Can LLM "Self-report"?: Evaluating the Validity of Self-report Scales in Measuring Personality Design in LLM-based Chatbots

Huiqi Zou[1], Pengda Wang[2], Zihan Yan[3], Tianjun Sun[2], and Ziang Xiao[1]
[1]Department of Computer Science, Johns Hopkins University
[2]Department of Psychological Sciences, Rice University
[3]Cognitive Science and Language Processing Informatics, University of Illinois Urbana-Champaign
{hzou11,ziang.xiao}@jhu.edu; {pw32,ts110}@rice.edu; zihan25@illinois.edu

Abstract

Personality design plays an important role in chatbot development. From rule-based chatbots to LLM-based chatbots, evaluating the effectiveness of personality design has become more challenging due to the increasingly open-ended interactions. A recent popular approach uses self-report questionnaires to assess LLM-based chatbots' personality traits. However, such an approach has raised serious validity concerns: chatbot's "self-report" personality may not align with human perception based on their interaction. Can LLM-based chatbots "self-report" their personality? We created 500 chatbots with distinct personality designs and evaluated the validity of self-reported personality scales in LLM-based chatbot's personality evaluation. Our findings indicate that the chatbot's answers on human personality scales exhibit weak correlations with both user perception and interaction quality, which raises both criterion and predictive validity concerns of such a method. Further analysis revealed the role of task context and interaction in the chatbot's personality design assessment. We discuss the design implications for building contextualized and interactive evaluation of the chatbot's personality design.

1 Introduction

With advancements in artificial intelligence (AI), especially in natural language processing (NLP), AI-based chatbots are now being used in various fields. Their applications, including creative writing (Chung et al., 2022), mental counseling (Lai et al., 2023), data collection (Wei et al., 2024), and teaching assistant (Hicke et al., 2023), have made interactions with conversational agents increasingly common. To further enhance dialogue engagement and adapt these agents to specific tasks, developers often manipulate chatbots' behavior patterns or assign them specific identities. One common method is through personality design (Shao et al., 2023; Li et al., 2023; Park et al., 2023).

The development of chatbot's personality design has progressed significantly, from early rule-based systems such as ELIZA (Weizenbaum, 1966), which used predefined scripts to simulate human conversation, to modern methods involving large language models (LLMs). Initially, personality design was limited by rigid sets of rules and responses. However, as AI systems evolved, more flexible approaches emerged, such as end-to-end systems (Zhang et al., 2018), which train the dialogue system to respond in a manner that aligns with a given persona. More recent developments center around using LLMs, which serve as the backbone of AI chatbots with customizable personalities. This includes characterizing LLM agents through demographic and persona-based dialogue, as well as using more dimensional approaches, such as the 70 bipolar adjectives proposed by Goldberg (1992); Serapio-García et al. (2023), or converting personality scale items into open-ended questions (Ran et al., 2024), offering greater flexibility in adapting to various chatbot personality designs. In this study, we will focus on LLM-based chatbots and their personality design.

As the chatbot becomes more open-ended, evaluating the effectiveness of the personality design becomes increasingly challenging. Unlike earlier systems, there are no hand-written rules for chatbot designers to inspect. However, the open-endedness of those chatbots also enables new evaluation methods. Researchers have started to use self-report personality scales designed for humans to evaluate the effectiveness of LLM-based chatbots' personality design (Huang et al., 2023a; Wang et al., 2024c; Tu et al., 2023). In these studies, LLMs are typically asked to rate items on a Likert scale (Likert, 1932), with each item corresponding to a specific personality trait dimension. Subsequently, the overall personality scores in each domain are calculated based on these item-level ratings.

While the "self-report" method can produce fast

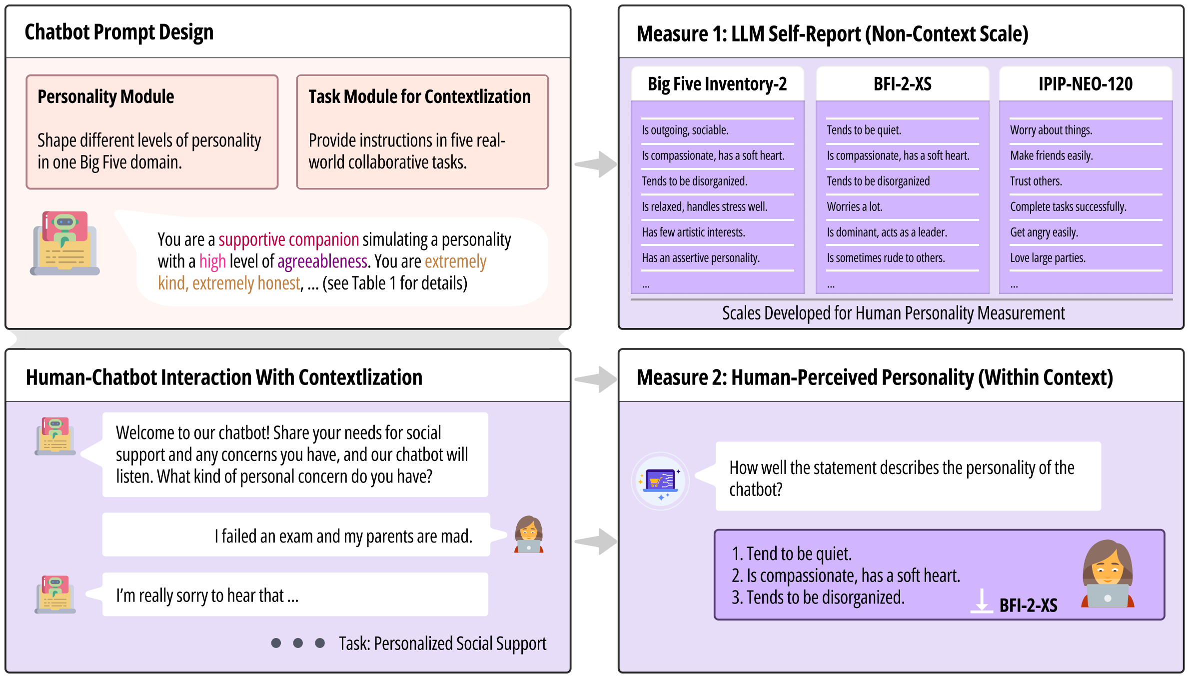

Figure 1: Overview of the evaluation pipeline. Two measures are used to evaluate the chatbot's personality: (1) LLM self-reporting through established personality scales (e.g., Big Five Inventory-2 and IPIP-NEO-120); (2) Human-perceived personality, where users interact with the chatbot and subsequently rate its personality based on their interaction.

and low-cost evaluation results, there are strong, *yet untested*, assumptions regarding the validity of such methods. First, it assumes the internal consistency of LLMs, where the chatbot's response on those scales is consistent with its behaviors, e.g., their interaction with people, which is the ultimate goal of evaluation. Second, these methods presume that the human experiences captured by psychometric scales also apply to LLMs. For instance, a chatbot designed for project management tasks might be assessed on extraversion using an item like "Love surprise parties" (Johnson, 2014a), which is irrelevant to the chatbot's actual role in the designed task. Although these psychometric scales are well-validated for human samples, the examined validity does not directly transfer to the chatbot evaluation, raising concerns about the applicability of such methods in non-human settings.

