# Effects of Embodiment and Personality in LLM-Based Conversational Agents

Sinan Sonlu*
Bilkent University

Bennie Bendiksen †
University of Massachusetts Boston

Funda Durupinar ‡
University of Massachusetts Boston

Uğur Güdükbay §
Bilkent University

Figure 1: The four alternative 3D agent models used in the study expressing high (left group) and low (right group) traits.

**ABSTRACT**

This work investigates the effects of personality expression and embodiment in conversational agents. We extend a personality-driven conversational agent framework by integrating LLM-based conversation support to provide information about contemporary scientific topics. We describe a user study built on this system to evaluate two opposing personality styles using three models: a dialogue-only model that conveys personality verbally, an animated human model that expresses personality only through dialogue, and an animated human model expressing personality through dialogue and expressive animations. The users perceive all models positively regarding personality and learning outcomes; however, models with high personality traits are perceived as more engaging than those with low personality traits. We provide an analysis of personality perception, learning, and user experience.

**Index Terms:** Five-factor personality, Generative Pre-trained Transformer (GPT), Large Language Model (LLM), Conversational agent, Dialogue, Character animation.

## 1 INTRODUCTION

Virtual agents have tremendous opportunities to provide personalized, on-demand experiences in domains such as education and healthcare, particularly in immersive virtual environments. Their function is more than just relaying information; they can socially connect with users, establish rapport, and motivate them [46]. The advancement of Large Language Models (LLMs) has significantly increased virtual agents' power, enabling them to both understand and respond to natural language queries. As LLMs can effectively assume various roles and personalities, virtual agents with LLM-driven dialogue capabilities have the potential to offer customized experiences for users with diverse preferences and needs. Such potential necessitates a deeper understanding of agent characteristics that make them more likable and effective in conversation. This work investigates two of these characteristics: embodiment and personality. Users perceive embodied agents as more trustworthy, engaging, and socially present compared to disembodied agents [18, 4]. Studies report that conversational agents increase motivation and enjoyment in Virtual Reality (VR) learning applications [37]. However, the implications of personality-enhanced behavior in LLM-based conversational systems are not yet fully understood. This work explores how virtual agents' embodiment and personality expression affect the self-assessment of learning, engagement, and quality outcomes in a conversational educational application, where agents act as "learning objects" that support the learning of specific concepts [17].

We extend an existing personality-driven conversational agent framework [43] with LLM-based conversation support tailored for an educational scenario. We run a user study wherein participants interact with the system by typing their questions about a conversation topic to which the agent responds verbally. We use 3D agent models as they display a more comprehensive range of gestures and complex facial expressions [7] compared to their 2D counterparts. Furthermore, 3D avatars are found to be more compelling and impactful [27] as they improve presence, immersion [59], and the overall experience in virtual environments [9, 38]. They naturally fit within VR scenes and are more likely than 2D agents to be perceived as part of the VR experience [51].

We focus on the combined effect of extraversion and agreeableness, as gestures and facial expressions convey these traits more effectively than the other three personality factors [43]. We refer to this combination as the agent's 'personality style.' A high-trait style combines high extraversion and agreeableness, resulting in a friendly, lively, and energetic agent. In contrast, a low-trait style reflects low extraversion and agreeableness, creating a more reserved and less approachable agent. This focus on combining traits with the same polarity allows us to simulate the 'Big One' effect [29], which captures an overall sense of positive personality. We favor this approach over comparing individual traits separately, as our conversational agent system expresses different traits with varying degrees of effectiveness. For instance, comparing a highly expressive extraverted agent with one that expresses a less prominent factor, such as openness, would not result in a fair comparison due to these variations in expressiveness.

We assess embodiment using three models (see Fig. 2): a dialogue-only model and two models with 3D humanoid bodies.

*e-mail: sinan.sonlu@bilkent.edu.tr
†e-mail: b.bendiksen001@umb.edu
‡e-mail: funda.durupinarbabur@umb.edu
§e-mail: gudukbay@cs.bilkent.edu.tr

2642-5254/25/$31.00 ©2025 IEEE
DOI 10.1109/VR59515.2025.00094

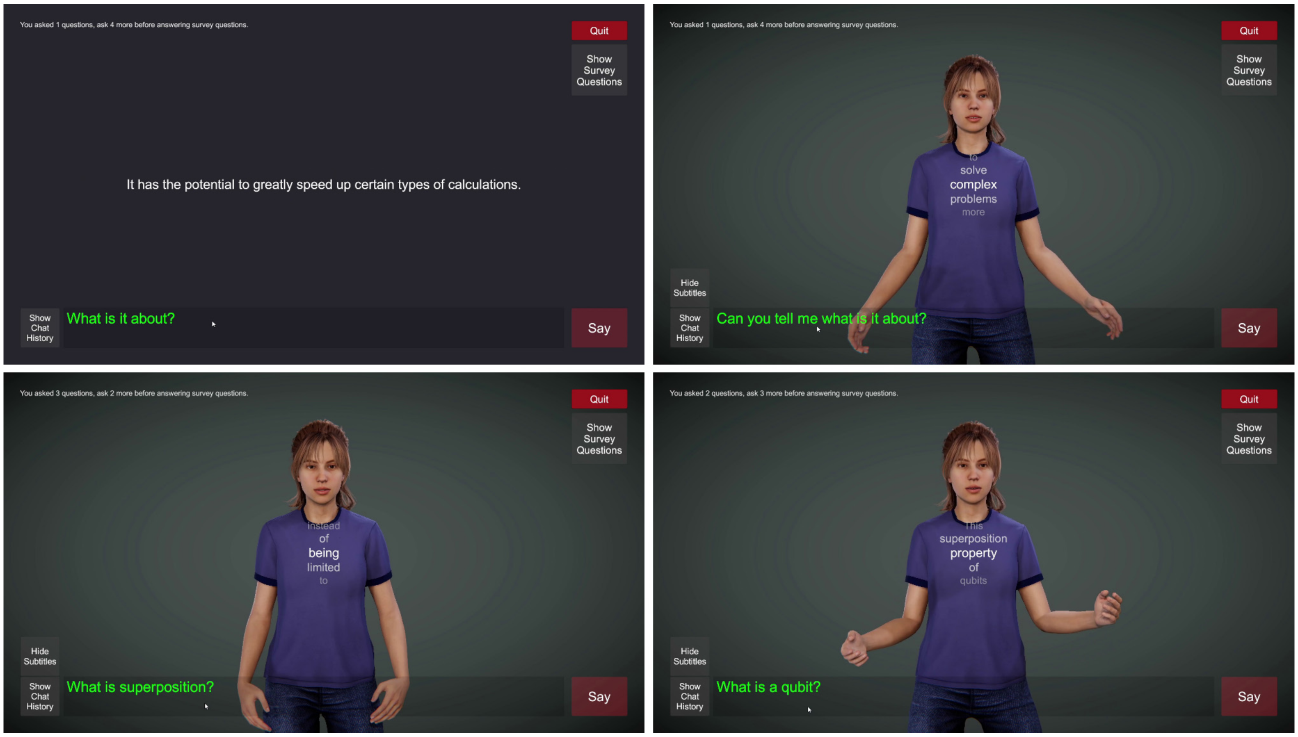


Figure 2: Screenshots of different models and variations: Model D (top left, dialogue only), Model A (top right, non-expressive, same animation for high and low traits), Model E-Low (bottom left, low-trait expressive agent), and Model E-High (bottom right, high-trait expressive agent).

