Computer Animation and Virtual Worlds

WILEY

SPECIAL ISSUE PAPER

Talk With Socrates: Relation Between Perceived Agent Personality and User Personality in LLM-Based Natural Language Dialogue Using Virtual Reality

Mehmet Efe Sak | Sinan Sonlu | Uğur Güdükbay

Department of Computer Engineering, Bilkent University, Ankara, Turkey

Correspondence: Uğur Güdükbay (gudukbay@cs.bilkent.edu.tr)

Received: 16 April 2025 | **Accepted:** 11 May 2025

Funding: This work was supported by the Technological Research Council of Turkey (TÜBİTAK) under Grant No. 122E123.

Keywords: conversational agent | five-factor personality | large language model | personality perception | virtual reality

ABSTRACT

Large Language Models (LLMs) offer almost immediate human-like quality responses to user queries. Conversational agent systems support natural language dialogues utilizing LLM backends in combination with Text-to-Speech (TTS) and Automatic Speech Recognition (ASR) technologies, enabling life-like characters in virtual environments. This study investigates the relationship between user personality and perceived agent personality in LLM-based natural language dialogue. We adopt a Virtual Reality (VR) setting where the user can talk with the agent that assumes the role of Socrates, the famous philosopher. To this end, we utilize a three-dimensional (3D) avatar model resembling Socrates and use specific LLM prompts to get stylistic answers from OpenAI's Chat Completions Application Programming Interface (API). Our user study measures the agent's personality and the system's ease of use, quality, realism, and immersion concerning the user's self-reported personality. The results suggest that the user's conscientiousness, extraversion, and emotional stability have a moderate effect on certain personality factors and system qualities. User conscientiousness affects the perceived ease of use, quality, and realism, while user extraversion affects perceived agent conscientiousness, system realism, and immersion. Additionally, the user's emotional stability correlates with perceived extraversion and agreeableness.

1 | Introduction

The rise of LLMs enabled the easy design of character dialogue for conversational agent systems. Historically, Artificial Intelligence (AI) characters required using handcrafted mechanisms with predefined rules to match user dialogue patterns to agent responses. One popular choice was Artificial Intelligence Markup Language (AIML) [1], an Extensible Markup Language (XML) dialect that can set predefined variables to recall user-given information in its responses. AIML was used in the award-winning ALICE [2], supporting over 40,000 knowledge categories, that is, response contents. Although techniques like text-mining helped craft AIML models with ease [3] and AIML-based chatbots are still used in domain-specific cases [4], building an AIML-based system that well generalizes to open-ended conversations requires too much effort. Although AIML supports

Abbreviations: ASR, automatic speech recognition; LLM, large language model; TTS, text-to-speech; VR, virtual reality.

Mehmet Efe Sak and Sinan Sonlu are joint first authors.

randomness, creating response variation through such randomness is labor-intensive, and thus, adding personality to AIML-based agents is challenging [5].

LLMs' data-driven nature removes the need for handcrafting dialogue patterns and exhibits variation inherently. Being trained on very large corpora, LLM responses to user queries easily create the illusion of human understanding [6]. LLM-based chatbots offer highly generalizable dialogues that can be specialized to different fields using specific training data or prompting [7]. LLMs can help with story character construction [8] and authoring simulation-based dynamic plots [9], emphasizing the rich styles of the generated content. While certain problems like hallucinations can cause unreliable responses and thus introduce a risk for the use in certain areas [10], LLMs offer highly human-like and accurate responses well suited to general dialogue. LLMs offer personalized interactions and flexibility in education [11], enable collaboration with nurses in outpatient reception [12], and efficient summarization [13]. This highly versatile aspect of LLMs makes them indispensable to the recent conversational agents [14]. LLM-based agents can assume various roles or personalities [15], which can be used for improving engagement in conversational agent dialogue [16]. The perceived traits of LLMs are important for creating an overall positive conversational experience [17].

People perceive personality in stylistic dialogue [18]. Studies aim to express different personalities in virtual agents through utilizing different nonverbal features [19], animation [20], gesture [21], and dialogue [22] for enhancing variation and immersion. The works focusing on accurate personality expression often utilize heuristic-based cues to alter stimuli to express specific styles. Emotional words and grammar choices can control the user's perceptions of linguistic style and personality; for example, filled pauses and tag questions can be systematically inserted into dialogue to express introversion [23]. LLMs exhibit such styles in a data-driven manner. Recent studies focus on evaluating the personality of LLM responses [24]. LLMs can generate personality-enriched dialogue through prompting [25], and highly detailed prompts can result in accurate personification [26, 27]. On the other hand, long prompts could harm efficiency [28], increasing the "thinking" time unnaturally. LLMs are valuable for additional personality-related tasks, including augmentations to increase accuracy in text-based personality detection [29] and user experience enhancements in personality questionnaires [30]. In this work, we design a Socrates persona through LLM prompting and use it to generate real-time dialogue in our VR-based embodied conversational agent system. Through a user study, we analyze the effect of the system on personality perception, immersion, and social presence in correlation with user personality.

Our system inputs the user's speech, converted into text using OpenAI's Whisper Model [31]. Then, we utilize OpenAI's Chat Completions API to generate the agent's response, which is turned into speech using OpenAI's TTS solution. We utilize the Onyx voice to accompany the articulated Socrates model. We use the Oculus Lipsync Library to extract the corresponding mouth shapes from the generated speech, which controls the 3D model's shape keys for appropriate speech animations. The dialogue between the user and the agent is based on turns; the agent does not respond unless a user query is received and responds with one generation at a time. The user can interrupt the agent's speech using the controller. We implemented our system using Unity for VR and tested it on Meta Quest 2 (Meta Platforms Inc.). Our preliminary study was conducted at a university's philosophy festival, where we collected feedback on improving the application. Then, we conducted a user study with students to focus on the perceived personality of the Socrates agent and system performance in correlation to self-reported user personality.

The personality of the self can influence how people perceive others [32, 33]. This phenomenon inspired us to analyze the agent's perceived personality with the user's. Our conversational agent assumes specific personality traits by adopting Socrates' persona through LLM prompting; we investigate if the perceived factors for this persona are consistent and whether the perceived factors are affected by the user's self-personality. We employ a user study where participants freely talk with the agent, asking open-ended questions. After interacting with the agent, the participants answer the survey questions, in which we collect information on the participant's self-reported personality, the agent's perceived personality, realism, and social presence. We analyze the correlation between the measured qualities and report the significant findings. The results suggest a direct correlation between the user's emotional stability and the agent's perceived extraversion-agreeableness. We also observe an inverse relationship between the user's extraversion and the agent's perceived conscientiousness. The higher the user's extraversion, the better the perceived system's realism and immersion are. User conscientiousness has an inverse relation with the perceived ease of use and a direct correlation with the system's quality and realism. The system's performance received positive ratings in general.