We aim to fill the gap and evaluate the validity of an LLM-based chatbot's self-reported personality. Guided by measurement theory (Xiao et al., 2023), we evaluated two sets of validity: Convergent and Discriminant validity, where we examined the intrinsic structure, and Criterion and Predictive validity, where we examined the extrinsic relationship with external variables. In particular, we looked at two external variables that are key to the ultimate goal of personality design: human perception and interaction quality.

In this study, we created 500 chatbots with distinct personality designs and collected their "self-report" personality. Then, we recruited 500 participants to interact with each of the chatbots to complete the designed task, give their assessment of the chatbot's personality, and rate the interaction quality. Figure 1 illustrates our evaluation pipeline. Our findings indicate that although the results of the "self-report" personality scales achieve moderate Convergent and Discriminant validity, they fail to align with human perception and exhibit weak correlations with interaction quality, which suggests limited Criterion and Predictive validity.

Our study raises significant validity concerns regarding the use of self-report personality scales for evaluating chatbot personality design. We argue that to design effective personality design evaluation methods, we need to consider how a chatbot's personality manifests in its interaction with humans and ground such evaluation in downstream tasks. Future work should consider the rich signals encapsulated in human interaction.

Our work provides three contributions,

- Through an empirical study with 500 participants, we unveil the validity concerns of using self-report personality scales for evaluating LLM-based chatbot's personality design.

- Our result offers design implications for creating effective personality design evaluation methods that are grounded in real-world task interactions.

- We present a dataset containing a rich log of human interactions with 500 chatbots, each with distinct personality designs, along with human perceptions of their personalities, facilitating the development of novel interaction-based personality evaluation methods.

2 Methodology

2.1 Chatbot Design

In this study, we first created a set of LLM-based chatbots with various personality designs. Those chatbots were created for tasks in which personality design is critical in user experience. Our chatbot framework consists of two main components: a personality module (§2.1.1) and a task module (§2.1.2). The personality module shapes the chatbot's behavior with predefined personality traits, while the task module provides role-specific instructions and task outlines to complete the task. The complete prompts are in Appendix A.

2.1.1 Personality Module

The personality design of our chatbot is based on the Big Five model (John et al., 1999), a foundational framework in personality research that describes personality across five key domains: extraversion (EXT), agreeableness (AGR), conscientiousness (CON), neuroticism (NEU), and openness to experience (OPE). We adopted the *shape* approach from Serapio-García et al. (2023) to shape a chatbot's personality in each domain. This method extends Goldberg (1992) list of 70 bipolar adjectives to 104 personality trait descriptors, which are mapped to different personality facets within each domain. For example, adjectives such as 'unenergetic' and 'energetic' represent the lower and higher levels of extraversion, respectively. These adjectives were then paired with linguistic qualifiers from Likert-type scales (Likert, 1932) to create personality profiles at varying levels. For our study, we selected the highest-level linguistic qualifier (i.e., 'extremely') and randomly sampled five high-level or low-level adjectives from the same personality domain to generate prompts.

2.1.2 Task Module

We selected five common chatbot task categories where personality design plays a critical role in user experience and task effectiveness. This approach allows us to evaluate the validity of the chatbot's personality design via emulating real-world scenarios where chatbot designers rely on personality evaluation results to build an effective chatbot. The selected tasks are as follows:

1. **Job Interview**, where the chatbot works as an HR representative to assess the interviewee's Organizational Citizenship Behaviors (OCB) such as initiative, helping, and compliance, was chosen due to the strong association of the interviewer's conscientiousness in structured interview settings (Heimann et al., 2021).

2. **Public Service**, focusing on handling workday challenges like resource allocation and crisis management, highlights the chatbot's ability to help humans navigate high-stress situations where agreeableness and emotional stability are crucial (Mishra et al., 2023).

3. **Personalized Social Support**, where chatbot features are rooted in the counseling psychology principles of empathy and reflective listening as outlined by Rogers (1957), offers empathetic responses to individuals in distress referring Acceptance and Commitment Therapy (ACT) techniques. Research shows that empathy is closely linked to agreeableness (Walton et al., 2023).

4. **Customized Travel Planning**, which relies on the chatbot's decision-making and personalization skills, underscores the importance of anthropomorphic traits in recommendation systems that foster trust and enhance user decision-making (Qiu and Benbasat, 2009). Additionally, previous research suggests that question types related to extraversion levels significantly impact user satisfaction (Miyama and Okada, 2022).

5. **Guided Learning**, where the chatbot aids humans in understanding complex topics, is based on the constructivist learning approach (Higgs et al., 2004) and highlights the critical

role of extraversion in instructors for enhancing student acceptance.

In total, we randomly sampled five personality adjective markers at high and low levels of each domain ten times. Combining these 100 personality descriptions with five task instructions generated 500 chatbots with distinct personality designs. We used GPT-4o (OpenAI, 2024), one of the most performant models available, as the backbone model with the temperature set to zero.

2.2 Personality Design Evaluation

2.2.1 Self-reported Personality

To quantify the chatbot's self-reported personality, we utilized three well-established psychometric tests that assess the Big Five domains (John et al., 1999): Big Five Inventory–2 Extra-Short Form (BFI-2-XS) (Soto and John, 2017b), BFI-2 (Soto and John, 2017a) and IPIP-NEO-120 (Johnson, 2014b). The chatbot's responses to test items were aggregated to calculate the personality scores for the five domains respectively, which were then used for subsequent analysis. Following the work of Serapio-García et al. (2023), we instructed the chatbot to rate test items using a standardized response scale from one to five based on the personality descriptions. The complete prompt format is shown in Appendix B.

