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

Anonymous ACL submission

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 agents. Each observer is configured with a specific relational context (e.g., family member, friend, or coworker) and engages the subject LLM in dialogue before evaluating its behavior across the Big Five dimensions. We show that these observer-report ratings align more closely with human judgments than traditional self-reports and reveal systematic biases in LLM self-assessments. We also found that aggregating responses from 5 to 7 observers reduces systematic biases and achieves optimal reliability. Our results highlight the role of relationship context in perceiving personality and demonstrate that a multi-observer paradigm offers a more reliable, context-sensitive approach to evaluating LLM personality traits.

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). Also, 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, self-report questionnaires are the most commonly used, assessing personality through individuals' responses to standardized questions about their thoughts, emotions, and behaviors. Existing LLM personality assessment methods also rely heavily on self-report questionnaires, in which an LLM is prompted to answer a set of standardized questions. Despite their extensive use, researchers have raised questions about the reliability of using self-reports for LLM personality assessments, particularly in maintaining stable personality traits in different contexts (Gupta et al., 2023; Dorner et al., 2023; Wang et al., 2024). There are also 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 the LLM's test results remains unclear.

In this work, we explore alternative methodologies for LLM personality assessments. We start from the intuition that others (such as friends, family, or colleagues) can provide valuable insights into a subject's personality traits, sometimes even more accurate than the subject themselves. In psychology, this intuition leads to the personality assessment method of informant-report (Vazire, 2006). Instead of using the self-report rating of the subject, informants other than the subject are asked to give ratings of the subject's personality. In this fashion, it is possible to obtain a more objective assessment of the subject's personality profiles.

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).[1] For a given subject agent, we first prepare N observer agents, each assigned a specific

[1] The code & data will be made available upon publication.

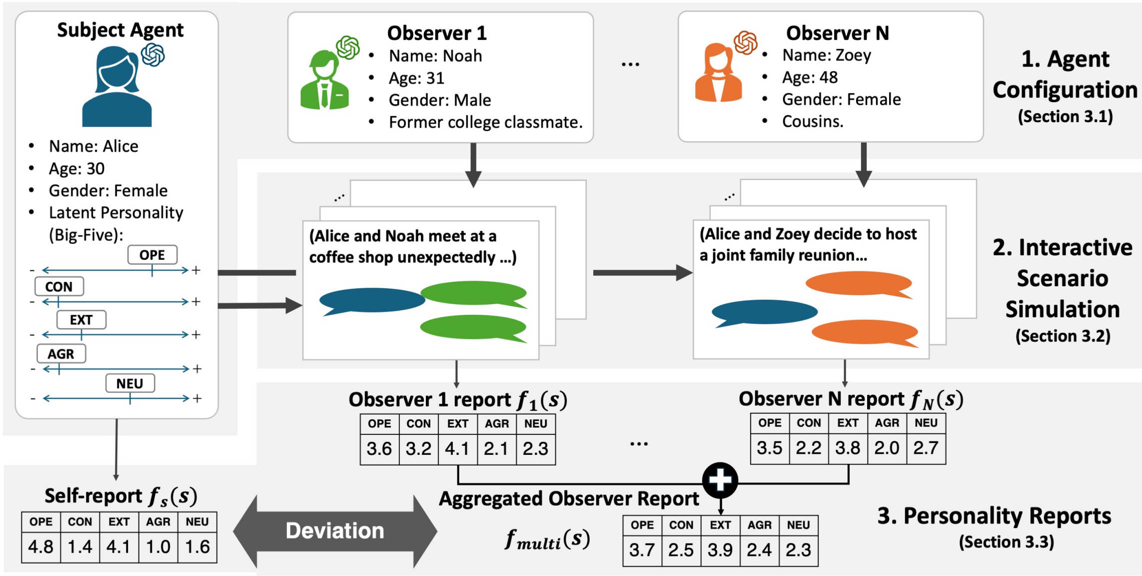


Figure 1: Overview: multi-observer LLM agents for Big Five personality assessment.

relationship with the subject, such as college classmates or cousins. Based on this relationship, an array of interactive scenarios is generated automatically. For each scenario, we perform a simulation in which the subject and the observer agents engage in a dialogue based on the scenario. After that, each observer agent is instructed to take a questionnaire and give ratings on the subject's personality from their perspective based on the dialogues obtained during the simulation process. Finally, all observer reports are aggregated to give a final collective assessment of the subject.

The experiments showcase that observer-report ratings align better with human ratings, while the self-report ratings reflect a high correlation with the injected personality prompt instead of real behaviors. Further, we empirically show that LLMs possess systematic biases in self-reporting their personality on some personality dimensions. We also analyze the influence of the number of observers (N) and the relationship between the observer and the subject. The analysis reveals the effectiveness of aggregating multiple observer responses to yield more robust personality ratings.