This work's contribution includes an open-source VR conversational agent system utilizing LLM-based dialogue [1] and an in-depth analysis of the resulting system's performance in terms of personality expression, ease of use, response quality, realism, and immersion. We also analyze the influence of user personality on the perceived agent personality and system performance, emphasizing the interrelated nature of these perceptual qualities. The results of our study could inspire future work to focus on expressing specific personality factors in VR-based conversational agents to improve the overall perception.

2 | Related Work

2.1 | Large Language Models

Natural Language Processing (NLP), a subfield of AI, aims to build machines that can understand and communicate using human language. One major goal of this field is to achieve human likeness in conversational situations. LLMs show promising results in general conversation subjects [34]. In contrast to its predecessors, one key advantage of the LLM architecture is its use of the Transformer Attention Mechanism [35], which enables parallelization and can capture long-range dependencies in text. LLMs are used in many subjects, including education [36] and autonomous agents [37]. As with all machine learning models, high amounts of training data benefit LLMs, and currently, good quality text-based corpora are easy to collect in large sizes [38].

LLMs help with data labeling [39]; they can mediate user input to enhance other learning models in different fields, such as image generation [40].

The linguistic choices of LLMs can lead to personality expression; the users of LLM-based conversational systems can observe such cues as apparent personality traits. Rather than using naive prompting, utilizing specific traits associated with each personality dimension helps improve the expressivity of LLMs [25]. LLMs can assume the roles of consistent characters through prompting and generate responses that follow the linguistic style of that persona [15]. Integrating expert generators trained to exhibit different personality traits can improve the personality accuracy of the generated LLM responses [41]. The power of LLMs regarding personality and consistent persona expression inspired us to represent a historical figure in our application to analyze the consistency of the perceived system qualities. Specifically, we analyze how user personality affects the observed system qualities in VR-based conversational agents. While other studies, too, utilize LLM-based dialogue in conversational agent systems [42, 43], we differ in our application and analysis.

2.2 | Conversational Agents

Conversational agents precede data-driven language models. One major goal of conversational agents is to enable natural interaction with the user to assist with various tasks, and to this end, understanding user input with high accuracy is essential. Older conversational agent systems utilize AIML or Traditional NLP to map user input patterns to agent answers; more recent architectures utilize fully data-driven models. The appearance and nonverbal behavior of the conversational agents is an important factor in their realism [44]. Successful conversational systems often utilize appropriate gesturing and voice [45]. Certain types of gesturing could affect the personality traits observed in animated characters [21]. For example, fast movements are usually associated with extraversion, and reserved gestures signal introversion. The perceived personality traits of human animation can be altered using systematic adjustments focusing on different cues [20]. In addition to dialogue, voice, and motion, facial expressions are successful communicators of personality [22]; for example, agents that smile frequently score high in agreeableness.

VR-based agents improve social presence [46], immersive VR settings help agents better utilize gaze and spatial orientation to shape the conversational roles of human users [47]. Conversational agents are used in many different areas, including healthcare [48], tour guiding [49], education [50], and virtual assistance [51]; they can support learning in VR simulations [52] and help teach social conversational protocols [53]. Conversational systems can input user queries as speech [54] or text [55]; additional information such as the user gaze or facial expressions [56] enables more accurate responses. Nonverbal features such as posture and head direction can reveal the user's attention and psychological state during conversation [19], which can be utilized to adjust the agent's dialogue to create a more interesting interaction. In this work, we utilize speech-based user input, where the conversational agent responds to the user's queries in a voice-based manner. The agent's mouth animations accompany its speech, and we use generic talking animations for the rest of the body during the speech. While listening, the agent assumes an idle animation and looks at the user.

2.3 | Personality

Personality examines the observable characteristics that make up the individual. Theories group the common traits under different categories; two popular personality categorizations are the Five-Factor Model [57] and the Myers-Briggs Type Indicator (MBTI) [58]. In this work, we assume the Five-Factor Model of personality, which examines the individual based on five orthogonal dimensions:

- **Openness:** Reflects personality's imaginative, creative, and philosophical aspects. High openness relates to being intellectual, curious, and artistic. Low openness is associated with being traditional, conventional, and predictable. Being related to the intellectual aspect of personality, openness is hard to convey in short animated sequences through nonverbal communication [22]. Showing comprehension of complex subjects and asking curious questions can help express openness in dialogue.
- **Conscientiousness:** Comprises self-control, responsibility, and reliability. High conscientiousness corresponds to being diligent, hard-working, organized, and responsible. Low conscientiousness is associated with impulsive, careless, and disorganized behavior. Expressive animation systems rely on noisy motion that appears careless to represent low conscientiousness and smooth motion to express high conscientiousness [20]; thoughtful language could signal high conscientiousness in dialogue [59].
- **Extraversion:** The most common trait group among personality theories; measures sociability, assertiveness, and activeness. High extraversion is reflected as being confident, outgoing, energetic, and talkative. Low extraversion relates to being reserved, quiet, and passive. This factor is represented through the extent of gesturing and speed in animation [20, 21]. Dialogue models can focus on the talkative nature to represent high extraversion; short answers can signal low extraversion.
- **Agreeableness:** Describes the understanding, caring, and kindness to others. High agreeableness corresponds to being friendly, sympathetic, helpful, and considerate. Low agreeableness is manifested as cold, impolite, and rigid behavior. Using polite words can express high agreeableness in dialogue; happy facial expressions similarly represent high agreeableness [22]. Perceived agreeableness affects the likability of the virtual agents [60], contributing to the overall system performance.
- **Emotional Stability:** Explains how the individual regulates emotions and responds to threats. Some studies use it as neuroticism using the opposite polarity. High emotional stability relates to calm and relaxed behavior. Low emotional stability corresponds to being anxious, prone to negative emotions, and depressed. Relaxed movements express high emotional stability; indecisive and quick movements signal low emotional stability in animation [22].

3 | The System

We implemented our conversational agent system in Unity with the XR Interaction Toolkit to support VR. We render a simple 3D scene with a temple and a few trees in the background to reserve computing power for the animated agent. We aimed to keep a stable high frame rate to make the VR experience comfortable; low and unstable frame rates tend to cause a simulation sickness [61]. Our build target is the Meta Quest 2 VR headset, which we used in our user study, but the system is compatible with similar VR devices that support Unity XR. We utilize OpenAI's TTS, ASR, and Chat Completions APIs to handle voice-based natural conversation. These APIs require a stable internet connection but can generate almost immediate responses. We summarize our system in Figure 1. We utilize turn-based dialogue where the agent responds to user speech using TTS.