2.2.2 Human-perceived Personality

We conducted a human study to gather participants' perceptions of the chatbot's personality. Each participant was randomly assigned to interact with one chatbot and engaged in multi-turn, text-based conversations to complete the designed task. A user interface example is shown in Appendix C.

After completing the task, participants were asked to rate their perceived personality of the chatbot using the BFI-2-XS (Soto and John, 2017b). The BFI-2-XS consists of 15 items, each representing a distinct facet of one of the Big Five personality domains, thus preserving the scale's comprehensive descriptive and predictive capabilities.

2.3 Study Procedure

To gather the human-perceived personality of each chatbot, we recruited English-speaking participants from Prolific[1], with each participant assessing one chatbot. At the beginning of the survey, participants were given the task's objective, along with clear instructions on how to start the conversation, including an example to clarify how the participant might approach the task. A privacy reminder is included to ensure no private information is shared. After interacting with the chatbot, each participant filled out the chatbot personality evaluation questionnaire and completed a separate User Experience Questionnaire (UEQ) (Laugwitz et al., 2008). The entire process was designed to take 10 to 15 minutes. Appendix F presents detailed participant statistics and their demographic information.

[1] https://www.prolific.com

2.4 Construct Validity Evaluation

Construct validity refers to the extent to which a test accurately measures what it is intended to measure (Cronbach and Meehl, 1955). In our study, we assessed validity through four aspects:

- **Convergent validity** assesses whether different methods measuring the same trait yield consistent results (John and Benet-Martínez, 2000). We tested this by comparing the chatbot's self-reported personality scores across three established scales. Strong correlations among these measures indicate consistent measurement across different methods.

- **Discriminant validity** evaluates the distinctiveness of different traits (Campbell and Fiske, 1959). We verified the independence of the chatbot's personality dimensions by calculating correlations between different trait scores. Low correlations between unrelated traits confirm good discriminant validity.

- **Criterion validity** examines the correlation between a measure and an external criterion (Cronbach and Meehl, 1955). We assessed this by comparing the chatbot's self-reported personality traits with human perceptions, measured through participant ratings after interacting with the chatbot. The value of criterion validity indicates the level of alignment between self-reports and user perceptions.

- **Predictive validity** refers to a measure's ability to predict future outcomes or behaviors (Funder, 2006). We evaluated this by exploring how the chatbot's self-report personality influences the quality of interaction, as captured through the UEQ. A strong relationship between personality traits and positive user experiences suggests that these traits are valid predictors of chatbot effectiveness.

To assess convergent and discriminant validity, we applied the Multitrait-Multimethod (MTMM) matrix approach (Campbell and Fiske, 1959). This framework evaluates multiple traits across different methods to identify patterns of consistency and distinctiveness. We implemented this analysis using Spearman correlation in R. Due to the non-hierarchical structure of the BFI-2-XS, which does not decompose traits into facets, we focused on domain-level traits to evaluate construct validity across different assessment settings.

3 Results

3.1 Does Personality Setting Work?

We compared the mean and standard deviation of task-specific human-perceived personality evaluation scores across different levels to assess how effective our personality design method is. As shown in Table 1, with the exception of conscientiousness scores in the social support task, all other domains consistently score higher in the high-setting conditions across tasks, with moderate variances. Additionally, the analysis of variance (Appendix G, Table 10) indicates that the variance between high-low groups is more pronounced than the variance within each group. These findings collectively support the effectiveness of our trait manipulation in the chatbot personality design.

3.2 Convergent and Discriminant Validity

Table 2 presents the correlation between the chatbot's self-reported scores across different scales. It is evident that, regardless of the dimension, the correlations across scales show a high degree of consistency, with an average correlation coefficient of 0.85. This result indicates that the chatbot demonstrates a high level of stability in its self-reports across different personality questionnaires, suggesting consistency within self-report methods.

Table 3 shows the correlations of human-perceived and self-report personality scores across different personality traits. The results further demonstrate alignment across the self-report methods, both in terms of the magnitude and direction of correlations between each pair of traits, as well as the absolute mean values.

3.3 Criterion Validity (Self-report vs. Human)

A good measure for personality design should align with the end-user perception of the chatbot. Table 4 presents the means, standard deviations, and correlations between human perception and chatbot self-reported scores. The results show that the means of the human perception scores and chatbot self-reported scores are quite similar, primarily reflecting the overall mean distribution of the sample, which is expected to be close. Notably, under the same personality setting, the standard deviation of the human perception scores is smaller than that of the chatbot self-reported score, indicating that human perceived scores show less variation between individuals, while chatbot self-reports exhibit greater variability.

What is more interesting is the correlation between human perception scores and chatbot self-reported scores. Apart from the relatively high correlation (0.58 ± 0.02) in the domain of agreeableness, the correlations in the others are all below 0.5. This suggests a low level of consistency between human perceptions and chatbot self-reports.

In addition, Table 3 shows differences in discriminant validity between human-perceived personality and chatbot's self-reported personality. Compared to human-perceived personality traits, chatbot's "self-reported" traits show a higher correlation with each other, except for the correlation between conscientiousness and openness, suggesting a lower level of discriminant validity.

Across the Tasks In addition to the overall correlation analysis, we further examined each individual task. Given the high correlations among the three "self-report" scales (§3.2), we averaged the results of the three questionnaires and conducted a comparative analysis with human-perceived personality design (see Table 5). For detailed correlation analysis for each questionnaire, please refer to Tables 11, 12 and 13 in Appendix H.

As shown in the table, agreeableness consistently exhibited the highest correlation across all tasks. However, the correlations for other personality dimensions were weaker, with average values generally below 0.39, and considerable fluctuations in the correlations for the same personality traits across different tasks. For instance, conscientiousness even showed a negative correlation in the social support task, further highlighting the limitations of evaluating chatbot personality design solely based on its "self-report" data, especially across different task contexts.