All the models display conversation text concurrently with audio feedback. We evaluate the efficacy of different modalities and personality styles through an independent-subjects user study. The study randomly presented each participant with a high or low personality variant of each model. The dialogue-only model and one of the embodied models express personality only through text, and the other embodied model expresses personality through body movements, facial expressions, and gaze, in addition to text. During the study, we collected ratings about the system for self-assessment of learning, quality, and engagement, as well as the perceived personalities of the agents. Additionally, we obtained user feedback through open-ended questions.

Although the system parameters were selected to express certain personality traits, we considered potential variations in participants' perceptions of the agents' personalities. For instance, a high-trait agent might also be perceived as emotionally stable, or a dialogue-only agent might be viewed as conscientious, even though these traits were not intentionally highlighted. Therefore, we collected users' perceptions of the agents' personalities across all five dimensions of the FFM for each of the three models.

This work aims to answer the following research questions:

**RQ1.** Is there an effect of personality style on perceived personality?
**RQ2.** Is there an effect of model type on perceived personality?
**RQ3.** Is there an effect of model type on learning outcomes, i.e., assessment of the agent as a learning object?
**RQ4.** Is there a correlation between learning outcomes and personality perception?

Based on the findings of the previous studies, we formulate the following hypotheses:

**H1.** Learning object ratings will be higher for the embodied agents than the dialogue-only agent, reflected as higher scores in H1a. self-assessment of learning; H1b. quality; H1c. engagement. Since the literature indicates a more positive approach towards embodied agents than disembodied ones, we expect a similar tendency in our application; the embodied agents will be more engaging and effective [37].

**H2.** Agents expressing high extraversion and agreeableness will be rated higher for the outcomes than the agents expressing low extraversion and agreeableness, reflected as higher ratings in H2a. learning; H2b. quality; H2c. engagement. We formulate this hypothesis based on the documented preferences of users for highly agreeable chatbots [50].

In addition to investigating these questions through quantitative analysis, we identify common themes and individual differences across participants by an in-depth qualitative analysis of their responses to open-ended questions. Furthermore, we provide our system as an open-source virtual tutoring application with conversational virtual agents that exhibit desired personality traits via motion and language. An overview of our study is depicted in Fig. 3. Our system combines existing approaches to personality expression and embodiment in a novel way that can be further studied and extended. Our data and code are available in our public repository [1]

## 2 Related Work

Involving multiple computing fields, this work on conversational agents combines personality expression and LLM-based dialogue generation. Although the current system involves a desktop application, as the user study necessitates, the implementation is straightforward to adapt to an immersive VR setting, which brings advantages and challenges, as described in this section.

**Conversational Agents.** Conversational agents use computational linguistics techniques to interpret and respond to user statements in ordinary natural language; understanding and exhibiting emotions

[1] https://github.com/sinansonlu/LLM-Agent

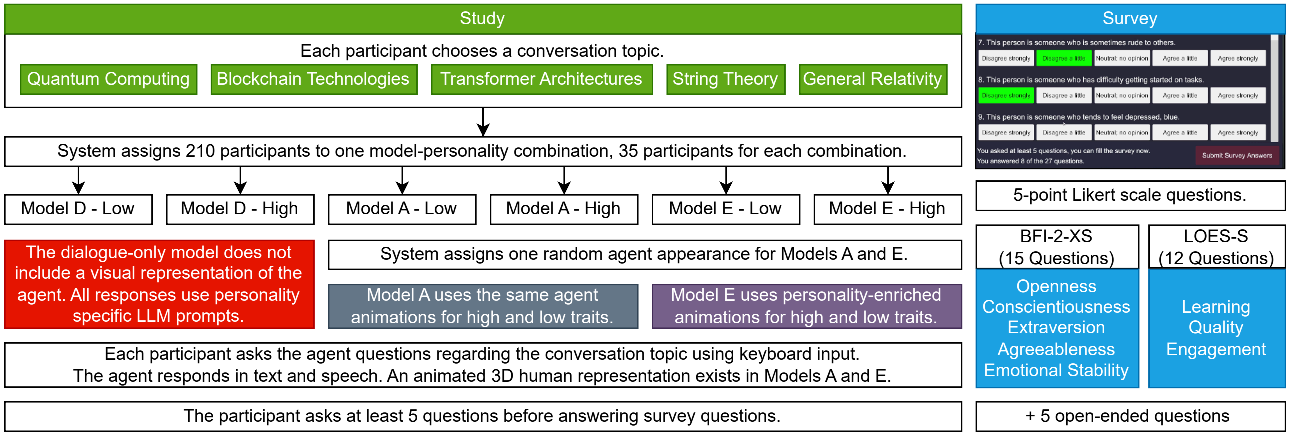


Figure 3: The overview of our study.

and personality are essential for successful natural language conversations [5]. For example, the same query may require different interpretations based on the user's mood. Similarly, the same response can be perceived differently based on the agent's body language and facial expression. Users perceive conversational systems that give relevant answers to their questions as more human-like and engaging [41]. Studies synthesize gesture animation to accompany speech [34], which can help achieve motion appropriateness. Co-speech gestures offer means of tracking conversation topics; specific gestures can suggest changes of subject [21]. Conversational agents can respond to user language, gesture, and affect [20]; they can recognize gaze, speech, and facial expressions.

**VR Agents.** VR-based systems enable a natural connection to the virtual environment, isolating users from the real world and surrounding their views with computer-generated imagery that helps improve the social presence of virtual agents [13]. Co-presence and realism of VR agents can moderate social facilitation, which influences the user's performance of the given tasks [45]. VR systems with eye-tracking enable gaze and emotion estimation [14], and agents can imitate the user's emotions using the collected information, which enhances affective human-agent communication [33]. For instance, multi-modal human agents in VR shopping experiences have been shown to improve users' warmth, communication, trust, and satisfaction [57]. Gaze behaviors and spatial orientation of VR agents can shape the conversational roles of human users as speakers, addressees, bystanders, and overhearers [36]. VR systems also enable full-body pose estimation from limited sensors [3], which can help integrate the user's posture as an input. VR-based agent systems generally offer more varied user input and induce a better sense of presence, improving interaction quality and empathy. Realistic VR agents have been associated with reports of high enjoyment and significant knowledge acquisition in pedagogical settings. However, their behavioral realism, as exhibited by gesturing, eye contact, speech, and lip synchronization, was shown to diminish factual knowledge acquisition [37]. Perception of VR agents is complicated: factors such as their rendering styles and personality expression behaviors interact in complex ways [58].