User speech is recorded using the device's internal microphone while the trigger button of the controller is pressed. We utilize text on a sphere object tracking the controller's position to give feedback about the system's current state: Waiting for input, recording, or thinking. When the user releases the trigger button, the system sends the voice recording to the speech recognition API to receive the transcription. If no meaningful words are recognized, the agent responds with a message indicating that the user should repeat their words. If the system successfully receives the user's speech transcription, we send it to the Chat Completions API with the user's role. The system prompt includes the message *"Act as Socrates. Your name is Socrates. Answer questions as Socrates."* for the responses to assume the Socrates persona. We observe that OpenAI's LLM successfully gives correct information about Socrates and talks in a Socrates-like philosophical manner without further prompting.

We use OpenAI's LLM-based text generation model GPT-3.5 Turbo. We do not have a token limit for the user input or agent output; however, the user speech recording has an upper bound of 60 s to prevent long speech recordings from introducing a substantial delay to the transcribing part of the pipeline. We utilize a message history of 10 so that the generated responses remember the last 10 dialogue lines. This helps limit the token usage as the API requires resending the message history for each turn of generated dialogue. We clear the message history once the VR headset is removed so that each participant can start the conversation from the beginning. In practice, we observe that most user speech recordings are less than half a minute, and the latency between the end of the user's speech recording and the agent's speech start is approximately 2 to 4 s. During this waiting time, the agent continues the idle animation, the thinking stage. This latency is shorter when the generated response or the user recording is shorter. The agent's response text is turned into speech using OpenAI's Speech Synthesis API, utilizing the Onyx voice. The 3D model of Socrates is designed in Blender and supports skeletal animation. The mouth of the model is controlled through shape keys corresponding to different visemes. We use the Oculus Lipsync Library to extract visemes from the agent's speech at runtime, which controls the facial shape keys of the Socrates model. The scene with the Socrates model from the user's perspective on Meta Quest 2 is shown in Figure 2; we utilize semi-realistic rendering due to the device's limitations.

Users freely talk with Socrates, and there is no additional system interaction to keep the focus on the conversation. The user can look around from the same position, but cannot move. The system is suitable for both standing and sitting user positions. There is no limit to the number of dialogue turns; the users can decide when to end the interaction by removing the VR headset. The initial test of the system was done during a university's philosophy festival, where students used the system during the event. During this initial test, we collected open-ended feedback from the users. Then, we performed a user study to evaluate the perceived

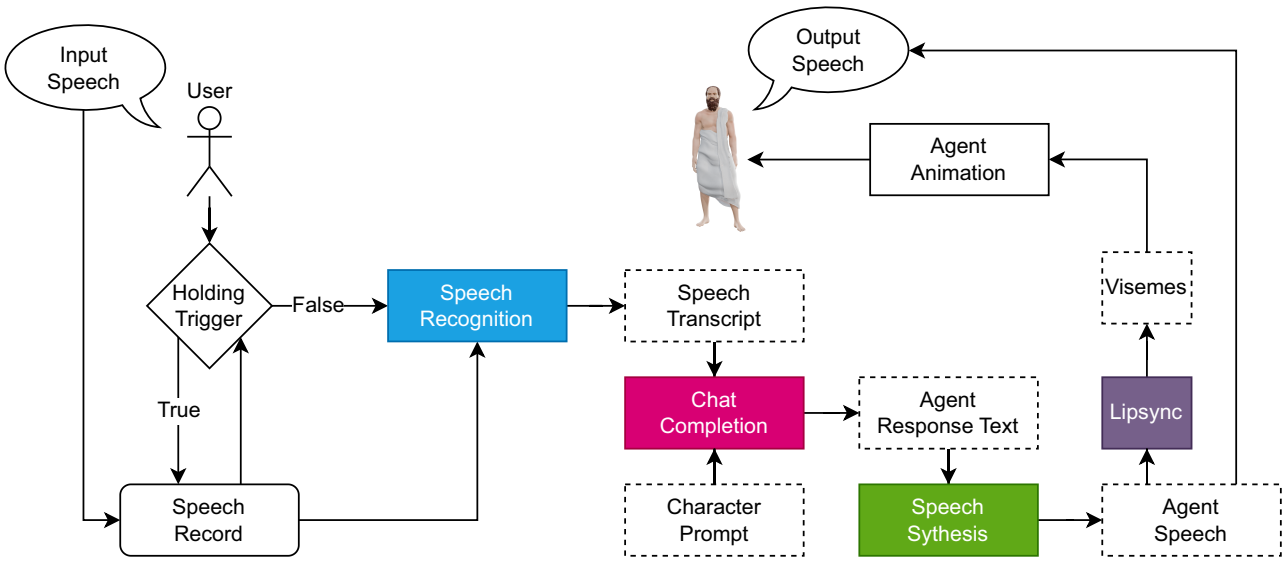

FIGURE 1 | The framework of our system. The user controls the speech recording process by holding the controller trigger. Releasing the trigger sends the recorded speech to the recognition system. The speech transcript is sent to the Chat Completions API with the predefined Socrates character prompt. The generated response is used to synthesize the agent's speech, which is also used in the lipsync system to generate visemes that drive the agent's mouth animation. The user can interrupt the process at any time by starting the recording; if no speech is recognized, the agent asks the user to speak again.

FIGURE 2 | Screenshot of the system running on Meta Quest 2. We include a temple and a few trees to increase immersion, but limit the background details to keep a stable frame rate. Most of the rendering power is reserved for the animated figure; although a performance load, we keep the shadows detailed for increased realism. A static environment texture is used for the sky.

agent personality and overall performance in correlation with user personality.

4 | Evaluation

We evaluate the performance regarding the personality perception of the LLM-based conversational agent, the system's ease of use, quality, realism, and immersion. We also collect the user's self-reported personality to observe any correlation between the measured aspects and the user's personality. We conducted the user study as part of a university's philosophy event. We used Turkish as the conversation language as this was the native language of the participants; however, OpenAI's ASR, TTS, and Chat Completions APIs support many other languages, including English. 24 users interacted with the system and answered our survey questions. Our survey included a total of 32 questions using their official translations: 10 for measuring the personality of Socrates using the Ten-Item Personality Inventory (TIPI) [62] (the same 10 questions repeated for the user's personality), 3 to measure the ease of using the system, 3 to measure the quality of the agent's answers, 4 to measure the realism of the agent, and 2 to measure the user's immersion. The questions are shown in Table 1; the questions regarding ease of use, quality, realism, and immersion are adapted from other studies involving virtual environments and digital characters [63]. While adapting the user study questions, we considered VR applications that utilize a single agent with which the user can interact. Since our application does not involve physical interaction, we did not include questions regarding interaction quality but focused on the appeal of the agent and the environment. The measured realism is separated into agent model, movement, voice, and environment. The ease-of-use questions focused on whether the participant encountered any problems during the interaction. The questions that measure the conversation quality focus on the naturalness of the dialogue.