2 Related Work

The *Big Five Inventory* (BFI) remains the most commonly used framework for evaluating personality traits, capturing the traits of openness, conscientiousness, extraversion, agreeableness, and neuroticism (John et al., 1999). Several variants of the BFI have also been utilized, including the BFI-2 (Soto and John, 2017a), BFI-2-XS (Soto and John, 2017b), and IPIP-NEO-120 (Johnson, 2014). Other prominent frameworks include the *Myers-Briggs Type Indicator* (MBTI) (Myers, 1962), the *HEXACO Personality Inventory* (Lee and Ashton, 2004), and Goldberg's bipolar adjective markers (Goldberg, 1992a). Researchers have also investigated multi-rater assessments of personality traits. For instance, (Connelly and Ones, 2010) showed how accurately various observers (e.g., friends, family members, coworkers) could rate an individual's personality traits and how these ratings compare to self-assessments. Similarly, (Mount et al., 1994) explored the relationship between coworkers' personality ratings and job performance. Furthermore, (Vazire, 2010) investigated which traits are better judged by the self versus others, providing insights into the conditions under which external observations might outperform self-reports.

There is evidence for an alignment between self- and other-reports of personality. 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. Importantly, moderate discrepancies emerged only when the informants were strangers, implying that people tend to be more critical of individuals they do not know well. These results have important implications for personality assessment and contexts where self-enhancement motives may play a role.

A comprehensive meta-analysis by (Connelly and Ones, 2010) underscores the advantages of incorporating other-reports alongside self-reports for a richer and more predictive understanding of personality. Their findings indicate that accuracy in other-reports varies by trait. That is, extraversion and conscientiousness are rated most accurately, while emotional stability (neuroticism) and agreeableness are more difficult to identify, particularly for observers who lack familiarity with the target. One explanation for lower accuracy in agreeableness lies in its high evaluativeness, which can make ratings more subjective. Moreover, family and friends provide the most accurate other-ratings, while coworkers, despite frequent interactions, tend to offer less accurate assessments.

Regarding predictive validity, (Connelly and Ones, 2010) also found that other-ratings often outperform self-ratings in predicting academic and job performance; multiple other-raters further enhance predictive power relative to a single informant. In contrast, self- and other-ratings exhibit comparable validity for forecasting first impressions. This led to the conclusion that there is "extraordinary value" in collecting other ratings of personality while emphasizing the importance of using multiple raters to mitigate individual bias, the importance of well-acquainted observers, the subtlety in evaluating less visible traits, and the importance of specific context and purpose of the assessment. 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 persona 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 of personality assessment. We first introduce the configuration of the subject and observer agents in Section 3.1. We then utilize these agents to simulate scenarios (Section 3.2) before observers give the personality assessment on the subject (Section 3.3).

3.1 Agent Configuration

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

Agent Profile Each agent is given a randomly generated basic profile that contains **Name**, **Age**, and **Gender**. Setting specific agent names facilitates smooth interaction in the simulation phase, and incorporating random age and gender of the agents results in a greater diversity of observers. For both subject and object agents, a randomly generated basic profile is assigned.

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 5-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 is an integer within the range of $[1, 6]$, indicating the strength level of a corresponding Big Five personality dimension d.

Following previous work, we construct a personality instruction of ψ based on personality markers (Serapio-García et al., 2023). For each dimension d, we pick m personality markers that reflect the personality strength ψ_s^d. For instance, if the agent s has an extraversion trait of strength $\psi_s^{EXT} = 2$, which is on the lower side of the spectrum, some possible choices of personality markers are "timid, silent, 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 conducted by individuals who have a close relationship with the subject. Here, we follow Kim et al. (2019) and generate relationships within one of the following **relation contexts**: Family, Friend, or Workplace. Based on the agent profiles and a designated relation type, an inter-agent relationship that matches

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

the profiles is generated. Here, we utilize a separate LLM to generate the relationships automatically.

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. Based on this motivation, we conduct simulations of a diverse set of scenarios involving the subject and the object agent to elicit their 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. In the following, we summarize the process.

Scenario Generation We next generate a set of K scenarios involving the subject and observer based on their relationship. Specifically, we generate diverse scenarios that can elicit behaviors of the subject agent that signals 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 Further, we perform a simulation in which the subject and the observer agents engage in a dialogue based on the generated scenarios. For each scenario, we first 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 always 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 utterance, the agents are asked to specify whether the dialogue is over or if they wish to leave the conversation. The simulation is terminated if both agents reply that the dialogue should be over.

3.3 Personality Reports

We utilize a questionnaire to assess the personality of a specific 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 (Huang and Hadfi, 2024).