**Personality Expression in Agents.** Studies use nonverbal behavior elements to convey personality [39] and leverage high-level motion meanings to express the target traits [1]. For instance, PERFORM establishes a link between Laban Movement Analysis (LMA) parameters and the perceived personality of virtual human characters [10]. Personality-specific voice, dialogue, and facial expressions help distinguish opposing personality traits in expressive conversational agents [43]. Perceived personality influences users' attitude; users are more willing to trust and listen to serious-looking, assertive agents [56]. Gesture performance in combination with language highly influences perceived personality [30]. Linguistic elements such as the ratio of phrases, words of emotion, and exclamations correlate with personality traits [24]; similarly, rendering style plays an important role in personality perception [54].

**LLM-Based Agents.** LLMs started to play critical roles in innovative technologies [16]. For instance, LLM-based vocalized agents aid students in foreign language learning in Augmented Reality (AR) environments [49]. Additionally, LLMs enhance patient experiences during consultation, diagnosis, and management in healthcare [53]. Although LLMs can generate highly sophisticated responses, they lack access to dynamic content. Consequently, LLM-based agent systems often focus on isolated tasks, such as answering questions based on pre-existing knowledge or performing data-driven qualitative analysis [55]. LLMs such as GPT exhibit consistent personality cues and offer customization for assuming different personalities [15]. Systems can predict different personality types in LLMs using certain prompts [26], supporting LLMs can capture language-based personality cues. Distinguishable knowledge levels of LLM-based virtual agents influence perceptions of intelligence, rapport, and willingness for future interaction [52].

## 3 Method

This section describes a user study[2] to assess the impact of different modalities and personality parameters on the perceived personality of agents and self-assessed learning parameters. For this, we designed an application employing a conversational agent that gives information about contemporary scientific topics through turn-based dialogue.

### 3.1 System

To run our study, we updated the personality-driven conversational agent platform by Sonlu et al. [43], an open-source, multi-modal system for animating 3D conversational virtual agents through controlling facial expressions and body movements based on an input personality. The platform modifies a base animation via joint rotation and animation speed adjustments, noise addition, and inverse kinematics-based gesture changes following the LMA mappings defined in PERFORM [10]. Facial animation involves mouth movements during speech, frequent blinks associated with low emotional stability, and blend shape updates to express emotions associated

[2]Bilkent University Ethical Committee for Human Research approved the study with the decision number 2023_11_05_01.

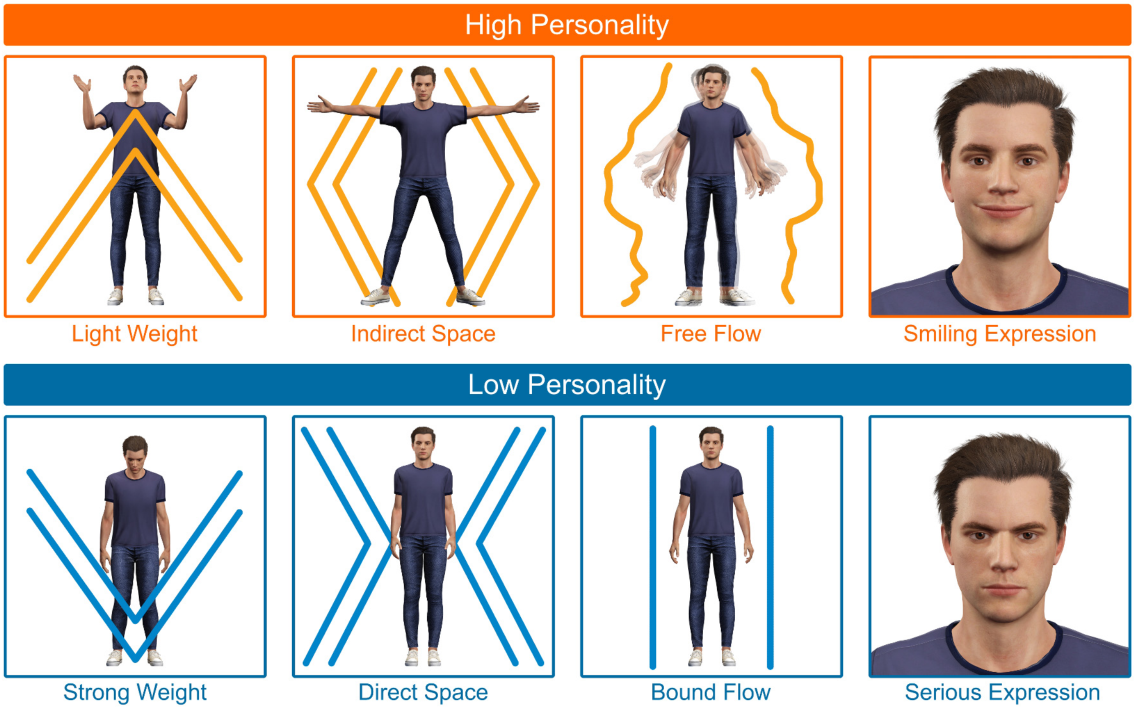


Figure 4: Summary of the LMA-based motion and facial expression adjustments applied to the agent in Model E.

with specific personality factors. The input personality determines the agent's default facial expression. For example, an agreeable agent tends to smile by default with each turn of its dialogue. We designed 3D human models for the current study using Reallusion Character Creator. To introduce a measure of diversity, we created four characters: two female and two male, each with light and dark skin tones, as depicted in Fig. 1.

Our updated system differs from the existing personality-driven platform regarding how it handles dialogue. The previous work used IBM Watson Assistant to extract the intent from user queries mapped into domain-specific handcrafted dialogue lines. This work replaces the dialogue logic with an LLM-based text generation model, GPT-3.5 Turbo, facilitated by OpenAI's Chat Completions Application Programming Interface (API), eliminating the need for manual dialogue crafting. When users type their prompts (e.g., ask a question), the system returns an answer coherent with the input personality description. We limited the token number to 750 for the output text to keep the conversation concise. Unlike the previous platform, which used Watson Text-to-Speech API for speech generation, our current system employs Microsoft® text-to-speech functionality, producing an almost immediate response to vocalize the agent's answer. This local solution also lets us determine the currently spoken word we use to display partial subtitles. Since the generated responses could be pretty long, we followed a dynamic approach where five words centering the currently spoken word were displayed on top of the agent in models with visual representation. The subtitles were on by default, but the users could disable them if they were distracting.

We used a temperature of 0.9 to promote diverse outputs from GPT while maintaining the information's reliability. Temperatures above 1 introduce creativity; however, they lead to hallucinations, conflicting with the aim of the information-based system. Chat Completions API takes as input a "messages" parameter consisting of message objects, where each object has a role of "system", "user", or "assistant" and content. For the role of "system", we give the following messages as input for different agent personalities and teaching topics:

- Act as an extraverted teacher teaching about <*topic*>, give friendly and polite answers.
- Act as an introverted teacher teaching about <*topic*>, give short and unfriendly answers.

The system sends the role prompt and the last five dialogue messages to produce a response that the agent speaks. Dialogue messages alternate between the user and the assistant. We limited the number of messages that form the agent's memory to five in order to manage costs, eliminate context drift, and prevent users from repeatedly asserting incorrect information until the LLM incorporates it.