After interaction with the system, participants rate the survey questions on a 5-point Likert scale. We first observe the answers by considering the mean scores for the user personality, the perceived agent personality, and the system qualities. The mean scores for the user personality questions show how well the participant pool represents the population. The perceived agent personality and the system quality are tightly bound with the LLM performance and the overall integration of the agent. We use Pearson correlation to analyze the relationship between the measured system parameters and the user personality. We also check if there is a correlation between the measured personality factors using the Kaiser-Meyer-Olkin (KMO) [64] test.

We illustrate the Likert scale histograms for the user personality questions in Figure 3. We observe that the participants mostly agree with the questions with direct scoring and disagree with the reverse-scored questions, showing that they generally score high for all the personality factors. A few participants have indecisive answers to each personality question. All 32 questions have significant answers with $p < 0.001$ compared to uniform randomness.

We illustrate the Likert scale histograms for the agent's perceived personality questions in Figure 4. We observe that Socrates received relatively high agreement for the conscientiousness factor. Reverse-scored factors, except for agreeableness, received strong disagreement. Questions 11 and 17 received a relatively high number of indecisive answers, suggesting neutral agreeableness and extraversion. In general, the agent's personality has positive connotations for all traits. We also observe that the users do not reflect their personality to the agent; this is mostly apparent in questions 1–11 for extraversion and 7–17 for agreeableness. The participants have much higher ratings for these factors, while the agent is more neutral.

We illustrate the Likert scale histograms for the remaining questions that measure the system's performance in Figure 5. We observe a greater tendency toward neutral. The questions measuring ease of use received the most agreement. The participants strongly disagree that the experience is uncomfortable (Q23). Participants slightly agree that the answers resemble a real person's (Q24), but how the agent answers them is like a robot (Q25). Users mostly did not have problems communicating with the agent (Q26). The questions measuring the system's realism received mostly slightly positive answers (Q27–30); the agent's most realistic part was its voice and pronunciation, with no disagreement for this question (Q29). The immersion was slightly positive but very close to neutral (Q31 and 32).

For an easy-to-read view, we map Likert scale answers to the integer range $[-2,2]$ and combine the ratings of the questions based on their categories. For personality-related questions, we use the standard scoring of TIPI, where each answer contributes to the corresponding factor with a positive or negative sign. Similarly, we sum the answers to the remaining questions with a sign based on whether they have positive or negative connotations,

TABLE 1 | Survey questions used in the evaluation focusing on different measurements. For measuring personality, we use TIPI [62]. The first 10 questions measure the user's personality, and the next 10 measure the agent's personality; the subject of the question changes between myself/Socrates. Each TIPI question measures one personality factor using direct or inverse scoring; these are indicated in parentheses and are not shown in the survey. The rest of the questions are adapted from studies involving virtual characters [63] and focus on ease of use of the system, the answer quality of the agent, realism, and immersion.

Category	No	Question
Personality	1/11	I see myself/Socrates as extraverted, enthusiastic. (Extraversion)
	2/12	I see myself/Socrates as critical, quarrelsome. (Reverse Agreeableness)
	3/13	I see myself/Socrates as dependable, self-disciplined. (Conscientiousness)
	4/14	I see myself/Socrates as anxious, easily upset. (Reverse Emotional Stability)
	5/15	I see myself/Socrates as open to new experiences, complex. (Openness)
	6/16	I see myself/Socrates as reserved, quiet. (Reverse Extraversion)
	7/17	I see myself/Socrates as sympathetic, warm. (Agreeableness)
	8/18	I see myself/Socrates as disorganized, careless. (Reverse Conscientiousness)
	9/19	I see myself/Socrates as calm, emotionally stable. (Emotional Stability)
	10/20	I see myself/Socrates as conventional, uncreative. (Reverse Openness)
Ease	21	I had no problems talking to Socrates.
	22	I adapted to the virtual reality experience easily.
	23	The experience in the virtual environment made me uncomfortable.
Quality	24	Socrates' answers resembled a real person.
	25	Socrates was giving answers like a robot.
	26	Socrates did not understand what I said correctly.
Realism	27	Socrates looked realistic.
	28	Socrates' movements were realistic.
	29	Socrates' pronunciation was correct, and his voice was realistic.
	30	The virtual environment was realistic.
Immersion	31	I felt like I was in the same environment as a real person.
	32	When the experience ended, I felt like I had returned to the "real world" after a journey.

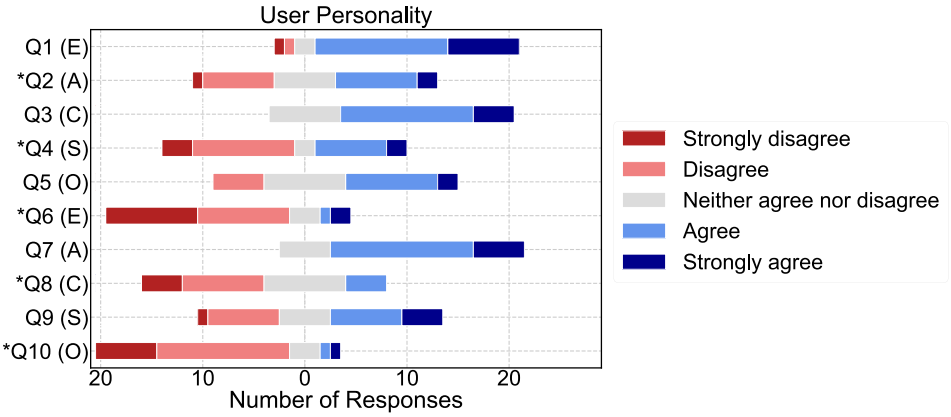

FIGURE 3 | Likert-scale plots for the first 10 questions measuring the user personality. Questions marked with * measure the corresponding personality factor reversely. Each factor is represented with one letter: Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (S), and Openness (O).

for example, for a negatively scored question like "Q23. The experience in the virtual environment made me uncomfortable." Strong disagreement (−2) contributes to a positive score.