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", "sympathize with others' feelings". For each statement, the subject agent s is instructed to rate how accurately the statement describes itself using a 5-Likert scale (from "1 = very inaccurate" to "5 = very accurate"). See Appendix A.2 for details. 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 rating score of its related statements as the final assessment score $f_s^d(s)$, which is the dth dimension of the subject's self-report $f_s(s)$.

Individual Observer-Report We obtain the subject's personality assessment from the perspective of 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 the case of observer reports, the observer agent is asked to rate how each statement fits the description of the subject agent. To get the assessment from the perspective of the observer agent, the dialogues generated from 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 ratings. In this work, we simply take the average value of the observer reports as $f_{multi}(s) = \frac{1}{N}\sum_{n=1}^{N} f_n(s)$. This aggregated observer report reflects the collective reports 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 adopt GPT-4o as LLM (Hurst et al., 2024). We

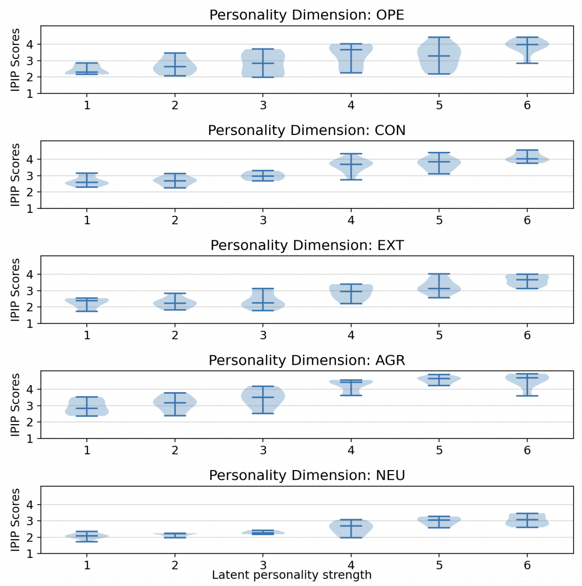


Figure 2: Observer report ratings across different latent personality strength levels.

also conducted experiments based on Qwen2.5 and Llama-3, found in Appendix A.3.

Scenario Generation We use the GPT-4o model to generate inter-agent relationships and scenarios. We perform scenario simulation based 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. For each pair of subject and observer agents, we generate $K = 5$ scenarios and conduct the simulation.[3]

Personality Questionnaire We adopt the International Personality Item Pool (IPIP) personality test (Goldberg, 1992b), which is a widely used personality inventory designed for assessing the Big Five personality traits. The questionnaire consists of $M = 50$ statements, with each statement related to one of the Big Five personality dimensions.[4]

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 an increasing tendency of observer rating scores (**IPIP Scores**) with the increase of latent personality strength level.

	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

Correlation with Latent Personality Profiles We compare the self- and observer-report ratings in terms of their correlation with the latent personality strengths in each Big Five dimension. We precisely define the **latent-observer** case as correlating the subject's latent personality ψ_s with the aggregated observer report $f_{\text{multi}}(s)$. On the other hand, the **self-observer** case correlates the subject's self-report $f_s(s)$ with the aggregated observer report $f_{\text{multi}}(s)$. The first row of Table 1 shows the resulting correlation values. For all personality dimensions, the self-report ratings show a correlation coefficient exceeding 0.9, showing a near-perfect positive correlation with the latent profiles. On the other hand, the correlation strength between observer-report ratings and the latent personality strength levels is lower.

Correlation with Human Ratings We crowdsourced a small-scale human rating dataset in which human annotators are asked to provide their personality ratings of a subject agent based on the dialogues between the subject and observer agents.[5] We then calculated the agreement between the human and self- and observer-report ratings, respectively. The lower section of Table 1 shows the Spearman's rank correlation coefficients. For the openness, agreeableness, and neuroticism dimensions, the observer-report ratings correlate more with human ratings. While self-reports correlate more with human ratings for conscientiousness and extraversion, the difference is marginal. Notably, self-report exhibits a negative correlation for openness, implying a reversal in rank ordering relative to human ratings. With these results, we verify the validity of the proposed observer-report personality assessment method, which aligns better with human ratings than the self-report method. The results also raise concerns about the reliability of the

[3] Refer to Appendix A.2 for the details of the prompts.

[4] The list of 50 items and the scoring schemes can be found at `https://ipip.ori.org/newBigFive5broadKey.htm`.

[5] See Appendix A.5 for the data collection process.

self-report method. The relatively lower correlation with human ratings indicates that the near-perfect correlation with the latent personality profile only reflects the personality instruction prompts, but not the actual behavior of the agent.