### 3.2 Stimuli

We designed a $3 \times 2$ independent subjects study to compare three models—D, A, and E—each tested with high and low values of agreeableness-extraversion combination. Model D is the **d**ialogue-only setting, where the system's answers were shown on screen sentence-by-sentence concurrently with audio playback. Model A included an **a**nimated 3D model of the agent, randomly chosen among four alternatives. Model A involved the virtual human animated without any personality-based alterations. In Models D and A, personality was conveyed only through synthesized dialogue. Model E was similar to Model A but incorporated the **e**xpression of personality through face and body movements.

In Model E (see Fig. 4), motions that display high extraversion and agreeableness involve the LMA parameters of Indirect Space, Light Weight, and Free Flow. These correspond to multi-focal spatial attention, delicate, lifted-up movements, and uncontrolled and fluid motion. Because high extraversion and agreeableness are associated with opposite Time Efforts (Sudden vs. Sustained), we left the Time component of the animations unaltered. The animations expressing low extraversion and agreeableness involve Direct Space, Strong Weight, Bound Flow, and neutral Time, corresponding to single-focused, heavy, and controlled movements. The facial expression of a highly extraverted and agreeable agent is relaxed and happy, with occasional smiles and direct eye contact. In contrast, an agent characterized by lower levels of these traits displays a more tense facial expression, avoiding eye contact.

Fig. 2 displays screenshots of different models. We name each variation with its model name and whether they express high or low traits. For example, E-High refers to the variation where we express high trait personality using the model that uses both text and animation-based cues. We display a single image for Models A and D as they are visually similar in high and low variations. In Fig. 2, the E-Low variation has hands close to the body with a slightly more slanted posture, and the E-High variation has hands further from the body with a more upright posture.

### 3.3 Study Design

The study involved a conversation with a virtual agent to get information about a scientific subject. We presented participants with six options and asked them to select the least familiar topic. The topics were quantum computing, blockchain technologies, transformer architectures, quantum mechanics, string theory, and general relativity. The selected topic was provided to the GPT model as part of the system role prompt to guide a focused conversation. The application required that participants pose the agent at least five questions to learn about the topic, with no upper limit on the number of questions. The participants could interact with the system up to the one-hour time limit assigned by Prolific. Upon completing their queries, participants could proceed to answer survey questions. They could review the survey questions or the chat history at any point during the study. Before the study, the participants were informed that they would interact with a conversational agent system to rate its performance in personality expression, immersion, and learning.

The survey questions appeared in two groups. The first group included 27 questions on a 5-point Likert scale, where 15 questions measured the perceived personality of the agent using the extra-short form of the Big Five Inventory–2 (BFI-2-XS) [44]; we considered emotional stability as negative neuroticism for all factors to indicate traits with positive connotations on high values. 12 questions measured self-assessment of learning, quality, and engagement using the Learning Object Evaluation Scale for Students (LOES-S) [17]. In LOES-S, learning-related questions are about the self-assessment of learning and how much the learning object, i.e., the tool in question, helped teach the subjects a new concept. Quality assesses the instructional design, ease of use, organization, and help features. Engagement evaluates how much the subjects liked the tool and whether they found it motivating.

The second group of questions required open-ended input to receive detailed participant feedback. Completing both groups of questions directed the participants to the user study completion page, where they received a link for task approval. We avoided using specific pre- and post-test questions that could hinder the free-form dialogue between the participant and the agent.

### 3.4 Participants

We used the crowd-sourcing service Prolific to recruit participants. The study page informed the participants that they would interact with our conversational system to rate its success regarding personality perception, immersion, and learning. Before running the study, each participant was directed to a website to test whether they had installed the correct text-to-speech package. Only those with the supported system configurations could continue with the study. 210 unique participants (99 female, 95 male, 16 not specified) rated our system, with each alternative evaluated by 35 individuals, which provides a medium effect size (Cohen's $f = 0.26$) for both main effects and their interaction and power of 0.80 at a significance level of 0.05 for independent-subjects Analysis of Variance (ANOVA).

Each participant interacted with only one version of the system, where the average interaction time was $19.74 \pm 9.25$ minutes. This time excludes the introduction, where participants read about the task and download the application, but includes the time spent on answering survey questions. The average participant age was $28.80 \pm 8.57$. Upon entering the system, participants were shown an introduction message about the study details, where we also informed them about the data collected and the study's aim to measure the system's performance; we emphasized that the study did not aim to measure their knowledge in any manner.

## 4 Quantitative Analysis

### 4.1 Data Organization and Exploratory Analysis

BFI-2-XS includes three questions for each personality factor, some of which are inversely proportional to the measured dimensions. Responses were assigned integer values on a 5-point Likert scale, ranging from -2 to 2. We calculated the signed sum of these values to derive personality scores between -6 and 6, which were then re-scaled back to the range $[-2,2]$. Similarly, LOES-S has five questions measuring learning, four questions measuring quality, and three questions measuring the engagement of the learning object. We calculated the sum per measurement type and mapped the corresponding ranges into $[-2,2]$ to report the corresponding means.

For exploratory analysis, we display box plot diagrams of each model regarding perceived personality and LOES-S scores for learning, quality, and engagement (see Fig. 5). The diagrams indicate positive mean scores for all personality factors. The models received specifically high ratings for conscientiousness. The plots also show high positive ratings for learning, quality, and engagement, with mean engagement scores slightly higher for high personality variations than low personality ones. We can also observe that model E-High, followed by A-High, represents high conscientiousness, agreeableness, and emotional stability better than the other models. In the next section, we perform descriptive analysis to identify potential statistically significant effects of the models and personality styles on the output variables.

### 4.2 Variance Analysis

To investigate the impact of model type (D, A, and E) and personality style (high or low) on the measured qualities, we ran seven two-way analysis of variance (ANOVA) models and one non-parametric alternative model (Welch's ANOVA). Welch's ANOVA was utilized to assess the quality scores of LOES-S across model type and personality style, given that the assumption of equal variances was violated, as indicated by a Bartlett test.