Since each category has a different number of questions, we normalize the signed sums into the [−2, 2] range to report the corresponding box plots in Figure 6. We observe that the ratings are generally on the positive side. The users are rated high on extraversion and slightly neutral in emotional stability. The agent's emotional stability is much higher than that of the participants. Socrates is perceived as relatively high in openness and conscientiousness; these factors were found hard to convey in

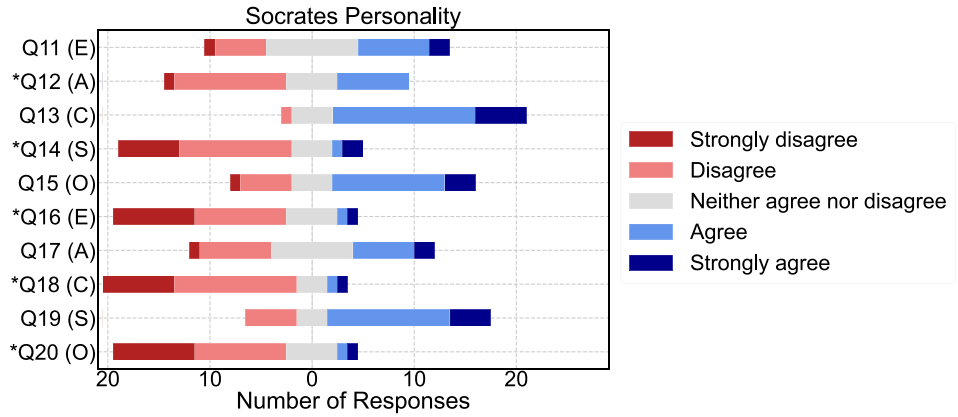

FIGURE 4 | Likert-scale plots for the second 10 questions measuring the agent personality. Questions marked with * measure the corresponding personality factor reversely. Each factor is represented with one letter: Extraversion (E), Agreeableness (A), Conscientiousness (C), Emotional Stability (S), and Openness (O).

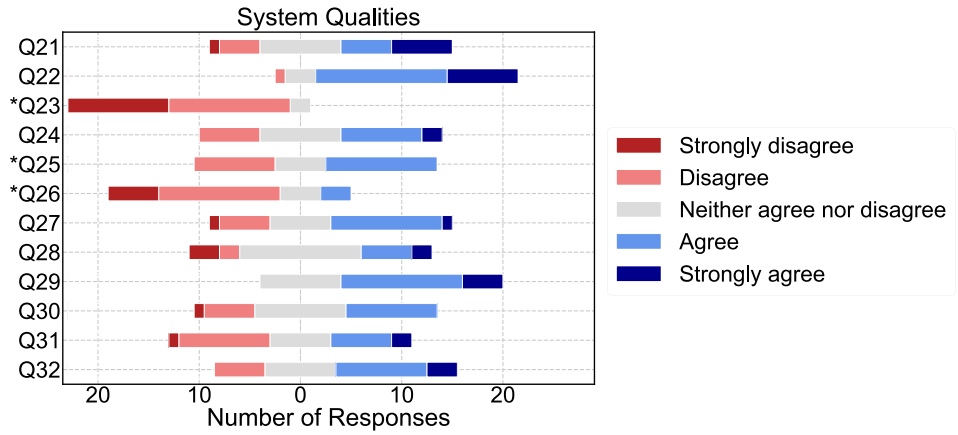

FIGURE 5 | Likert-scale plots for the second 10 questions that measure the system performance. Questions marked with * measure the corresponding quality in reverse. Thus, a successful system should receive disagreement to those questions.

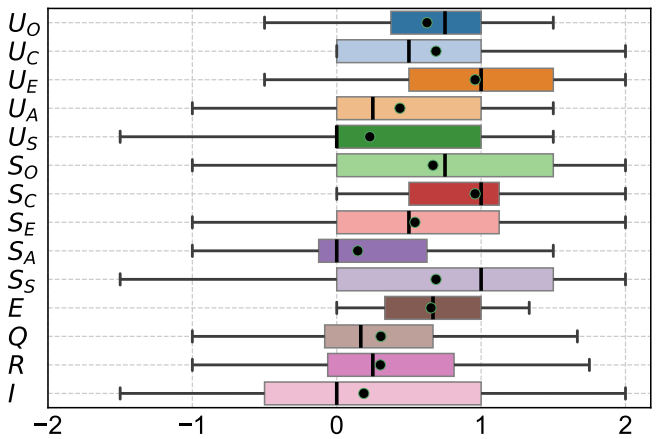

FIGURE 6 | The box plots of the measured aspects. The measured personality of the User (U) and Socrates (S) are reported in terms of Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Emotional Stability (S). We also report Ease of Use (E), Response Quality (Q), Realism (R), and Immersion (I). Measurements are signed sums for the related categories mapped into the [−2, 2] range; black lines depict the median, and the black dots show the mean.

earlier studies that do not utilize LLMs [20–22], we believe the opportunity to talk with the agent freely could emphasize the intellectual aspect of the agent. This could also be due to the persona the agent assumes. The participants were instructed to rate the agent rather than predict the personality of a real Socrates; however, we observed that the agent's perceived traits resemble a philosopher's, with a high rating in the intellectual aspect. Performing the study in a VR setting could also have helped users to focus better on the agent's apparent traits, which previous studies analyzing virtual agent personality lacked.

We hypothesize that how users perceive the system is affected by their personalities, following previous findings in psychology literature [32, 33]. To this end, we analyze the Pearson correlation between user personality and study measurements and report the correlation coefficients in Table 2. The sign of the coefficient indicates whether the variables are directly or inversely proportional, and its magnitude shows how strong the correlation is. The coefficients in the range [0.4, 0.6] are assumed as moderate. We observe that the user's extraversion has a moderate effect on the observed conscientiousness of the agent. Similarly, the user's emotional stability affects the perceived agent's extraversion and agreeableness. Significant correlations for the system qualities mostly relate to the user's conscientiousness and extraversion;

TABLE 2 | Pearson correlation between user personality and system measurements, values marked with * indicate $p < 0.05$. We report the correlation between the personalities of the User (U) and Socrates (S) in terms of Openness (O), Conscientiousness (C), Extraversion (E), Agreeableness (A), and Emotional Stability (S). The correlations to the system's Ease of Use (E), Response Quality (Q), Realism (R), and Immersion (I) are also reported.

	U_O	U_C	U_E	U_A	U_S
S_O	0.371	−0.078	0.344	−0.018	−0.104
S_C	0.104	0.329	−0.409*	−0.033	−0.114
S_E	−0.102	0.366	−0.161	−0.084	0.410*
S_A	0.119	0.216	0.206	0.062	0.474*
S_S	0.134	−0.025	−0.089	−0.119	−0.105
E	0.324	−0.406*	0.021	−0.070	−0.266
Q	−0.021	0.572*	0.110	0.171	0.242
R	0.361	0.594*	0.509 ∗	0.181	0.146
I	0.157	0.325	0.409 ∗	0.167	0.232

the user's conscientiousness moderately correlates with the system's perceived ease of use, quality, and realism. Extraversion of the user has a moderate effect on the perceived realism and immersion.