5.2 Impact of Multiple Observers

We now examine how the number of observers influences the 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 1 to 15, for each Big Five dimension. Conscientiousness improved with additional observers, with both correlation types increasing sharply to five observers. Self-observer correlations stabilized at 0.90, while latent-observer correlations stabilized at 0.85, making it the personality trait with the highest agreement. Extraversion and agreeableness demonstrated an increased correlation pattern with additional observers before both correlation values converged at 0.85. Neuroticism displayed the most interesting convergence pattern, initially variable, with correlations reaching similar levels of 0.85 for both correlation types when incorporating around 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 rate, 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, which is 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. Conscientiousness in particular yields the highest self–observer agreement (Connelly and Ones, 2010). In contrast, openness is associated with internal characteristics such as imagination, aesthetic sensitivity, curiosity, etc. Due to the low trait visibility, openness often demonstrates the lowest agreement (Vazire, 2010). Our results align with these findings, with the high agreement for conscientiousness and low agreement for openness. However, for neuroticism, another trait with lower trait visibility, we do not identify a lower self-observer agreement.

	OPE	CON	EXT	AGR	NEU
Mean Deviation	0.20	**0.39***	-0.12	**0.91***	-0.19
Cohen's d (LLM)	0.24	0.46	-0.13	1.07	-0.26
Cohen's d (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.

Wisdom of the Crowd Another finding is that the benefits of additional observers generally stabilize after 5-7 observers. Below 5-7 observers, we observe something akin to the wisdom of the crowd effect, where groups outperform top individuals by pooling diverse insights and aggregating responses from multiple sources to generate superior outcomes compared to relying on a single model (Burton et al., 2024; Guo et al., 2024; Fleenor, 2006). This insight sets 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 as an optimal number for practical assessment purposes. 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 five individuals on average (Dunbar et al., 2015; Hill and Dunbar, 2003; Roberts and Dunbar, 2011). This observation aligns with research showing that intimate bonds (family, close friends, or trusted colleagues) offer the most revealing insights into an individual's personality. Real-world social structures demonstrate that a core group of just a few relationships is sufficient to capture the most profound knowledge of a person's traits.

5.3 Self-Observer Deviations

We calculate the differences between aggregated self- and observer-report scores. We also identify systematic biases between self- and observer-report ratings via mean deviation $\frac{1}{N}\sum_{s=1}^{N} f_{multi}(s) - f_s(s)$. Positive values indicate higher observer ratings than self-ratings, and negative values indicate lower observer ratings. Zero or near-zero values imply close agreement. The non-systematic biases introduced by individual observers will be averaged out in the statistical aggregation process (Simmons et al., 2011; Steyvers et al., 2014). This phenomenon could be linked to the wisdom of the crowd phenomenon in which aggregating multiple independent judgments often produces esti-

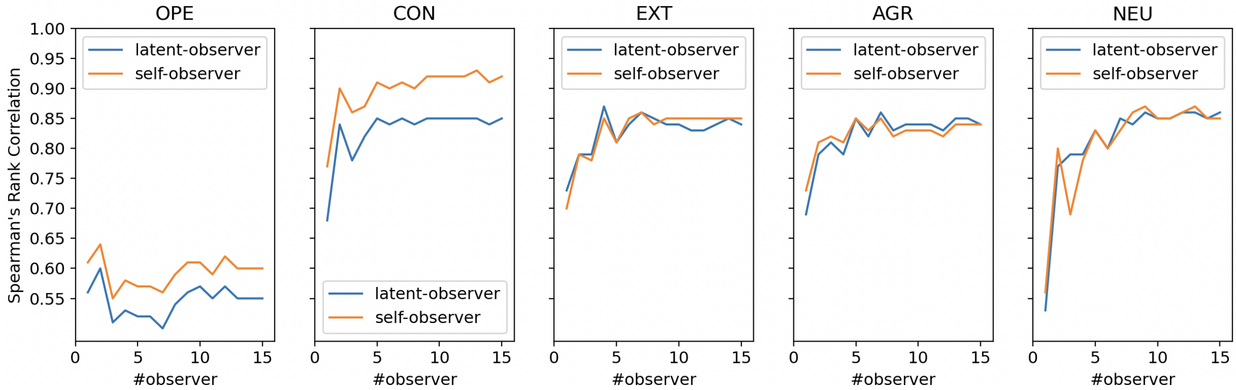


Figure 3: Spearman's Rank Correlation coefficients between latent-observer and self-observer ratings as a function of the number of observers for each Big Five trait.

mates superior to individual experts, as random errors tend to cancel each other out when combined (Fleenor, 2006). On the other hand, the mean deviation between the observer- and self-report shows the systematic biases that cannot be averaged out in the aggregation process. We note systematic patterns in how observers perceive personality traits compared to self-perceptions (Table 2). Agreeableness, in particular, shows the most significant positive deviation (0.91 point) and the widest spread of ratings, indicating that observers consistently rate individuals as more agreeable than they rate themselves, though with substantial variability in the magnitude of this difference. Conscientiousness also demonstrates moderate positive deviations (0.39), suggesting a consistent tendency for observers to rate these traits slightly higher than self-ratings. In contrast, openness, extraversion, and neuroticism exhibit minor mean deviations with magnitudes no larger than 0.2, indicating that almost no systematic bias exists for these personality dimensions.