Apart from the non-parametric model, which combined model type and personality style as a single factor, all other models examined the influence of model type and personality style on outcome mean individually and any potential interaction between these factors. With balanced and sufficiently large sample sizes ($n$=35) across factor combinations and evidence for equal variances across factor levels (as measured by Bartlett's test), all outcomes except for quality were appropriate for ANOVA modeling. To control the familywise error rate at 0.05, we employed the Hommel method to adjust for multiple testing across all model terms, including one post-hoc analysis. Unlike the conservative Bonferroni correction, the Hommel method offers increased statistical power. Tab. 1 presents significant terms for all ANOVA runs before and after the correction for multiple testing. Adjusted-for ANOVA tests revealed significant effects of personality style on engagement, openness, extraversion, agreeableness, and emotional stability. Although conscientiousness

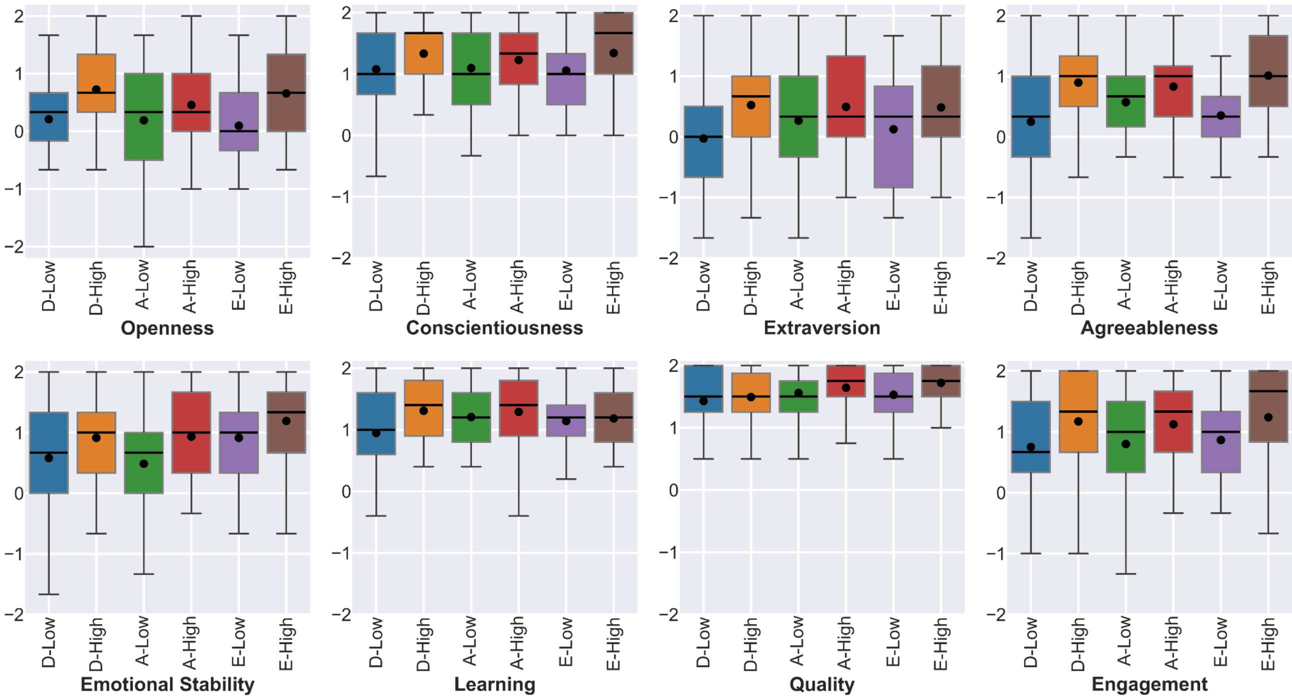


Figure 5: Box plots of each variation's BFI-2-XS and LOES-S scores.

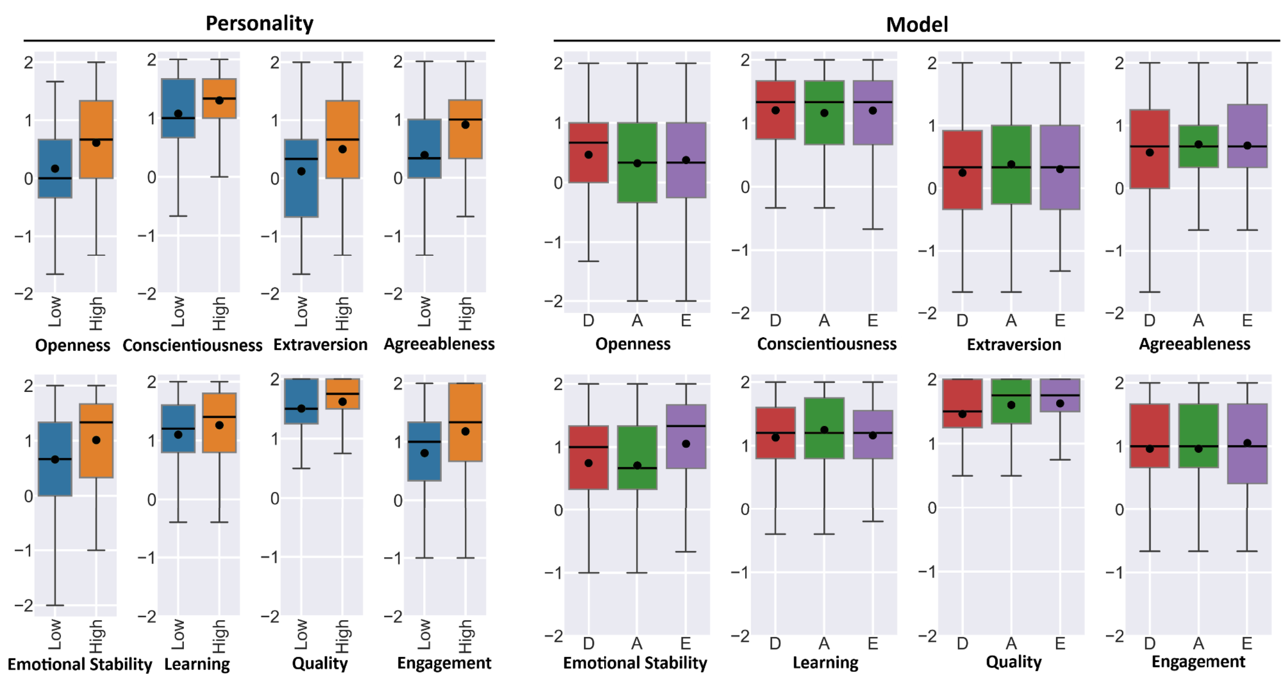


Figure 6: Box plots for BFI-2-XS and LOES-S scores assuming personality (Low-High) and model (D-A-E) styles.

initially carried a significant finding for personality style, this term's statistical significance dropped after multiple testing corrections. The main effect of the model type was initially significant for emotional stability, but the effect did not remain significant after the Hommel procedure. The box plots assuming personality and model groups are depicted in Fig. 6.

The effects of agent gender and skin color were not among the hypotheses. So, we randomly selected one 3D agent model among four different appearances, which also determined the agent's voice to support variety. We do not observe a significant effect due to the agent's gender or skin color, which confirms previous work [8].

