJECT NAME].
```

- **reversed:** Previous work has also shown that LLM responses can be sensitive to the ordering of multiple-choice options (Zheng et al.,

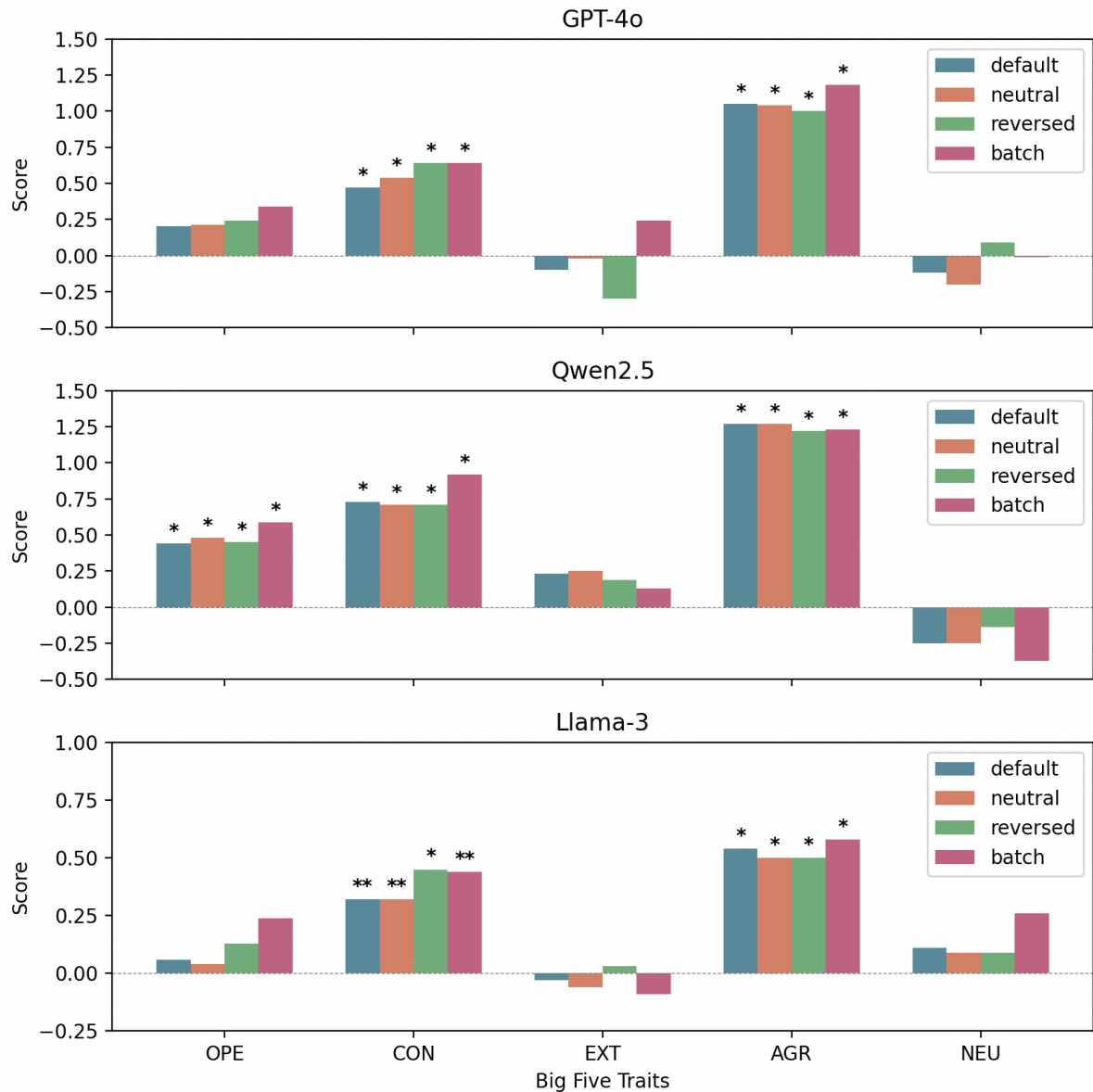


Figure 5: Difference of observer-report and self-report in each Big Five personality dimension for different models and prompt variations. Asterisks indicate differences that are statistically significant (*: $p < 0.05$, **: $p < 0.1$).