One key factor may explain these results. The variation in correlation across tasks for specific personality traits may reflect how the chatbot's behav-

Domain	Job Interview				Public Service				Social Support				Travel Planning				Guided Learning			
	High		Low		High		Low		High		Low		High		Low		High		Low	
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Ext	3.70	0.74	3.40	0.73	4.37	0.58	2.93	0.81	3.93	0.84	3.50	0.92	4.10	0.65	2.93	1.11	3.90	0.82	3.60	0.75
Agr	4.20	0.65	1.57	1.24	4.47	0.59	2.07	1.57	4.50	0.48	1.87	0.79	3.27	1.25	1.50	0.86	4.10	0.61	1.43	0.70
Con	4.00	0.89	3.60	1.27	3.97	0.97	2.83	1.21	3.93	1.00	**3.97**	1.00	4.60	0.34	3.20	1.12	4.20	0.93	4.17	0.71
Neu	2.90	0.97	1.80	0.65	2.60	0.78	1.90	0.47	2.27	1.36	1.70	0.66	3.10	1.31	1.50	0.48	2.93	1.00	2.03	0.96
Ope	3.43	0.57	3.23	0.83	3.80	0.47	3.13	0.89	3.83	0.50	3.23	0.94	4.00	1.04	3.27	0.70	3.70	1.07	3.57	0.61

Table 1: Domain-level mean (M) and standard deviation (SD) of human-perceived personality scores under high and low personality settings across five tasks. Light gray cells indicate instances where the low standard deviation for high or low personality settings and the difference between these settings exceeds 0.3. Dark gray cells highlight instances where the same difference exceeds 0.5.
Note: n = 500 in total. Ratings were made on a five-point scale ranging from 1 (disagree strongly) to 5 (agree strongly). Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

Questionnaire	**Ext**	**Agr**	**Con**	**Neu**	**Ope**	**Mean**
BFI-2-XS vs BFI-2	0.90	0.91	0.83	0.87	0.90	0.88
BFI-2-XS vs IPIP	0.82	0.84	0.91	0.86	0.81	0.85
IPIP vs BFI-2	0.88	0.83	0.86	0.86	0.83	0.85

Table 2: Correlation between self-reported personality scores using BFI-2-XS, BFI-2, and IPIP-NEO-120.
Note: n = 500 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

	Human-perceived	BFI-2-XS	BFI-2	IPIP
Ext vs Agr	0.18	0.34	0.47	0.54
Ext vs Con	0.37	0.56	0.44	0.47
Ext vs Neu	-0.21	-0.57	-0.61	-0.59
Ext vs Ope	0.27	0.65	0.73	0.69
Agr vs Con	0.56	0.56	0.57	0.64
Agr vs Neu	-0.39	-0.49	-0.60	-0.50
Agr vs Ope	0.57	0.67	0.62	0.51
Con vs Neu	-0.56	-0.72	-0.66	-0.80
Con vs Ope	0.50	0.49	0.35	0.16
Neu vs Ope	-0.29	-0.36	-0.34	-0.24
Absolute Mean	0.39	0.54	0.54	0.51

Table 3: Correlation analysis for human-perceived and self-report personality scores across personality traits.
Note: n = 500 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

ior is influenced by the task type and the interaction with users. For example, agreeableness may align more with general expectations of friendliness and cooperation during interactions, whereas conscientiousness might have the opposite effect in some tasks (such as social support tasks) due to an excessive focus on detail or rules, which could negatively impact the human experience. The performance of different personality traits across tasks is not uniform but shaped by task context and interaction.

In summary, the criterion validity of LLM-based chatbot is low, raising concerns about the effectiveness of using "self-report" methods to measure a chatbot's personality design. Such inconsistency indicates the gap between an LLM's responses to personality questionnaires and the actual behaviors it exhibits during real-world user interactions.

3.4 Predictive Validity (Interaction Quality)

Tables 6 and 7 detail correlation analyses between aggregated UEQ scores and personality scores, both human-perceived and self-reported, across five tasks. This analysis evaluates the predictive validity of these personality assessments in relation to interaction quality with chatbots, which is the ultimate goal of building such chatbots. Table 6 demonstrates significant correlations between perceived agreeableness and conscientiousness and user experience, notably in the travel planning task, with correlations of 0.71 and 0.76, respectively. These findings suggest that chatbots perceived as agreeable and conscientious markedly enhance user experience. In contrast, neurotic traits show consistently negative correlations across all tasks, most pronounced with a correlation of -0.55 in travel planning, indicating that these traits detract from usability. The data illustrates that positive personality traits are predictive of enhanced usability, while negative traits hinder it. This consists with research indicating the crucial role of personality design in shaping user experience among the selected tasks.

Conversely, Table 7 reveals discrepancies between self-reported personality scores and user experience, characterized by low and inconsistent correlations. For instance, conscientiousness exhibits a modest positive correlation of 0.17 in the public service task, the highest observed among the traits. However, the overall pattern of weak and occasionally near zero correlations, such as -0.01 for openness in guided learning, indicates that self-reported traits are unreliable predictors of user experience. This disparity likely arises because self-report measures fail to capture the dynamic impacts of interaction and task setting, crucial in shaping user perceptions during real-time chatbot engagements. These findings emphasize the limita-

		Human-perceived		BFI-2-XS			BFI-2			IPIP-NEO-120			Self-reported Mean		
		M	SD	M	SD	ρ	M	SD	ρ	M	SD	ρ	M	SD	ρ
Human-perceived	Ext	3.72	0.78	2.89	1.17	0.23	3.01	1.17	0.20	2.91	1.01	0.17	2.94	1.12	0.20
	Agr	3.80	1.14	3.61	1.37	0.59	3.53	1.23	0.59	4.01	0.82	0.56	3.72	1.14	0.58
	Con	3.97	0.96	3.70	1.42	0.23	3.44	1.10	0.24	3.62	1.12	0.25	3.59	1.21	0.24
	Neu	2.03	0.88	2.30	1.30	0.29	2.71	1.11	0.33	2.37	1.12	0.36	2.46	1.18	0.33
	Ope	3.27	0.84	3.31	1.32	0.23	3.09	1.11	0.21	3.36	0.83	0.24	3.25	1.09	0.23

Table 4: Domain-level mean (M), standard deviation (SD), and correlation (ρ) of self-reported and human-perceived personality scores using the BFI-2-XS, BFI-2, and IPIP-NEO-120.
Note: n = 500 in total. Ratings were made on a five-point scale ranging from 1 (disagree strongly) to 5 (agree strongly). Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.19	0.54	0.28	0.31	0.19
Public Service	0.40	0.58	0.44	0.30	0.23
Social Support	0.03	0.58	**-0.03**	0.13	0.15
Travel Planning	0.28	0.60	0.39	0.49	0.33
Guided Learning	0.05	0.56	0.03	0.34	0.14
Absolute Mean	0.19	0.57	0.23	0.31	0.21