Table 1: Two-way ANOVA significant findings on model type and personality style ($n = 210$). Statistically significant factors ($p < 0.05$) after p-value adjustment are emphasized in bold.

| Outcome | Factor | F | p-value | Adj. p-value |
|---|---|---|---|---|
| **Eng.** | Pers. Style | 9.502 | .002 | **.042** |
| **Opn.** | Pers. Style | 4.474 | $<.001$ | **.002** |
| Cons. | Pers. Style | 5.68 | .018 | .290 |
| **Ext.** | Pers. Style | 10.148 | .002 | **.031** |
| **Agr.** | Pers. Style | 25.541 | $<.001$ | **$<.001$** |
| **E.S.** | Pers. Style | 9.708 | 0.002 | **0.038** |
| E.S. | Model Type | 3.681 | .027 | .417 |

### 4.3 Correlations Analysis

We report the Pearson Correlation between LOES-S and personality factors in Tab. 2; correlation coefficients higher than 0.4 are considered moderate. Perceived openness, conscientiousness, extraversion, and agreeableness all positively correlate to the LOES-S scores, albeit some weakly. The highest correlations are for conscientiousness, particularly for Model D. Quality and engagement scores are strongly correlated ($> 0.6$), and learning is moderately correlated, nearing the threshold for a strong correlation. For Model D, openness and agreeableness also have moderate correlations with all the learning parameters. In general, engagement is moderately correlated with all factors except emotional stability. The expressed emotional stability is weakly proportional to each parameter. However, its only statistically significant correlations are for learning in Model D and quality in Model A. Overall, the correlations are the strongest for Model D and weakest for Model E.

Table 2: Pearson correlation ($r$) between perceived personality and **Le**arning, **Qu**ality, and **En**gagement. $^*$ indicates $p < .05$, $^{**}$ indicates $p < .001$. The cell colors transition from weak to strong correlation.

| Model | Cor. | O | C | E | A | ES |
|---|---|---|---|---|---|---|
| D | $r_{Le}$ | .422** | .582** | .396** | .446** | .349* |
| | $r_{Qu}$ | .437** | .654** | .279* | .325* | .190 |
| | $r_{En}$ | .417** | .608** | .287* | .422** | .109 |
| A | $r_{Le}$ | .351* | .348* | .456** | .398** | .208 |
| | $r_{Qu}$ | .215 | .378* | .204 | .150 | .425** |
| | $r_{En}$ | .504** | .332* | .505** | .541** | .233 |
| E | $r_{Le}$ | .266* | .405** | .281* | .184 | .046 |
| | $r_{Qu}$ | .334* | .133 | .056 | .305* | .054 |
| | $r_{En}$ | .401** | .353* | .389** | .353* | .052 |

### 4.4 Manipulation Check

We performed a manipulation check with a separate group of 57 users (27 female, 30 male, age average=$31.17 \pm 8.99$) recruited from Prolific. The results show that participants can distinguish the high variants from their low counterparts for all models. We displayed short segments that compare high and low variants of each model in random order without indicating which segment expresses the high or low variant. Each participant rated all three models and an alternative Model E where the dialogue is not personality-specific. The participants either chose one of the samples as the high trait variant or indicated that both segments have an equal chance to be the high variant. The ratios where participants preferred the intended sample as the high variant are D: 45%, A: 54%, and E: 75%, all above the chance threshold of 33.3%. The cases where the participant chose the opposite sample are rare, with D: 12%, A: 7%, and E: 8%. The difference from uniform random distribution is significant for all models with $p < .01$. Additionally, participants chose the correct high trait variant 73% of the time when personality-specific dialogue was omitted in Model E, and only 17% of the participants preferred the opposite sample.

### 4.5 Qualitative Analysis

#### 4.5.1 Theme Extraction

We analyzed user responses to open-ended questions regarding (1) their understanding of the conversation topic, (2) why the topic is important, (3) if they learned anything new, (4) if the system's behavior was influential on their learning experience, and (5) if they found the conversation interesting. Two independent researchers tagged user answers with at least 95% agreement for each theme. Tab. 3 shows the number of themes for each variation, determined based on the following criteria:

1 **Benefit / No Benefit to Learning Experience** reflects the effect of the system behavior on the participant's learning experience.
2 **Learning / No Learning** captures whether the participant learned anything new interacting with the system.
3 **Interesting / Not Interesting** considers whether the participant found the system interesting or not.

Table 3: Theme analysis results. The numbers depict the occurrence of each theme for each variation. The cell colors indicate transition from low to high over 35 participant answers per variation.

| Variation | D-Low | D-High | A-Low | A-High | E-Low | E-High |
|---|---|---|---|---|---|---|
| Benefit | 15 | 19 | 17 | 22 | 19 | 21 |
| No Benefit | 9 | 4 | 8 | 6 | 8 | 7 |
| Learning | 23 | 25 | 21 | 26 | 25 | 29 |
| No Learning | 6 | 6 | 5 | 2 | 7 | 2 |
| Interesting | 15 | 22 | 19 | 20 | 21 | 20 |
| Not Interesting | 3 | 2 | 0 | 1 | 2 | 1 |

#### 4.5.2 Theme Analysis

**Benefit.** Most participants indicated that the system behavior improved their experience; they emphasized increased engagement due to having a human-like agent: *"It was more engaging to have a human avatar instead of a blank screen or other representation." (P3:A-High)* Even for the dialogue-only variations, the experience was more enjoyable due to interaction: *"I think I enjoyed learning using the system more than I would have if I were reading on my own in a book or Google." (P21:D-High)* Being able to ask questions using natural language and receive to-the-point answers was found to be beneficial: *"It saved me some time from having to Google specific terms and read long texts on them, giving me the key points to get a basic understanding." (P17:E-High)*

A human-like interaction with a non-human system can help people with anxiety to experience interactive learning: *"It can be useful a lot, and I'd like to use it because it is quite calming down* *the person who has anxiety." (P27:E-High)* Participants who reported no benefit from the system often cited the difficulty of the answers: *"I could have learned more on Wikipedia or Google. The system's answers were too difficult to understand for me." (P169:D-Low)* Although the high-trait models were generally found to be more beneficial, some participants noted that their responses were too lengthy: *"The responses were quite long, sometimes too long. Some of the things it said could've been left out as they didn't provide any useful information; it was just 'flavor text.' I also read very fast, so waiting for it to stop talking was a bit boring. Other than that, it was a very positive experience. " (P42:D-High)*

In addition to length, a major difference between the high and low variations was the LLM's word choices. The high-trait variations use motivational language, supporting the user instead of just giving the answer: *"Yes, the system had a motivational tone that lifted my energy towards learning about something I had zero knowledge about." (P37:D-High)* Such language can help create a more interactive and thus motivational experience: *"Yes, I liked the easy-to-learn explanations and also the motivational part 'that is indeed a fantastic question'." (P106:D-High)* On the other hand, the low-trait variations responded with shorter sentences that some users preferred: *"It was not interesting per se; however, it was very informative and straight to the point." (P2:E-Low) "Conversation was simple and quick. The agent gave me short and simple answers to my question that are easy to understand" (P170:E-Low)*

Participants found variations with embodied agents slightly more beneficial and interesting, they reported a positive influence of having an embodied human agent for the high-trait variations: *"Somewhat, seeing a human-like face made it easier to memorize the information" (P163:A-Low) "Interacting with a humanoid entity is more engaging than reading a book." (P143:A-High) "Yes! I really enjoyed it; it seemed very human. It was like talking to an expert; it can answer any question you have instantly." (P12:A-High) "It was really exciting to see a person/character in front of the screen. Of course, it has an influence and really affected my learning process positively." (P134:E-High) "I think the animations and the looks of the character were motivating, and this could help with the learning in general." (P207:E-High) "Body movement and text to speech allowed to be more engaged in the conversation." (P184:E-High) "Having someone explaining a subject to you in human form generates a curiosity that is similar to listening to an enthusiastic teacher. As someone with a low attention span, the agent kept me engaged in the conversation and sparked further interest." (P1:E-High)* While most participants focused on the body movements, a few reflected on the facial expressions and their positive effect in the E-High variant: *"Slight facial "expressions" was noted and kind of felt like it made an impact, to be fair." (P11:E-High)*