All the significant correlations, except for the effect on ease of use, are directly proportional. User conscientiousness is inversely proportional to the system's ease of use; highly conscientious users had problems talking to Socrates and did not easily adapt to the VR experience. This could have been caused by highly conscientious individuals being more attentive to the details; the semi-realistic rendering of the system could have caused a hard-to-adapt setting. The effect of conscientiousness on the measured system quality and realism can be explained with the same attention to detail. The tendency to use formal language in highly conscientious individuals [59] could help them to receive more quality answers, ultimately affecting the perception of realism. In contrast, individuals who are low in conscientiousness use informal language and thus could have received lower-quality answers.

The relationship between users' extraversion and immersion shows that extroverted individuals experience more immersion in VR settings, which confirms previous findings [65]. This could have improved the perceived realism of the system; the more immersed the user is, the more realistic the virtual environment will become. Users' extraversion also affects the perceived conscientiousness with inverse proportion; the social aspect of such individuals could help them engage in the conversation more and cause the agent's answers to appear less formal. It is also possible that the energetic nature of the high extraversion users prevents them from focusing on the conversation while judging conscientiousness; such users could have focused on the animations more to rate this trait. The emotional stability of the user has a moderate correlation with perceived extraversion and agreeableness. The relaxed nature of emotionally stable users could have triggered more positive answers from the agent, resulting in a high extraversion-agreeableness appearance. In contrast, anxious inputs from individuals low in emotional stability could cause the agent to respond more negatively, affecting its extraversion and agreeableness.

There can be correlations between the perceived personality factors of virtual characters, and this is likely because digital characters may lack sufficient information regarding different communication aspects. For example, if the character is not talking, certain intelligence-based personality factors like openness can be judged concerning other personality dimensions like extraversion [21]. To this end, we apply the Kaiser-Meyer-Olkin (KMO) [64] test to check if Principal Component Analysis (PCA) applies to our five-factor personality data as a measure for correlation. We obtain a KMO value of 0.595 for the user personality factors and 0.301 for the perceived agent personality. Following the consensus that $KMO < 0.5$ is unsuitable for PCA, we observe a fairly low correlation between the agent's personality factors. This suggests the participants deduced each factor of the agent's personality in an orthogonal manner. This could be due to the verbal capabilities of the agent. Additionally, interacting with the agent in a VR setting may improve personification, helping users to observe different personality factors better.

5 | Conclusion and Future Work

This work focused on analyzing perceived personality in LLM-based conversational agents in VR. We designed a Socrates persona using OpenAI's Chat Completions API and a fully animated 3D model. Our user study involved participants talking with Socrates freely using natural language, to which Socrates responded in synthesized speech. We measured user and perceived agent personalities, ease of use, response quality, realism, and immersion. The participants generally observed Socrates as high in conscientiousness and emotional stability, followed by slightly high openness and extraversion. The agent is found to be neutral in agreeableness. The system is rated high in ease of use, and we observe slightly high quality, realism, and immersion.

The Pearson correlation analysis suggests that the user's extraversion inversely influences the perceived conscientiousness of the agent with a moderate effect; this factor also directly correlates with the system's perceived realism and immersion. The user's emotional stability moderately affects the perceived extraversion and agreeableness. User conscientiousness has an inverse correlation with the system's perceived ease of use and a direct correlation with the measured quality and realism. These findings could inspire future work to improve certain design aspects of VR conversational agents to focus on specific user types. For example, high system realism can be targeted at highly conscientious users. Similarly, high extroversion users could better relate to more immersive environments. Expressing high extraversion and agreeableness with the agent could help communicate with users who are high in emotional stability. We also found an inverse relation between the user's extraversion and the agent's perceived conscientiousness. Less conscientious behavior could help relate to users with high extraversion; in contrast, highly conscientious agent behavior could better communicate with introverted users.

The participants perceived the system's performance as positive in general; however, there is room for improvement regarding

quality, realism, and immersion. Future studies could utilize more advanced rendering to increase realism and more ways to interact with the agent to improve immersion. Currently, the user can only input speech and cannot freely move inside the virtual environment, which could hurt immersion. Interacting with the agent using gestures can be an interesting future direction. The agent can react to the user's facial expression tracked in VR [66]. The device's tilt could indicate the user's posture [67], which could predict the user's current mood. Similarly, the user's speech style can be utilized so that the agent can respond with appropriate language. For example, if the user seems sad, the agent could use a more encouraging tone. Such improvements could help digital characters better connect with users of various personalities.

Acknowledgments

We thank M.D. Mete Akel and Dr. Tufan Kiymaz for giving us the initial idea for initiating the project and providing the necessary equipment. We also thank Erhan Tunalı and the Graphic Design Department of Bilkent University for providing the 3D Socrates model.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Endnotes

[1] The public repository is available on https://github.com/sinansonlu/socrates-vr, including the system and analysis.

References

1. F. A. Mikic, J. C. Burguillo, M. Llamas, D. A. Rodríguez, and E. Rodríguez, "CHARLIE: An AIML-Based Chatterbot Which Works as an Interface Among INES and Humans," in *Proceedings of the Annual Conference of the European Association for Education in Electrical and Information Engineering (EAEEIE '09)* (IEEE, 2009), 1–6.

2. R. S. Wallace, *The Anatomy of ALICE* (Springer, 2009).

3. G. De Gasperis, I. Chiari, and N. Florio, "AIML Knowledge Base Construction From Text Corpora," in *Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing* (Springer, 2013), 287–318.

4. B. Omarov, Z. Zhumanov, A. Gumar, and L. Kuntunova, "Artificial Intelligence Enabled Mobile Chatbot Psychologist Using AIML and Cognitive Behavioral Therapy," *International Journal of Advanced Computer Science and Applications* 14, no. 6 (2023): 137–146, https://doi.org/10.14569/IJACSA.2023.0140616.

5. A. M. Galvão, F. A. Barros, A. M. Neves, and G. L. Ramalho, "Adding Personality to Chatterbots Using the Persona-AIML Architecture," in *Advances in Artificial Intelligence–IBERAMIA 2004: 9th Ibero-American Conference on AI, Puebla, Mexico, November 22–26, 2004. Proceedings*, vol. 9 (Springer, 2004), 963–973.

6. T. J. Sejnowski, "Large Language Models and the Reverse Turing Test," *Neural Computation* 35, no. 3 (2023): 309–342.

7. J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "ChatGPT and Large Language Model (LLM) Chatbots: The Current State of Acceptability and a Proposal for Guidelines on Utilization in Academic Medicine," *Journal of Pediatric Urology* 19, no. 5 (2023): 598–604.

8. H. X. Qin, S. Jin, Z. Gao, M. Fan, and P. Hui, "CharacterMeet: Supporting Creative Writers' Entire Story Character Construction Processes Through Conversation With LLM-Powered Chatbot Avatars," in *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)* (ACM, 2024), 1051.