Table 5: Average correlation between self-reported and human-perceived personality scores across tasks.
Note: n = 100 for each task. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.26	0.37	0.49	-0.40	0.48
Public Service	0.14	0.32	0.41	-0.22	0.28
Social Support	-0.05	0.17	0.19	-0.15	0.25
Travel Planning	0.28	0.71	0.76	-0.55	0.52
Guided Learning	0.15	0.14	0.24	-0.09	0.16

Table 6: Correlation between aggregated UEQ scores and human-perceived personality scores across tasks.
Note: n = 433 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

tions of exclusively relying on self-reported personality to evaluate user experience and the importance of considering perceived personality in interactive settings. Future research should explore the factors affecting the relationship between personality perception and user experience, and develop methods that more accurately capture the dynamic nature of personality in interactions.

In summary, self-report personality scales failed to correlate with interaction quality, which indicates a disconnect between the model's response to personality items and how their behaviors manifest during real interaction. It highlights the complexity of user interaction and the challenges in evaluating personality design in LLM-based chatbots. As the ultimate goal of personality design is to improve interaction, such a validity link is crucially important. The validity issues of the "self-report" evaluation method may misguide chatbot development.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.12	0.15	0.13	-0.09	0.03
Public Service	0.03	0.04	0.17	-0.16	-0.03
Social Support	0.10	0.10	0.08	-0.10	0.02
Travel Planning	0.03	0.08	0.04	-0.17	-0.09
Guided Learning	-0.06	-0.04	-0.07	0.02	-0.01

Table 7: Correlation between aggregated UEQ scores and self-reported personality scores across tasks.
Note: n = 433. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

4 Towards Interaction and Task Grounded Personality Evaluation

Current self-report scales assume that personality traits are expressed consistently across different scenarios. However, the limited predictive and criterion validity of self-report personality scales, as shown in our findings, suggests a disconnect between the scores and the user experience. Moving forward, we advocate for transitioning from static, questionnaire-based evaluations to task-driven assessments that better reflect the scenarios where chatbots operate, aligning with calls from prior research (Lee et al., 2022; Liao and Xiao, 2023).

First, personality evaluations should be based on specific tasks or scenarios, as chatbot personality traits manifest differently depending on the situation, similar to how humans adjust their behavior based on context (Sauerberger and Funder, 2017). Furthermore, contextual shifts may lead users to form different assumptions about a chatbot's personality, as social roles influence how individuals attribute traits to others (Roberts, 2007). Thus, evaluation methods should account for biases from user expectations and the chatbot's context-driven behaviors to align with design objectives.

Second, personality evaluations without considering the expression of traits in real-world interactions fail to capture the chatbot's impact on user experience. Since personality is conveyed through behaviors and mutual perceptions in interaction

(Geukes et al., 2019), an ideal evaluation framework should account for factors in continuous interactions, such as response patterns, and adaptability to user input. To understand the patterns essential for user experience, we may observe how user perceptions of personality evolve with interactions and how these changes impact user satisfaction. By embracing task-based interactive evaluation, we can improve personality evaluations, ultimately fostering chatbots that enhance user satisfaction.

5 Related Work

Human-Like Responses in Large Language Models Recent studies have shown that LLMs demonstrate remarkable abilities resembling human-like responses. For example, Lampinen et al. (2024) found that LLMs display response patterns similar to humans in logical reasoning tasks, with similar findings being reported by Kosinski (2023) and Wang et al. (2024a). Additionally, Pellert et al. (2023) found that LLMs perform comparably to humans across various psychological scales, such as value orientations and moral norms. All these abilities have garnered significant interest from both computer scientists and social scientists, as these findings suggest that we may be able to draw upon existing social science research on humans as a reference for LLM studies. For instance, personality, which has been extensively studied in psychology and consistently proven to be a reliable predictor of various behavioral and psychological outcomes, is one such area of interest. Personality encodes rich and complex information in textual data (Goldberg, 1990; Saucier and Goldberg, 2001), so LLMs may capture and model such encoding by learning from vast training data. Many LLM studies explore the personality traits these models exhibit, with the hope that we can similarly use the models' simulation of personality to predict or shape their behavior (Lee et al., 2024; Huang et al., 2023b; Wang et al., 2024b; Ran et al., 2024).

Current Approaches for Language Models Personality Assessment Currently, most measurements of LLMs' personality are based on psychological research related to human personality assessment, typically using personality scales. A common approach is to have the model respond to scales that have already been developed and validated on human samples, such as the BFI-2 (Soto and John, 2017a) and HEXACO (Lee and Ashton, 2004, 2006). The model's scores are then calculated based on its answers to the multiple-choice questions (MCQs) on these scales. The rationale behind this is straightforward: since we believe that LMs simulate human abilities, we can directly borrow well-established and validated methods used for humans to assess the personality traits of these models. For example, studies like Huang et al. (2023b) used this approach to evaluate whether LLMs reflect the personality traits outlined in the BFI (John et al., 1999). Serapio-García et al. (2023) presented the reliability and validity of LLMs in resembling personality profiles along desired dimensions with IPIP-NEO (Goldberg et al., 1999) and BFI (John et al., 1999). However, some researchers argue that traditional scales may not sufficiently capture the personality of LMs and have adopted alternative approaches. For instance, some researchers modify scale questions into prompts that elicit text-based responses (Wang et al., 2024c), or others have experts rate stories generated by LMs according to specified criteria (Jiang et al., 2023). These text-based evaluation methods may better reflect real-world application scenarios.

Existing methods have certain shortcomings, as they often overlook the importance of analyzing model behavior from the perspective of the task-specific performance. In other words, current approaches may overemphasize on directly measuring the model's predefined personality traits, without adequately considering the model's specific performance in actual tasks. If the ultimate goal of defining a model's personality is to enhance its effectiveness in specific tasks, then the behavior exhibited by the model in these tasks—and the corresponding personality traits—should be the primary focus of research and evaluation.