For the low-trait agents, participants noted that agent movements were monotonic and the visual representation brought no advantage: *"The person itself is extremely dull, there is no life if that makes sense, and the movements and gestures are extremely weird, the hand and arm movements are strange, I would rather have that taken away, but being able to ask any questions to a topic and a response provided immediately is amazing, really like that aspect." (P196:E-Low) "Maybe, I think if the system were "nicer" and less monotonic, the learning would be easier." (P147:E-Low) "I did not find the "graphics" to help. A chatbot would have basically had the same effect on me." (P191:A-Low)*

**Learning.** The results suggest that high-trait personality variations lead to improved learning outcomes. However, the influence of expressivity appears limited, as Model E shows only a slight improvement in learning performance compared to Model A.

The system inspired some participants to learn more about the topic: *"I did not know anything about this theory at all. After my conversation, I can proudly say that I am really into string theory. I learned the basic concept and the creators of the theory. I also asked how I could learn more, and the conversational agent suggested four different possible sources." (P8:A-High) "It really almost felt like talking to someone who knows quantum computing well. I especially appreciated the way it understood my questions, even though I felt a question or two were a bit vague. The system actually made me want to know more about the subject so I can ask better questions." (P11:E-High)* Since we asked participants to choose the subject they had the least information about, most of them reported having almost no prior knowledge of the conversation topic. A few participants indicated they already had sufficient knowledge of the subject, and they did not learn anything new.

Participants categorized in the "no learning" theme generally indicated the difficulty of the subject: *"Previously, I had no idea what transformer architecture was. But unfortunately, I still believe that I did not learn a great deal about this type of technology. In my opinion, these new technologies that use AI are very difficult to understand if a person has no background knowledge." (P20:E-High)* One detriment to learning could be the agent's short answers in the low-trait variations: *"Not much (learning) as the replies were brief, but I got a basic idea." (P101:E-Low)* Conversely, the lengthy answers of the high-trait variations could be distracting: *"The topics answered were on point, maybe a bit too long, and different questions had similar answers in common as part of it." (P41:E-High)*

**Interestingness.** Participants found the D-Low variation to be the least interesting, followed by A-Low. All the high-trait variations and E-Low were perceived as similarly interesting. The expressive gesturing in the E-Low variation may have mitigated the decrease in interest. Some participants who found the system interesting also reported a positive influence on learning: *"I found it fascinating, really interesting, and quickly increased my knowledge on the subject. I would have loved to do more and carry on asking questions to discover more about blockchain." (P199:E-High)* The experience's novelty could have resulted in some participants finding the study interesting: *"It was quite interesting. I did not know what to expect when entering the task, but I was pleasantly surprised and engaged in the entire experience. It would definitely be something I would use again if I could." (P175:A-High)* Participants found Model D less interesting but still beneficial: *"I think it wasn't interesting, but it taught me a topic I didn't know about." (P2:D-Low)*

## 5 Discussion

All variations received positive mean personality ratings across all traits; regardless of modality and personality expression, the agents were perceived positively as open, conscientious, extraverted, agreeable, and emotionally stable. Participant responses to open-ended questions also support this finding. The whole experience was favorably perceived even when low-trait personality variants, which were supposed to be less friendly, were employed. Previous studies show that people find interactions with virtual agents engaging, informative, and usable [41]. However, the positive responses could also be due to the "novelty effect", an initial fascination with new technology. To mitigate such effects, future work can employ techniques like extended tutorials and adaptive strategies [28].

Although mean ratings were positive for both, high-trait agents received higher scores than low-trait agents for all personality factors. High-trait personality styles were associated with increased openness, extraversion, agreeableness, and emotional stability ratings with statistically significant effects. The only factor that did not have a statistically significant relationship with style variation after multiple hypothesis testing was conscientiousness. Thus, for RQ1, we can conclude that personality style affects the perception of all the personality factors except conscientiousness. Among these, agreeableness had the highest effect size, followed by extraversion. This finding also helps validate the personality style expression adjustments in the system and the mappings of the LMA factors of Space, Weight, and Flow to extraversion and agreeableness. Participants' answers to open-ended questions suggest they cared about the virtual agent's "friendliness" and "niceness" or lack thereof. These results align with the previous reports that students generally prefer teachers high in extraversion, agreeableness, and conscientiousness [47]. The variance in conscientiousness is challenging to discern in a short scenario [43]; so, the lack of statistically significant effects of style on its perception is expected. However, conscientiousness received the highest scores as the perceived agent personality, which may imply a tendency to attribute reliability and organizational skills to educational agents.

Regarding RQ2, we found no statistically significant effect of model type on perceived personality. Similarly, for RQ3, no statistically significant effects of model type on LOES-S scores were observed. Thus, H1 was rejected as the absence or presence of visual representations did not impact learning outcomes assessed via LOES-S scores. Some participants indicated an indifference toward the graphical representation in their comments. However, the qualitative analysis suggests participants found the models with visual representation more interesting and motivational, which improves the overall experience in line with previous work [6].

For comparisons between high and low-trait styles, the evidence partially supported H2. Although both personality styles were positively rated across all models, the only significant difference was in the mean engagement scores. High-trait agents were more engaging than low-trait agents, supporting H2c. Participant responses to open-ended questions also confirm this finding. The absence of statistically significant differences in quality and learning scores of LOES-S across high and low-trait variants can be attributed to individual learning preferences. Some participants praised the directness of the low-trait agents, while others emphasized that the high-trait versions were motivational.

For RQ4, we found positive correlations between perceived personality factors and most learning outcomes for all the models. The highest correlations between learning outcomes and perceived personality traits were for Model D, followed by Models A and E. Among all personality dimensions, conscientiousness yielded the highest correlation with learning outcomes. This behavior is particularly evident in Model D, which suggests that the lack of a visual representation may have allowed participants to focus more on the informative aspects of the system.

## 6 Limitations and Future Work

LOES-S measures are based on self-reports, which is suitable for analyzing the usability and likeability of the application– it does not directly reflect the participant's comprehension of the subject. Objective measurement of comprehension requires more information on the participant's memory and cognition skills [2] and non-generic, subject-related questions focusing on the conversation itself. We leave the learning performance evaluation, including pre- and post-assessment questions, for future work.

We also plan to incorporate the opportunities provided by current VR technologies to collect objective sensor-based data. For instance, attention can be measured by gaze tracking [11], or empathic responses can be assessed through facial landmark analysis [35]. Signals such as head direction, gaze, pupil size, facial expression, voice, and upper body pose can be collected and used to shape agent responses to maintain users' attention and provide personalized responses.