9. Y. Wang, Q. Zhou, and D. Ledo, "StoryVerse: Towards co-Authoring Dynamic Plot With LLM-Based Character Simulation via Narrative Planning," in *Proceedings of the 19th International Conference on the Foundations of Digital Games (FDG '24)* (ACM, 2024).

10. J. Roberts, M. Baker, and J. Andrew, "Artificial Intelligence and Qualitative Research: The Promise and Perils of Large Language Model (LLM) 'Assistance'," *Critical Perspectives on Accounting* 99 (2024): 102722.

11. Q. Wen, J. Liang, C. Sierra, et al., "AI for Education (AI4EDU): Advancing Personalized Education With LLM and Adaptive Learning," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)* (ACM, 2024), 6743–6744.

12. P. Wan, Z. Huang, W. Tang, et al., "Outpatient Reception via Collaboration Between Nurses and a Large Language Model: A Randomized Controlled Trial," *Nature Medicine* 30 (2024): 2878–2885.

13. T. Xie, Y. Kuang, Y. Tang, J. Liao, and Y. Yang, "Using LLM-Supported Lecture Summarization System to Improve Knowledge Recall and Student Satisfaction," *Expert Systems with Applications* 269 (2025): 126371.

14. L. Liao, G. H. Yang, and C. Shah, "Proactive Conversational Agents in the Post-ChatGPT World," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23)* (ACM, 2023), 3452–3455.

15. G. Sun, X. Zhan, and J. Such, "Building Better AI Agents: A Provocation on the Utilisation of Persona in LLM-Based Conversational Agents," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CHI '24)*, vol. 35 (ACM, 2024), 6.

16. S. Sonlu, B. Bendiksen, F. Durupinar, and U. Güdükbay, "The Effects of Embodiment and Personality Expression on Learning in LLM-Based Educational Agents." arXiv preprint arXiv:2407.10993 (2024).

17. J. Wester, S. De Jong, H. Pohl, and N. van Berkel, "Exploring People's Perceptions of LLM-Generated Advice," *Computers in Human Behavior: Artificial Humans* 2, no. 2 (2024): 100072.

18. J. Cook and G. Salvendy, "Perception of Computer Dialogue Personality: An Exploratory Study," *International Journal of Man-Machine Studies* 31, no. 6 (1989): 717–728.

19. A. Cerekovic, O. Aran, and D. Gatica-Perez, "How Do You Like Your Virtual Agent? Human-Agent Interaction Experience Through Nonverbal Features and Personality Traits," in *Human Behavior Understanding: 5th International Workshop, HBU 2014, Zurich, Switzerland, September 12, 2014. Proceedings 5, Lecture Notes in Computer Science*, vol. 8749 (Springer, 2014), 1–15.

20. F. Durupinar, M. Kapadia, S. Deutsch, M. Neff, and N. I. Badler, "PERFORM: Perceptual Approach for Adding OCEAN Personality to Human Motion Using Laban Movement Analysis," *ACM Transactions on Graphics* 36, no. 1 (2016): 6.

21. H. J. Smith and M. Neff, "Understanding the Impact of Animated Gesture Performance on Personality Perceptions," *ACM Transactions on Graphics* 36, no. 4 (2017): 49, https://doi.org/10.1145/3072959.3073697.

22. S. Sonlu, U. Güdükbay, and F. Durupinar, "A Conversational Agent Framework With Multi-Modal Personality Expression," *ACM Transactions on Graphics* 40, no. 1 (2021): 7.

23. F. Mairesse and M. A. Walker, "Controlling User Perceptions of Linguistic Style: Trainable Generation of Personality Traits," *Computational Linguistics* 37, no. 3 (2011): 455–488.

24. A. Gupta, X. Song, and G. Anumanchipalli, "Self-Assessment Tests Are Unreliable Measures of LLM Personality," in *Proceedings of the 7th*

BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP (Association for Computational Linguistics, 2024), 301–314.

25. G. Jiang, M. Xu, S. C. Zhu, W. Han, C. Zhang, and Y. Zhu, "Evaluating and Inducing Personality in Pre-Trained Language Models," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)* (Curran Associates, Inc., 2024), 10622–10643.

26. Y. Tsubota and Y. Kano, "Text Generation Indistinguishable From Target Person by Prompting Few Examples Using LLM," in *Proceedings of the 2nd International AIWolfDial Workshop. Association for Computational Linguistics* (Association for Computational Linguistics (ACL), 2024), 13–20.

27. L. Sun, T. Qin, A. Hu, et al., "Persona-L Has Entered the Chat: Leveraging LLM and Ability-Based Framework for Personas of People With Complex Needs," 2024, https://arxiv.org/abs/2409.15604.

28. L. Li, Y. Zhang, and L. Chen, "Prompt Distillation for Efficient LLM-Based Recommendation," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)* (ACM, 2023), 1348–1357.

29. L. Hu, H. He, D. Wang, Z. Zhao, Y. Shao, and L. Nie, "LLM vs Small Model? Large Language Model Based Text Augmentation Enhanced Personality Detection Model," *Proceedings of the AAAI Conference on Artificial Intelligence* 38, no. 16 (2024): 18234–18242.

30. J. Lee, Y. Choi, M. Song, and S. Park, "ChatFive: Enhancing User Experience in Likert Scale Personality Test Through Interactive Conversation With LLM Agents," in *Proceedings of the 6th ACM Conference on Conversational User Interfaces (CHI '24)* (ACM, 2024), 36.

31. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the International Conference on Machine Learning (ICML '23)* (PMLR, 2023), 28492–28518.

32. S. Shrauger and J. Altrocchi, "The Personality of the Perceiver as a Factor in Person Perception," *Psychological Bulletin* 62, no. 5 (1964): 289–308.

33. H. Markus, J. Smith, and R. L. Moreland, "Role of the Self-Concept in the Perception of Others," *Journal of Personality and Social Psychology* 49, no. 6 (1985): 1494–1512.

34. Y. Chang, X. Wang, J. Wang, et al., "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology* 15, no. 3 (2024): 39, https://doi.org/10.1145/3641289.

35. A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is all You Need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)* (Curran Associates Inc., 2017), 6000–6010.

36. E. Kasneci, K. Seßler, S. Küchemann, et al., "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education," *Learning and Individual Differences* 103 (2023): 102274.

37. L. Wang, C. Ma, X. Feng, et al., *A Survey on Large Language Model Based Autonomous Agents*, vol. 18 (Frontiers of Computer Science, 2024), 186345.

38. M. Wan, T. Safavi, S. K. Jauhar, et al., "TnT-LLM: Text Mining at Scale With Large Language Models," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)* (ACM, 2024), 5836–5847.