Statistical Significance Test Further, we conduct a paired-samples t-test to examine the difference between the self- and observer-report ratings of each Big Five personality dimension. We also calculate the 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, we identify statistically significant differences for the AGR and CON traits ($p < 0.05$). Specifically, the self-report ratings for agreeableness (AGR) are significantly lower than the observer-report ratings with a large effect size ($d = 1.07$), which indicates that the system bias is larger than one full standard deviation. Conscientiousness trait also illustrates a statistically significant deviation between self- and observer-report ratings, but is more moderate in effect size ($d = 0.46$).[6] On the other hand, we do not identify significant systematic biases for the other three personality dimensions.

Comparison to Human Results We compare the deviation between self- and observer-reports based on our simulations to previous findings. Kim et al. (2019) conducted a meta-analysis to reveal the mean deviation between human self-report personality and observer-report personality (also called informant-report) of individuals with close relationships. Their work shows that there is only a slight deviation between self- and observer reports, with an effect size smaller than $d < 0.27$ across all Big Five dimensions. This aligns with our results for the personality dimensions of openness, extraversion, and neuroticism. On the other hand, we found that the LLM subject agents possess a significant systematic bias in self-reporting lower agreeableness and conscientiousness scores. Considering that systematic biases exist only in the case of LLM agents but not humans, we speculate that these biases might originate from the alignment training phase. The alignment training encourages LLMs to act according to users' preferences, resulting in an inherent bias in their self-reported personality. The result also suggests that LLM gives self-report personality assessments based on the personality instruction prompts (such as the personality markers used in this experiment), and the scores might not reflect the actual behavior of the agent.

[6]For Cohen's d, values around 0.2 indicate a small effect, 0.5 a medium effect, and 0.8 a significant effect (Cohen, 2013).

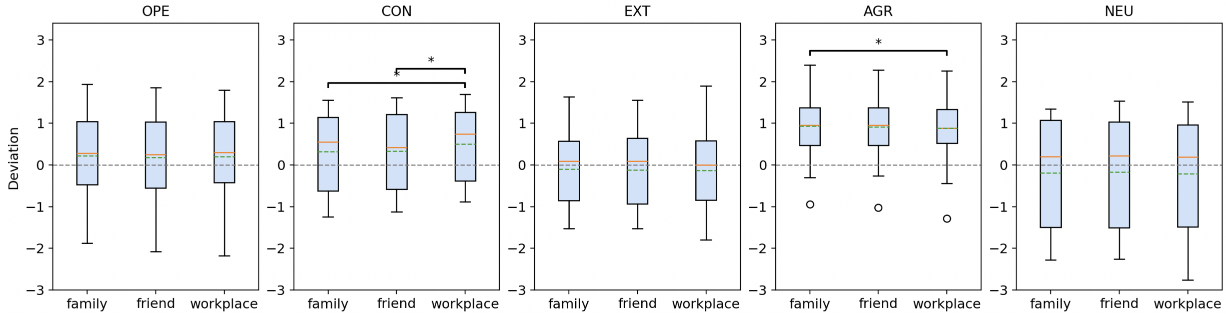


Figure 4: Mean differences between observer and self-reports across Big Five personality traits by relationship context. The orange line represents the median, while the green dotted line shows the mean. Relationships with statistically significant differences ($p - value < 0.05$) are highlighted with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

5.4 Impact of Relationship Type

We now look at the mean deviations between self-report and observer-report ratings under three relationship contexts as illustrated in Figure 4. We observed significant differences between the ratings for conscientiousness and agreeableness across different relationship types. These patterns suggest that relationship context particularly influences the perception of these traits. Conscientiousness showed significant differences between workplace and friend/family ratings. Observer agents in a workplace relationship scenario tend to give slightly higher conscientiousness ratings than observers of a family or friend relationship with the subject agent. Similarly, we observed differences between workplace and family ratings for agreeableness. Specifically, observer agents in a family relationship tend to give higher agreeableness ratings than observers in a workplace relationship. On the other hand, other personality dimensions demonstrated consistency across relationship types, with no statistically significant difference in ratings across different relationship types.

Context-Dependency of Personality Observers in different relational contexts assign different importance to specific traits. In particular, workplace observers showed distinctly different rating patterns compared to family and friends, particularly for conscientiousness and agreeableness. This divergence likely reflects the multi-faceted and context-dependent nature of personality. While an individual's personality is generally considered stable, the manifestation of personality may vary across different social contexts (Fleeson and Jayawickreme, 2015). An extroverted person, for example, may act even more outgoing in front of close friends but adopt a more reserved behavior at work. Since the observer reports are influenced by the different facets of the person, it is common to observe inconsistency in observer ratings across observers. 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 expect it to be easy to concentrate and so may increase his or her level of conscientiousness (Fleeson and Jayawickreme, 2015; Nasello et al., 2023).