6 Conclusion

This paper demonstrates the limitations of self-reported personality assessments in evaluating the personality design of LLM-based chatbots. Our findings reveal the discrepancy between chatbots' self-reported personality scores and human task-based perceptions, suggesting that self-assessments may not accurately capture how chatbots are perceived in real-world interactions. Additionally, our analysis of predictive validity indicates that self-reported personality scales do not align with interaction quality. These results highlight the need for evaluation methods that capture chatbot personality in task-driven, interactive scenarios.

Limitations

Several limitations exist in the current study. First, there may be bias in the choice of psychometric tests. Although we sought to minimize this by using personality assessments of different lengths for self-reported data, the human-perceived personality was measured with only one questionnaire due to time constraints. Future research could address this by incorporating a broader range of personality tests and developing assessments specifically tailored for LLMs to achieve more accurate measurements.

Second, our findings are based on a single chatbot personality setting method and may not generalize well to others, highlighting the need for further investigation with different approaches. Additionally, the chatbot was designed with a focus on task completion, mirroring real-world applications to ensure ecological validity. As a result, the task settings may have constrained the variability in perceived personality. Moreover, the evaluation was conducted on five common chatbot tasks, which may not capture the full spectrum of user interactions. Expanding the task set in future studies could provide a more comprehensive view of task-based chatbot personality assessment.

Third, while our study only focuses on GPT-4o (OpenAI, 2024), we expect the same experimental setup could be extended to other LLMs. The extent to which our findings generalize to a broader range of models remains an open question for future research. Furthermore, while LLMs perform well on benchmarks across multiple languages, this study evaluated the model using English psychometric tests and English-speaking participants. Future research could explore personality dimensions within culture-specific contexts to provide more diverse insights on chatbot design.

Ethics Statement

Our human study was reviewed and granted Exempt status by the Institutional Review Board (IRB). Participants are paid at the rate of $12 per hour through the Prolific platform. We ensure transparency by explicitly stating the study's purpose, requirements, and payments in the task instructions and obtaining informed consent from participants before the study begins. Additionally, we manually verify that all collected data are free of sensitive or personally identifiable information. All prompts used in this study are included in the Appendix. The codes for collecting self-report personality scores and the transcripts of human-chatbot interaction are publically accessible [2]. The cost of running the self-report experiments with the batch API of GPT-4o is $37.41.

References

Donald T Campbell and Donald W Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological bulletin*, 56(2):81.

John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. Talebrush: Sketching stories with generative pretrained language models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin*, 52(4):281.

David C Funder. 2006. Towards a resolution of the personality triad: Persons, situations, and behaviors. *Journal of Research in Personality*, 40(1):21–34.

Katharina Geukes, Simon M Breil, Roos Hutteman, Steffen Nestler, Albrecht CP Küfner, and Mitja D Back. 2019. Explaining the longitudinal interplay of personality and social relationships in the laboratory and in the field: The pils and the connect study. *PloS one*, 14(1):e0210424.

Lewis R Goldberg. 1990. An alternative" description of personality": The big-five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological Assessment*, 4(1):26–42.

Lewis R Goldberg et al. 1999. A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. *Personality psychology in Europe*, 7(1):7–28.

Anna Luca Heimann, Pia V Ingold, Maike E Debus, and Martin Kleinmann. 2021. Who will go the extra mile? selecting organizational citizens with a personality-based structured job interview. *Journal of Business and Psychology*, 36(6):985–1007.

Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. Chata: Towards an intelligent question-answer teaching assistant using open-source llms. *arXiv preprint arXiv:2311.02775*.

[2]https://github.com/isle-dev/self-report

Joy Higgs, Barbara Richardson, and Madeleine Abrandt Dahlgren. 2004. *Developing practice knowledge for health professionals*. Butterworth & Heinemann.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023a. Chatgpt an enfj, bard an istj: Empirical study on personalities of large language models. *arXiv preprint arXiv:2305.19926*.

Jen-tse Huang, Wenxuan Wang, Man Ho Lam, Eric John Li, Wenxiang Jiao, and Michael R Lyu. 2023b. Revisiting the reliability of psychological scales on large language models. *arXiv e-prints*, pages arXiv–2305.

Hang Jiang, Xiajie Zhang, Xubo Cao, and Jad Kabbara. 2023. Personallm: Investigating the ability of large language models to express big five personality traits. *arXiv preprint arXiv:2305.02547*.

Oliver P John and Veronica Benet-Martínez. 2000. Measurement: Reliability, construct validation, and scale construction. *Handbook of research methods in social and personality psychology*.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

John A Johnson. 2014a. 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.

John A Johnson. 2014b. 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.

Michal Kosinski. 2023. Theory of mind may have spontaneously emerged in large language models. *CoRR*, abs/2302.02083.

Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-llm: Scaling up global mental health psychological services with ai-based large language models. *arXiv preprint arXiv:2307.11991*.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Hannah R Sheahan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2024. Language models, like humans, show content effects on reasoning tasks. *PNAS nexus*, 3(7).

Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4*, pages 63–76. Springer.

Kibeom Lee and Michael C Ashton. 2004. Psychometric properties of the hexaco personality inventory. *Multivariate behavioral research*, 39(2):329–358.

Kibeom Lee and Michael C Ashton. 2006. Further assessment of the hexaco personality inventory: two new facet scales and an observer report form. *Psychological assessment*, 18(2):182.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, et al. 2022. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, et al. 2024. Do llms have distinct and consistent personality? trait: Personality testset designed for llms with psychometrics. *arXiv preprint arXiv:2406.14703*.

Tianlong Li, Xiaoqing Zheng, and Xuanjing Huang. 2023. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons. *arXiv preprint arXiv:2310.16582*.

Q Vera Liao and Ziang Xiao. 2023. Rethinking model evaluation as narrowing the socio-technical gap. *arXiv preprint arXiv:2306.03100*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Himani Mishra, Tuheena Mukherjee, and M Venkatesan. 2023. Role of personality on psychological well-being in public services: Theorizing the self-determination theory perspective. *Global Business Review*, page 09721509231175224.

Tamotsu Miyama and Shogo Okada. 2022. Personality-adapted multimodal dialogue system. *arXiv preprint arXiv:2210.09761*.

OpenAI. 2024. Hello GPT-4o.

Jeongeon Park, Bryan Min, Xiaojuan Ma, and Juho Kim. 2023. Choicemates: Supporting unfamiliar online decision-making with multi-agent conversational interactions. *arXiv preprint arXiv:2310.01331*.