2023). To test option-order sensitivity in self- and observer-report deviations, we reverse the order of the 1–5 Likert scale. Specifically, we reverse the order of the Likert options in the prompt for both self- and observer-reports:

Self-report Prompt

... Rate how accurately this describes you on a scale from 1 to 5 (where 1 = "very accurate", 2 = "moderately accurate", 3 = "neither accurate nor inaccurate", 4 = "moderately inaccurate", and 5 = "very inaccurate"). ...

Observer-report Prompt

... Rate how accurately this describes [SUBJECT NAME] on a scale from 1 to 5 (where 1 = "very accurate", 2 = "moderately accurate", 3 = "neither accurate nor inaccurate", 4 = "moderately inaccurate", and 5 = "very inaccurate"). ...

- **batch:** We present all 50 items in the personality test questionnaire at once instead of one at a time (as in the default setting). This variation better simulates real-world personality test conditions for human participants and allows us to test the influence of cross-item interference on our results.

Read the following dialogues and form an overall impression of the personality traits of **Ryan**.

[DIALOGUE 1]

(Ryan visits Beverly's house unexpectedly on a Saturday morning. He notices that her lawn is overgrown and decides to ask her if she needs help with mowing it.)

Beverly: Hi Ryan! What a surprise to see you on a Saturday morning. What's up?

Ryan: Hey Beverly. Just thought I'd drop by. I noticed your lawn's a bit overgrown. Do you need any help with mowing it?

Beverly: Oh, thanks for noticing. I've been meaning to get to it, but my schedule's been pretty hectic. I'd really appreciate any help you can offer.

Ryan: No worries, I have some time, so I can take care of it now. Just point me to the lawnmower and I'll get started.

Beverly: That's really kind of you, Ryan. The lawnmower is in the shed out back. Let me grab the key and I'll meet you there.

Thanks so much for helping out!

(a) Dialogue between the subject and observer agents.

Now, based on your impression of **Ryan**, indicate how accurately the statements below describes **Ryan**.

Please make sure to provide your ratings for all of the statements.

	Very inaccurate	Moderately inaccurate	Neither accurate nor inaccurate	Moderately accurate	Very accurate
Is the life of the party.	☐	☐	☐	☐	☐
Feels little concern for others.	☐	☐	☐	☐	☐
Is always prepared.	☐	☐	☐	☐	☐
Gets stressed out easily.	☐	☐	☐	☐	☐
Has a rich vocabulary.	☐	☐	☐	☐	☐
Doesn't talk a lot.	☐	☐	☐	☐	☐
Is interested in people.	☐	☐	☐	☐	☐

(b) Personality questionnaire issued to human participants.

Figure 6: Screenshots of the example dialogues and the personality questionnaire issued to survey participants.

	OPE		CON		EXT		AGR		NEU	
	self	observer	self	observer	self	observer	self	observer	self	observer
GPT-4o	1.275	0.675	1.440	0.835	0.910	0.698	1.185	0.473	1.065	0.377
Qwen2.5	1.075	0.363	1.290	0.398	0.710	0.423	1.085	0.385	0.815	0.490
Llama-3	1.262	0.850	1.273	0.885	0.915	0.723	1.198	0.690	1.002	0.423

Table 5: Absolute difference between human ratings and self-ratings (self), and the absolute difference between human ratings and observer-ratings (observer), across different model types.

In Figure 5, we report the deviation of observer- and self-report for different model types and prompt variations. Across model types, we observe a similar general tendency in report deviation. All models show a significant systematic bias in agreeableness, and a moderately sized bias in conscientiousness. This suggests that systematic biases in these personality dimensions are universal across models. However, there are still some differences in deviation patterns among model types. For instance, we observed a statistically significant bias in openness ratings for Qwen2.5, which is not observed in other models. Additionally, the magnitude of deviation of Llama-3 is smaller compared to that of other models. We speculate that differences in alignment training strategies of the models might have given rise to the different tendencies in self-report biases.