Finally, we found discrepancies in agreeableness and conscientiousness, indicating a higher degree of context dependence 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.

6 Conclusions

The study introduced a novel multi-observer framework for personality assessment in LLMs. We verify the validity of our proposed observer-report method, which yields more robust and context-sensitive personality evaluations than traditional self-report methods. Further analysis demonstrates the effectiveness of aggregating responses from multiple observer agents, which mitigates individual biases and yields more robust evaluations. robust and context-sensitive personality evaluations than traditional self-report methods. Our experiments reveal that relationship 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 relationship 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 counselling (Vowels et al., 2024), mental healthcare (Hua et al., 2024), mental therapy (Nie et al., 2024), teamwork (Arukgoda et al., 2023), where AI must adapt to individuals' varying personality traits.

References

AI@Meta. 2024. Llama 3 model card.

Anupama Arukgoda, 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 Xiangliang 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.

Michael K Mount, Murray R Barrick, and J Perkins Strauss. 1994. Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology*, 79(2):272.

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 Dumas. 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*.

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 pick a name from the 100 most common names in America[7]. The gender feature is assigned accordingly. For the age feature of the agents, a number is randomly selected from the range of 15 and 80.

For subject agents, additional personality instructions are also provided. We follow the setting of (Huang and Hadfi, 2024) and use the list of 70 bipolar adjective pairs proposed in the Big Five personality theory along with modifiers like 'very', 'a bit' to set different levels of personality traits.

In this fashion, we construct the text description of an agent's basic profile as follows:

Subject Agent Instruction

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

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

Observer Agent Instruction

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

A.2 Prompt Templates

The prompt for the relationship extraction is the following:

Relation Generation Prompt

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

Generate [$N] diverse [RELATION TYPE]
relations between X and Y. The generated
relations must be in the following format:
"X and Y are ..."
```

[7]The list could be found in https://www.ssa.gov/oact/babynames/decades/century.html

The prompt for the scenario extraction 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 where we can observe X's
personality.
2. DO NOT make presumptions about X's
personality in the scenario.
3. Generate a short text description of
the scenario. For each scenario, also
provide which of the Big 5 dimensions it
assesses.
```

During scenario simulation, we adopt the following instruction for the subject agent.

Simulation instruction

```
[SUBJECT BASIC PROFILE]
[LATENT PERSONALITY PROFILE]

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

The observer agent's instruction is similar but without the personality profile.

Finally, the prompt for self-report personality assessment and observer-report assessment is as follows:

Prompt for self-report

```
[SUBJECT BASIC PROFILE]
[LATENT PERSONALITY PROFILE]

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.
```

```
Prompt for observer report

[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 Variations of LLM Models and Prompt Formats

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

Model Variations For model variations, we consider the two other open-sourced models 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 works on LLM personality assessment reveal that self-report ratings are highly sensitive to variations in prompt phrasing. Here, we conduct a sensitivity analysis to observe whether the system biases between self- and observer-reports persist. We consider the following type of prompt variations.

- **default**: the default prompt setting introduced in Appendix A.1 and A.2.
- **neutral**: We introduce variation in the prompt of the subject agent instruction and the observer agent instruction. Specifically, we convert the persona-based style instruction to a more neutral tone as follows.

```
Subject Agent 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.
```

```
Observer Agent Instruction

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

- **reversed**: Previous work has also found that LLM responses can be sensitive to the ordering of multiple-choice options. Thus, we try to reverse the order of the 1-5 Likert scale to observe the option-order sensitivity of the deviation between self- and observer-report ratings. Specifically, we reverse the order of the Likert options in the prompt for self- and observer reports:

```
Prompt for self-report

...
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"). ...
```

```
Prompt for observer report

... 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 variable better simulates the real-world personality test settings of human participants and can test the influence of cross-item interference on our results.

In Figure 5, we report the deviation of observer-report and self-report of 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 slightly moderate level of bias in conscientiousness.