39. X. Wang, H. Kim, S. Rahman, K. Mitra, and Z. Miao, "Human-LLM Collaborative Annotation Through Effective Verification of LLM Labels," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '23)* (ACM, 2024), 303.

40. L. Qu, S. Wu, H. Fei, L. Nie, and T. S. Chua, "LayoutLLM-T2I: Eliciting Layout Guidance From LLM for Text-To-Image Generation," in *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)* (ACM, 2023), 643–654.

41. W. Li, J. Liu, A. Liu, X. Zhou, M. Diab, and M. Sap, "BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data," (2024), https://arxiv.org/abs/2410.16491.

42. B. Wang, G. Li, and Y. Li, "Enabling Conversational Interaction With Mobile UI Using Large Language Models," in *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '23)* (ACM, 2023), 432.

43. G. Li, H. Hammoud, H. Itani, D. Khizbullin, and B. Ghanem, "CAMEL: Communicative Agents for Mind Exploration of Large Language Model Society," *Advances in Neural Information Processing Systems* 36 (2023): 51991–52008.

44. Y. Ferstl, S. Thomas, C. Guiard, C. Ennis, and R. McDonnell, "Human or Robot? Investigating Voice, Appearance and Gesture Motion Realism of Conversational Social Agents," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents (IVA '21)* (ACM, 2021), 76–83.

45. D. Parmar, S. Olafsson, D. Utami, P. Murali, and T. Bickmore, "Designing Empathic Virtual Agents: Manipulating Animation, Voice, Rendering, and Empathy to Create Persuasive Agents," *Autonomous Agents and Multi-Agent Systems* 36, no. 1 (2022): 17.

46. M. Guimarães, R. Prada, P. A. Santos, J. Dias, A. Jhala, and S. Mascarenhas, "The Impact of Virtual Reality in the Social Presence of a Virtual Agent," in *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20)* (ACM, 2020), 23.

47. T. Pejsa, M. Gleicher, and B. Mutlu, "Who, Me? How Virtual Agents Can Shape Conversational Footing in Virtual Reality," in *Intelligent Virtual Agents: 17th International Conference, IVA 2017, Stockholm, Sweden, August 27–30, 2017, Proceedings 17* (Springer Nature, 2017), 347–359.

48. J. L. Z. Montenegro, C. A. de Costa, and R. de Rosa Righi, "Survey of Conversational Agents in Health," *Expert Systems with Applications* 129 (2019): 56–67.

49. S. Kopp, L. Gesellensetter, N. C. Krämer, and I. Wachsmuth, "A Conversational Agent as Museum Guide–Design and Evaluation of a Real-World Application," in *Intelligent Virtual Agents: 5th International Working Conference, IVA 2005, Kos, Greece, September 12-14, 2005. Proceedings*, vol. 5 (Springer, 2005), 329–343.

50. A. Kerry, R. Ellis, and S. Bull, "Conversational Agents in E-Learning," in *Applications and Innovations in Intelligent Systems XVI, SGAI '08*, ed. T. Allen, R. Ellis, and M. Petridis (Springer, 2009), 169–182.

51. M. Mekni, "An Artificial Intelligence Based Virtual Assistant Using Conversational Agents," *Journal of Software Engineering and Applications* 14, no. 9 (2021): 455–473.

52. C. P. Dai, F. Ke, N. Zhang, et al., "Designing Conversational Agents to Support Student Teacher Learning in Virtual Reality Simulation: A Case Study," in *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (ACM, 2024), 513.

53. S. Babu, E. Suma, T. Barnes, and L. F. Hodges, "Can Immersive Virtual Humans Teach Social Conversational Protocols?," in *Proceedings of the IEEE Annual International Symposium Virtual Reality (VR '07)* (IEEE, 2007), 215–218.

54. M. Porcheron, J. E. Fischer, M. McGregor, et al., "Talking With Conversational Agents in Collaborative Action," in *Companion of the ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)* (ACM, 2017), 431–436.

55. L. Laranjo, A. G. Dunn, H. L. Tong, et al., "Conversational Agents in Healthcare: A Systematic Review," *Journal of the American Medical Informatics Association* 25, no. 9 (2018): 1248–1258.

56. D. Aneja, R. Hoegen, D. McDuff, and M. Czerwinski, "Understanding Conversational and Expressive Style in a Multimodal Embodied Conversational Agent," in *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (ACM, 2021), 102.

57. J. M. Digman, "Personality Structure: Emergence of the Five-Factor Model," *Annual Review of Psychology* 41, no. 1 (1990): 417–440.

58. K. C. Briggs, *Myers-Briggs Type Indicator* (Consulting Psychologists Press, 1976).

59. F. Mairesse and M. A. Walker, *Can Conversational Agents Express Big Five Personality Traits Through Language?: Evaluating a Psychologically-Informed Language Generator* (Cambridge University Engineering, 2009).

60. S. H. Kang, J. Gratch, N. Wang, and J. H. Watt, "Agreeable People Like Agreeable Virtual Humans," in *Intelligent Virtual Agents: 8th International Conference, IVA 2008, Tokyo, Japan, September 1–3, 2008, Proceedings*, ed. H. Prendinger, J. Lester, and M. Ishizuka (Springer, 2008), 253–261.

61. J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H. N. Liang, "Effect of Frame Rate on User Experience, Performance, and Simulator Sickness in Virtual Reality," *IEEE Transactions on Visualization and Computer Graphics* 29, no. 5 (2023): 2478–2488.

62. S. D. Gosling, P. J. Rentfrow, and W. B. Swann, Jr., "A Very Brief Measure of the Big-Five Personality Domains," *Journal of Research in Personality* 37, no. 6 (2003): 504–528.

63. L. Bareišytė, S. Slatman, J. Austin, et al., "Questionnaires for Evaluating Virtual Reality: A Systematic Scoping Review," *Computers in Human Behavior Reports* 16 (2024): 100505.

64. C. D. Dziuban and E. C. Shirkey, "When Is a Correlation Matrix Appropriate for Factor Analysis? Some Decision Rules," *Psychological Bulletin* 81, no. 6 (1974): 358–361.

65. D. Weibel, B. Wissmath, and F. W. Mast, "Immersion in Mediated Environments: The Role of Personality Traits," *Cyberpsychology, Behavior, and Social Networking* 13, no. 3 (2010): 251–256.

66. S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa, "Eyemotion: Classifying Facial Expressions in VR Using Eye-Tracking Cameras," in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV '19)* (IEEE, 2019), 1626–1635.

67. D. C. Jeong, J. J. Xu, and L. C. Miller, "Inverse Kinematics and Temporal Convolutional Networks for Sequential Pose Analysis in VR," in *Proceedings of the IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR '20)* (IEEE, 2020), 274–281.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Data S1.** Supplementary Video Legend.