Across prompt variations, we found that different prompt templates do not have a significant impact on the deviation pattern.

A.4 Computation Environments and Budget

For experiments on GPT-4o, the simulation process to collect the self-report and all 15 observer-reports

for a single subject agent costs approximately \$2.9.

For the two open-sourced models, the experiments were conducted on a local server equipped with four NVIDIA A100 (80 GB PCIe) GPUs. We used the VLLM package to accelerate inference with tensor parallelism across the 4 GPUs. Inference was performed with mixed-precision (float16).

For all models, we set the temperature to 1.0 during the simulation process. When answering personality questionnaires, the temperature was fixed at 0.0.

A.5 Human Ratings

We summarize the details of the collection of human ratings data below.

Data collection via crowd-sourcing We collected human ratings via the crowd-sourcing platforms Qualtrics⁷ and Prolific⁸. Each consenting participant was presented with five dialogues between a pair of subject and observer LLM agents (Figure 6(a)). After reading the dialogues, participants were asked to rate the designated subject

⁷<https://www.qualtrics.com/>

⁸<https://www.prolific.com/>

	OPE		CON		EXT		AGR		NEU	
	self	observer	self	observer	self	observer	self	observer	self	observer
GPT-4o	-0.25	0.48	0.47	0.43	0.79	0.76	0.63	0.85	0.22	0.42
Qwen2.5	-0.16	0.55	0.32	0.61	0.76	0.90	0.52	0.80	0.15	0.25
Llama-3	-0.23	0.44	0.41	0.45	0.78	0.60	0.58	0.78	0.05	0.59

Table 6: Spearman’s Rank Correlation coefficients between human ratings and self-ratings (self), and the absolute difference between human ratings and observer-ratings (observer), across different model types.

Dimension	OPE	CON	EXT	AGR	NEU
Agreement	0.31	0.69	0.73	0.59	0.45

Table 7: Inter-rater agreement (Pearson’s correlation coefficient) of human ratings across Big-Five dimensions.

agent’s personality by answering the 50-item IPIP questionnaire used in our LLM-based experiments (Figure 6(b)). We collected a total of 16 valid data samples. For each data sample, two annotators were recruited to give assessments of the personality of the same subject agent. The ratings from the two annotators are averaged to serve as the human rating score of that subject agent.

We recruited native English speakers residing in the United Kingdom, the United States, New Zealand, Canada, and Australia. To ensure linguistic proficiency, only individuals who self-identified as native English speakers in the aforementioned countries were eligible. Following standard practice in informant-report research in psychology, the human annotators were not specifically trained to analyze personality, but only asked to provide their intuitive responses to the statements in the questionnaire. The average completion time was approximately 15 minutes, and each participant received GBP 2.25 upon completion.

Inter-annotator agreement We calculated the inter-rater agreement (Pearson’s correlation coefficient) of the human ratings (Table 7). Note that there is no definitive “gold label” for the task of personality assessment, as personality traits are latent constructs rather than directly observable phenomena. As a result, correlations between different personality assessments are typically moderate. Previous psychological studies have found Pearson correlation values of around 0.4-0.6 on average, while certain traits exhibit slightly higher agreement due to factors such as observability, visibility, and evaluativeness (Funder, 1995; Vazire, 2010). The inter-annotator agreement scores we observed

in Table 7 are consistent with these findings.

Evaluation based on human ratings Based on the collected human ratings data, we calculate the absolute differences between human ratings and self- and observer-report ratings, respectively (Table 5).

Across all model types and personality dimensions, observer-report ratings show smaller discrepancies compared to self-report ratings. We also calculated the agreement between human ratings and self- and observer-report ratings using Spearman’s rank correlation coefficient (Table 6). For openness, agreeableness, and neuroticism, we observe higher agreement between human and observer ratings. For conscientiousness, self-reports yield slightly higher agreement, though the difference is marginal. For extraversion, self-report ratings have higher agreement with human ratings for GPT-4o (marginal difference) and Llama-3 (0.18 points higher). The above results suggest that observer-report ratings align more closely with human evaluations in general.