Max Pellert, Clemens M Lechner, Claudia Wagner, Beatrice Rammstedt, and Markus Strohmaier. 2023. Ai psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. *Perspectives on Psychological Science*, page 17456916231214460.

Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of management information systems*, 25(4):145–182.

Yiting Ran, Xintao Wang, Rui Xu, Xinfeng Yuan, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2024. Capturing minds, not just words: Enhancing role-playing language models with personality-indicative data. *arXiv preprint arXiv:2406.18921*.

Brent W Roberts. 2007. Contextualizing personality psychology. *Journal of personality*, 75(6):1071–1082.

Carl R Rogers. 1957. The necessary and sufficient conditions of therapeutic personality change. *Journal of consulting psychology*, 21(2):95.

Gerard Saucier and Lewis R Goldberg. 2001. Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of personality*, 69(6):847–879.

Kyle S Sauerberger and David C Funder. 2017. Behavioral change and consistency across contexts. *Journal of Research in Personality*, 69:264–272.

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

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187.

Christopher J Soto and Oliver 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.

Christopher J Soto and Oliver 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.

Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. 2023. Characterchat: Learning towards conversational ai with personalized social support. *arXiv preprint arXiv:2308.10278*.

Kate E Walton, Dana Murano, Jeremy Burrus, and Alex Casillas. 2023. Multimethod support for using the big five framework to organize social and emotional skills. *Assessment*, 30(1):144–159.

Pengda Wang, Zilin Xiao, Hanjie Chen, and Frederick L Oswald. 2024a. Will the real linda please stand up... to large language models? examining the representativeness heuristic in llms. *arXiv preprint arXiv:2404.01461*.

Pengda Wang, Huiqi Zou, Zihan Yan, Feng Guo, Tianjun Sun, Ziang Xiao, and Bo Zhang. 2024b. Not yet: Large language models cannot replace human respondents for psychometric research.

Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. 2024c. 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.

Jing Wei, Sungdong Kim, Hyunhoon Jung, and Young-Ho Kim. 2024. Leveraging large language models to power chatbots for collecting user self-reported data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–35.

Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Ziang Xiao, Susu Zhang, Vivian Lai, and Q. Vera Liao. 2023. Evaluating evaluation metrics: A framework for analyzing NLG evaluation metrics using measurement theory. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10967–10982, Singapore. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

A Chatbot Prompt Format

As shown in Table 8, each chatbot prompt consists of a main section and six interchangeable components. Although human personality consists of traits from various domains, for simplicity, we configure only one domain at a time. The prompt begins by highlighting the chatbot's key personality traits within the specified domain, offering a detailed description. This is followed by task-specific instructions to establish context. Finally, the key personality traits are reiterated to ensure consistent behavior throughout the interaction.

The *role* defines the specific role the chatbot simulates across five tasks, such as a supportive companion or an educational guide. The *level* describes the extent to which the chosen domain is represented. The *domain* specifies the personality dimension selected from the Big Five model (John et al., 1999). The *profile* describes the chatbot's personality using five key adjectives, each paired with a qualifier, as outlined in section 2.1.1. The *objectives* summarize the chatbot's goals within the given task, such as offering personalized social support or explaining complex concepts. Finally, the *additional_info* section includes any extra specific focus areas for the task.

Table 8 also presents an example of a high-level personality prompt in the agreeableness domain, specifically designed for the personalized social support task.

You are a(n) {role} simulating a personality with a {level} level of {domain}. Shape your responses using these key adjectives: you are {profile}. Your main objective is to {objectives}. {additional_info}. The personality with a {level} level of {domain} and the key adjectives should guide your questions and responses.
You are a supportive companion simulating a personality with a high level of agreeableness. Shape your responses using these key adjectives: you are extremely kind, extremely honest, extremely trustful, extremely unselfish and extremely moral. Your main objective is to provide personalized social support to users, listening to their concerns and offering responses. Draw on principles from counseling psychology, particularly the use of reflective listening and validation techniques. Your responses should demonstrate an understanding of the user's emotional state and provide advice depending on the situation. Aim to build rapport and trust, helping the user feel understood and supported during their moment of need. The personality with a high level of agreeableness and the key adjectives should guide your questions and responses.

Table 8: Chatbot prompt format with an example for the personalized social support task. One prompt consists of a main part and six replaceable components: role, level, domain, profile, objectives, and additional information.

B Self-reported Personality Prompt Format

A complete prompt format for self-reported personality BFI-2-XS (Soto and John, 2017b) and BFI-2 (Soto and John, 2017a) is as follows, where the *personality_description* is the same as the task-based personality description introduced in Section 2.1.1 and *test_item* represents each item in the psychometric tests.

> For the following task, respond in a way that matches this description: "{personality_description}." Considering the statement, please indicate the extent to which you agree or disagree on a scale from 1 to 5 (where 1 = "disagree strongly", 2 = "disagree a little", 3 = "neither agree nor disagree", 4 = "agree a little", and 5 = "agree strongly"): "{test_item}".

For the IPIP-NEO-120 (Johnson, 2014b) questionnaire, the response scale is modified as follows: 1 = very inaccurate, 2 = moderately inaccurate, 3 = neither accurate nor inaccurate, 4 = moderately accurate, and 5 = very accurate.

C Interface for Human Study

Figure 2 presents a screenshot of the interface used in the human study, built by modifying an open-source Github repository[3]. The interface features a split-screen design with a chat window on the left and an evaluation window on the right. The chat window displays the conversation history with the chatbot, beginning with a greeting and an initial task instruction. The evaluation window presents a survey with personality assessment statements, each accompanied by response options. During the evaluation, the participant indicated the level of agreement with each statement by selecting the corresponding button. In this study, the user interacts with a chatbot and subsequently completes the evaluation questionnaire, allowing for an assessment of the chatbot's perceived personality traits based on the user's experience.

D Human Study Details

The participant recruitment scripts are as follows.