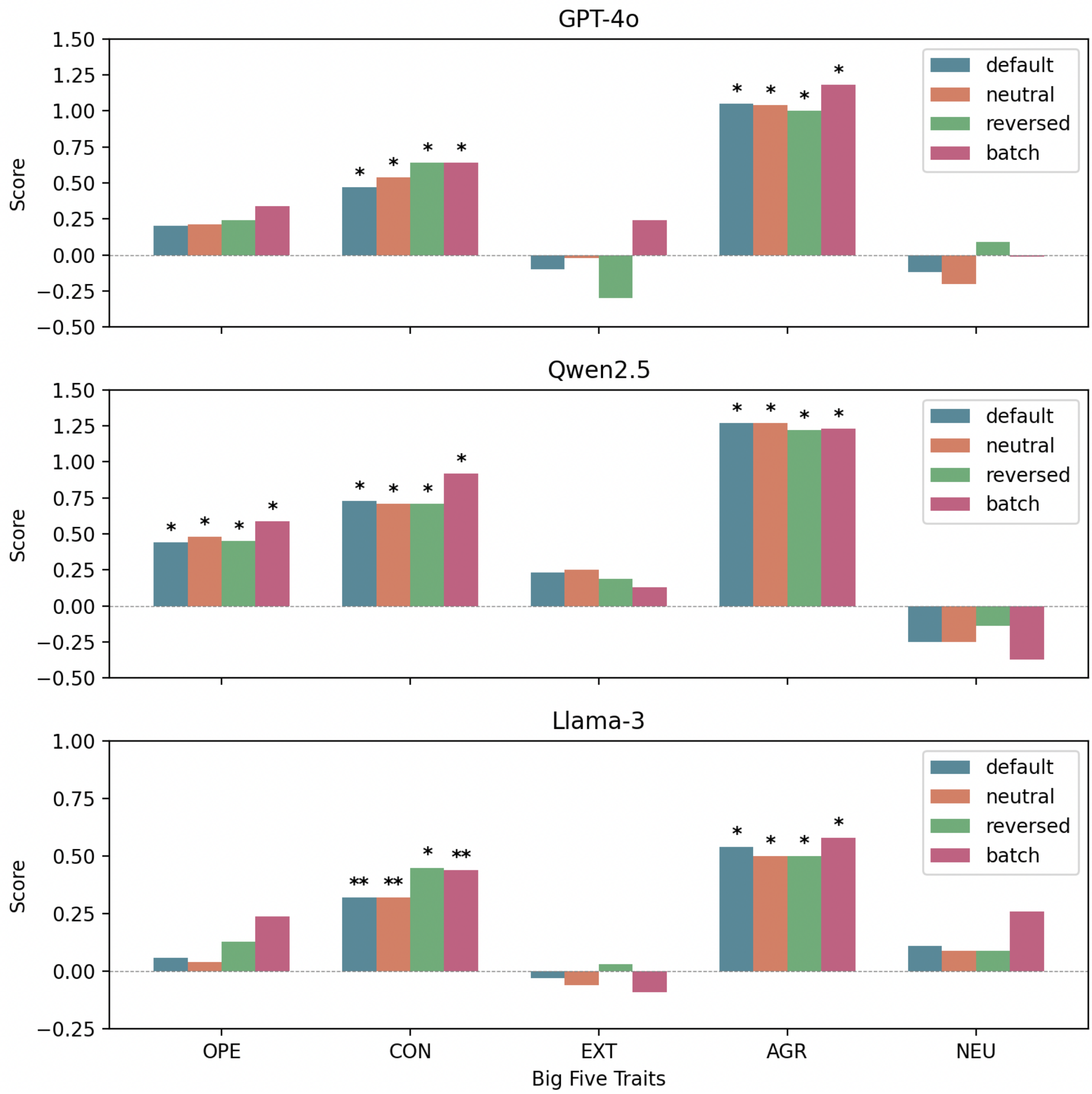


Figure 5: Difference of observer-report and self-report in each Big Five personality dimension for different models and prompt variations. Asterisks indicate differences that are statistically significant (*: $p < 0.05$, **: $p < 0.1$).

This indicates that the systematic biases in these personality dimensions are universal across model types. However, there are still some differences in deviation patterns among models. For instance, we also observe a statistically significant bias in openness ratings for Qwen2.5, which is not observed in other models. Also, we found that the magnitude of deviation of Llama-3 is smaller compared to other models. We speculate that the difference in the training process might have given rise to the slight difference.

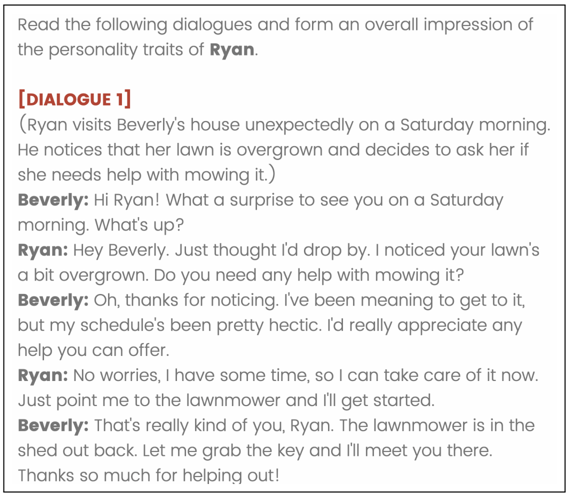
Across different variations of prompts, we found that using different prompt templates does not have a big impact on the deviation pattern.

A.4 Computation Environments and Budget

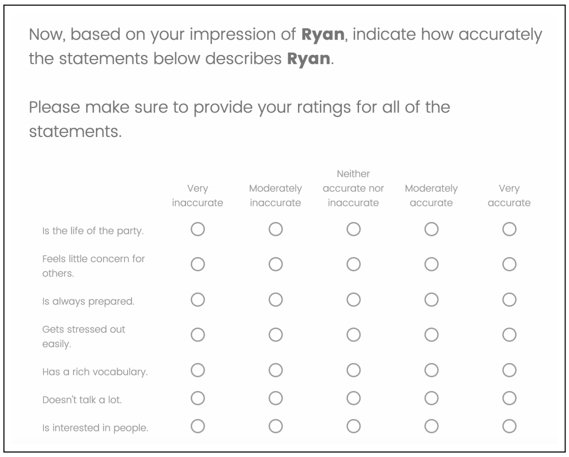
For the two open-sourced models, the experiments were conducted on a local server equipped with 4 NVIDIA A100 GPUs (80 GB PCIe) cards. Also, we use the VLLM package to accelerate inference with tensor parallelism across 4 GPUs. We used mixed-precision (float16) inference.

For GPT-4o, the simulation process to collect the self-report and observer-reports for one subject agent costs around 2.9$.

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



(a) Dialogue between the subject and observer agents.



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

A.5 Human Ratings

Data collection via crowd-sourcing We collected human ratings via crowd-sourcing platform Qualtrics[8] and Prolific[9]. Each consenting participant was presented with five dialogues between a subject and observer LLM agents (See Figure 6(a)). After reading the dialogues, the participant is asked to rate the designated subject 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.

We recruited 16 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. The average completion time was approximately 15 minutes, and each participant received GBP 2.25 upon complete submission.

In addition to detecting biases in self-reports, this work could detect biases across LLMs. We hypothesize that due to their different alignment training strategies, they exhibit different tendencies regarding trait clusters such as (AGR, OPE, CON), EXT, and NEU.

Evaluation based on human ratings Based on the collected human ratings data, we calculate the absolute difference between human and self- and observer-report ratings, respectively (Table 3). We can see that for all model types and personality dimensions, the observer-report ratings show a smaller discrepancy compared to self-report ratings.

We also calculate the agreement between human and self- and observer-report ratings, based on Spearman's rank correlation coefficient (Table 4). For the personality dimensions of openness, agreeableness, and neuroticism, we observe a higher agreement between human and observer ratings. For conscientiousness, we observe a higher agreement for self-report, but the difference is marginal. For extraversion, self-report ratings result in a higher agreement in the case of GPT-4o (marginal difference) and Llama-3 (0.18 points higher).

Based on the results above, we can conclude

[8] https://www.qualtrics.com/
[9] 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 4: 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.

that observer report ratings align better with human evaluation in general.

A.6 Case Study

In this section, we conducted a case study of our proposed method. Table 5 presents an example of the simulation dialogue between a pair of subject and observer agents, *Ethan* and *Jacob*, respectively. Starting from the automatically generated agent profiles and the relationship between them, scenarios are generated automatically. Then, we conduct a dialogue simulation of the interaction between the agents under these scenarios. For space limitations, we only present one example of the scenarios and its corresponding simulated dialogue. After the scenario simulation process, the observer agent is asked to give ratings on the subject agent's personality.

Table 6 shows the latent personality level given as instructions to the subject agent and the self- and observer-report scores. Specifically, the **Individual Observer-report** scores are the ratings the observer Jacob offers, and the **Aggregated Observer-report** scores are the average scores of all observers' ratings.

Table 5 and 6 present an intriguing case where an agent is assigned a "very disagreeable" personality (with the lowest AGR level of 1). We can see that the self-report score given by the subject agent itself gives a low agreeableness score (1.7), where the observer agents give a moderate agreeableness score (3.0). However, by observing the simulated dialogue in Table 5, we can see that the real behavior manifested by the subject agent *Ethan* does not reflect such extreme disagreeableness quality suggested by the self-report score. This discrepancy underscores the importance of examining behaviors beyond self-report questionnaires.

Subject Profile	`{name: Ethan, age: 29, gender: male}`
Observer Profile	`{name: Jacob, age: 52, gender: male}`
Relation Context	Workplace
Relation	Target is the Observer's mentee.
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 wellbeing and long-term goals while making the 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 5: Case study: mentor–mentee dialogue on project resource allocation.

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

Table 6: Latent personality level, self-report and observer-report scores for all Big Five dimensions corresponding to the example in Table 5.