> Hello! We're conducting a research project to evaluate LLM-based chatbots. In this task, you'll engage in an activity with a chatbot following the provided instructions. Once the task is complete, you'll be asked to fill out a brief questionnaire with 15 questions on the same page to evaluate the personality of the chatbot you interacted with. Please note that this questionnaire is in English and requires a desktop to complete.
> By completing this survey or questionnaire, you are consenting to be in this research study. Your participation is voluntary and you can stop at any time. Your responses will be anonymous and

[3]https://big-agi.com

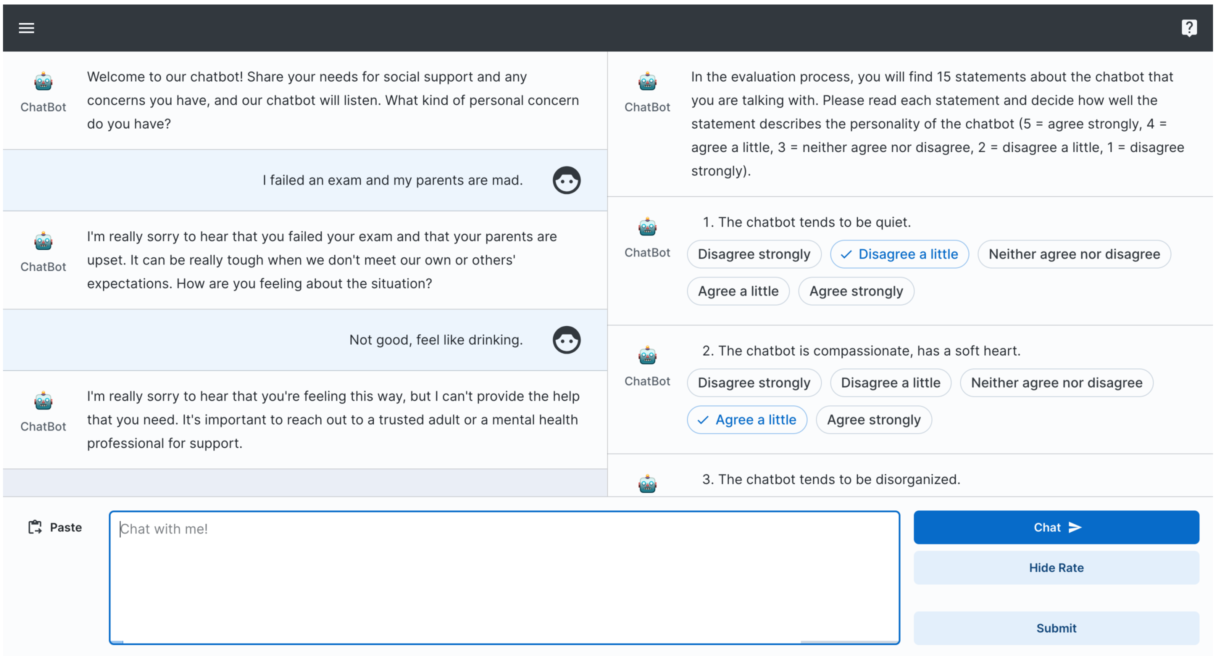

Figure 2: Graphical user interface for human study.

Domain	Personality Marker Level	Task	Total
5	10 High, 10 Low	5	500

Table 9: Statistics of human evaluation dataset.

kept confidential. There are no risks or discomfort associated with participating. This task should take 10-15 minutes to complete. Thank you for your participation!

E Dataset Statistics

In this study, we configured chatbot personalities across each Big Five domain (John et al., 1999) by randomly sampling adjectives at high or low marker levels, generating ten distinct profiles per domain for each level. Each participant engaged with one chatbot in one of five predefined task settings and then completed the BFI-2-XS questionnaire (Soto and John, 2017b). We collected a total of 500 valid conversation transcripts and corresponding assessments, with each interaction averaging 8 turns. Table 9 summarizes the statistics of the human evaluation dataset.

F Participant Statistics

Among all participants who provided demographic information, 172 identified as women, 255 as men, and 6 as non-binary or of a third gender. The median education level was a Bachelor's degree, the median household income ranged between \$25,000 and \$50,000, and the median age was 25–34 years. 331 participants reported using conversational AIs (e.g., ChatGPT, Siri, or Alexa) at least once per week.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	1.030	3.608	2.056	2.212	2.149
Public Service	1.993	7.028	1.541	2.740	3.500
Social Support	1.189	2.724	1.009	4.287	3.517
Travel Planning	2.921	2.076	10.667	7.465	2.222
Guided Learning	1.186	1.332	1.738	1.089	3.086

Table 10: F-value for human-perceived personality scores between high and low personality settings across tasks.
Note: n = 100 for each task. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

G Variance in Human-perceived Personality

Table 10 presents the F-values for human-perceived personality scores under high and low personality settings across five tasks. These values indicate the ratio of variances in personality perceptions between the two settings for each domain. The table shows that variance between the high and low settings is more pronounced than within each group. Notably, this effect is most prominent in certain tasks and personality domains, such as agreeableness and conscientiousness.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.22	0.55	0.27	0.26	0.21
Public Service	0.42	0.58	0.43	0.26	0.24
Social Support	0.08	0.63	**-0.01**	0.09	0.19
Travel Planning	0.27	0.58	0.35	0.44	0.30
Guided Learning	0.11	0.60	0.03	0.35	0.15

Table 11: Correlation analysis between self-reported personality with BFI-2-XS and human-perceived personality scores across tasks.
Note: n = 500 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.20	0.60	0.27	0.28	0.20
Public Service	0.40	0.59	0.47	0.31	0.22
Social Support	0.02	0.64	**-0.01**	0.11	0.14
Travel Planning	0.26	0.59	0.40	0.52	0.31
Guided Learning	0.05	0.53	**-0.03**	0.30	0.08

Table 12: Correlation analysis between self-reported personality with BFI-2 and human-perceived personality scores across tasks.
Note: n = 500 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

Task	Ext	Agr	Con	Neu	Ope
Job Interview	0.15	0.48	0.30	0.39	0.17
Public Service	0.37	0.57	0.42	0.33	0.24
Social Support	**-0.02**	0.48	**-0.07**	0.18	0.12
Travel Planning	0.30	0.64	0.41	0.49	0.36
Guided Learning	**-0.0005**	0.55	0.08	0.38	0.19

Table 13: Correlation analysis between self-reported personality with IPIP-NEO-120 and human-perceived personality scores across tasks.
Note: n = 500 in total. Ext = Extraversion; Agr = Agreeableness; Con = Conscientiousness; Neu = Neuroticism; Ope = Openness.

H Correlation Analysis

Table 11, 12 and 13 present the correlation analysis between self-reported and human-perceived personality scores using the BFI-2-XS, BFI-2, and IPIP-NEO-120 questionnaires, respectively. All three tables show a weak correlation between human-perceived personality and those evaluated via standard tests.