Another interesting direction would be to collect data on participant personalities to analyze how these traits influence their perception of the agents [25]. In our study, we did not collect personality data to avoid overwhelming the participants with too many questions. Future studies can include prolonged experiments to reveal interactions between self and observed personality and long-term effects.

Without an on-site user study, our conversational system relied on common hardware, so we avoided techniques requiring extra equipment or high-speed internet. One criticism regarding our system was the unnaturalness of the synthesized voice, potentially detrimental to perceived agreeableness [48]. Future studies can adopt state-of-the-art speech synthesis offered by data-driven services. Although no participant complained about typing their answers, a voice-based conversation input could feel more natural, and future studies could adopt automatic speech recognition to improve the user experience.

While embodied agents enable multi-modal communication and a more entertaining user experience than non-embodied ones [32], studies also report a decrease in user comprehension due to cognitive load [40]. Appropriate gesturing can mitigate such negative effects [9], improving the agent's perceived realism. High realism is essential to elicit positive effects on users, especially in VR [31]. Users generally prefer high behavioral realism in embodied agents [12], and recent human avatars can cross the uncanny valley to deliver high fidelity [42]. Future work involving more realistic agents can detect more nuanced behavioral changes. Our work was also limited in the diversity of the 3D models; different facial features [23, 19], or clothing [22] can influence personality perception. Evaluating these traits and other personality variations is left for future work.

Although the differences between high and low variations are statistically significant, they are not very pronounced in terms of mean differences. This could be due to the overall positive perception of the system; the low variants are perceived as neutral rather than low. This could also be due to participants interacting with only one system variation. Since the user study required a substantial amount of time, each participant only used one model's high or low version and did not have a reference point for comparison. This could have caused them to avoid making strong judgments.

## 7 Conclusion

We present a conversational system and user study to explore how personality perception and embodiment affect user experience and engagement in an informative setting. Using GPT-3.5 Turbo and realistic 3D human models, we created agents expressing high and low agreeableness and extraversion variations through dialogue and animation cues. We designed three types of agents: a disembodied agent expressing personality through dialogue, an embodied agent expressing personality only through dialogue, and an embodied agent expressing personality through dialogue and animation. We conducted a three-by-two independent-subjects user study with three agent models and the two personality variations, where each participant was asked to converse with an agent on a complex subject. After the conversation, participants rated their version of the system based on their perceived personality of the agent and the learning, quality, and engagement outcomes of the experience. The results indicate that the whole experience was rated favorably regardless of the model choice. Participants judged the agents as high in openness, conscientiousness, extraversion, agreeableness, and emotional stability. However, the degree of positive perception was lower in low-trait personality styles than in high-trait ones. Although the engagement score was higher for the embodied agent with expressive animations, we found no significant differences across the models for other learning outcomes. We hope that the findings of this work inspire future studies to utilize expressive animation and dialogue cues to improve the overall experience of conversational agents.


## Acknowledgments

This research is supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK) under Grant No. 122E123.

## REFERENCES


[1] J. Allbeck and N. Badler. Toward representing agent behaviors modified by personality and emotion. In *Embodied Conversational Agents at AAMAS*, 2002. 3

[2] T. P. Alloway and R. G. Alloway. Investigating the predictive roles of working memory and iq in academic attainment. *Journal of Experimental Child Psychology*, 106(1):20–29, 2010. 9

[3] T. Anvari and K. Park. 3d human body pose estimation in virtual reality: a survey. In *Proceedings of the 13th International Conference on Information and Communication Technology Convergence*, ICTC '22, pp. 624–628. IEEE, 2022. 3

[4] S. A. Aseeri and V. Interrante. The influence of avatar representation and behavior on communication in social immersive virtual environments. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*, VR '18, pp. 823–824. IEEE, 2018. 1

[5] G. Ball and J. Breese. Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost, and E. F. Churchill, eds., *Embodied Conversational Agents*, p. 189–219. MIT Press, Cambridge, MA, USA, 2000. 3

[6] A. L. Baylor. Promoting motivation with virtual agents and avatars: role of visual presence and appearance. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3559–3565, 2009. 9

[7] D. C. Berry, L. T. Butler, and F. De Rosis. Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, 63(3):304–327, 2005. 1

[8] J. C. Castro-Alonso, R. M. Wong, O. O. Adesope, and F. Paas. Effectiveness of multimedia pedagogical agents predicted by diverse theories: A meta-analysis. *Educational Psychology Review*, 33:989–1015, 2021. 7

[9] R. O. Davis. The impact of pedagogical agent gesturing in multimedia learning environments: A meta-analysis. *Educational Research Review*, 24:193–209, 2018. 1, 9

[10] 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(1), 2016. Article no. 6, 16 pages. 3

[11] A. Frischen, A. P. Bayliss, and S. P. Tipper. Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694, 2007. 9

[12] V. Groom, C. Nass, T. Chen, A. Nielsen, J. K. Scarborough, and E. Robles. Evaluating the effects of behavioral realism in embodied agents. *International Journal of Human-Computer Studies*, 67(10):842–849, 2009. 9

[13] 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, pp. 1–8, 2020. 3

[14] S. Hickson, N. Dufour, A. Sud, V. Kwatra, and I. Essa. Eyemotion: Classifying facial expressions in VR using eye-tracking cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision*, WACV '19, pp. 1626–1635. IEEE, 2019. 3

[15] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, vol. 36 of *NIPS '23*, 2024. Article no. 466, 9 pages. 3

[16] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, Article no. 102274, 9 pages, 2023. 3

[17] R. H. Kay and L. Knaack. Assessing learning, quality and engagement in learning objects: the Learning Object Evaluation Scale for Students (LOES-S). *Educational Technology Research and Development*, 57:147–168, 2009. 1, 5

[18] K. Kim, L. Boelling, S. Haesler, J. Bailenson, G. Bruder, and G. F. Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in AR. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality*, ISMAR '18. IEEE, 2018. doi: 10.1109/ismar.2018.00039 1

[19] R. S. S. Kramer and R. Ward. Internal facial features are signals of personality and health. *Quarterly Journal of Experimental Psychology*, 63(11):2273–2287, 2010. doi: 10.1080/17470211003770912 9

[20] N. Krishnaswamy, P. Narayana, R. Bangar, K. Rim, D. Patil, D. McNeely-White, J. Ruiz, B. Draper, R. Beveridge, and J. Pustejovsky. Diana's world: A situated multimodal interactive agent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(09):13618–13619, 2020. 3

[21] S. Laparle. Tracking discourse topics in co-speech gesture. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior: 12th International Conference, DHM 2021, Proceedings, Part I*, p. 233–249. Springer-Verlag, Berlin, 2021. doi: 10.1007/978-3-030-77817-0_18 3

[22] K. Legde and D. W. Cunningham. Evaluating the effect of clothing and environment on the perceived personality of virtual avatars. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, IVA '19, pp. 206–208. ACM, New York, NY, USA, 2019. 9

[23] A. C. Little and D. I. Perrett. Using composite images to assess accuracy in personality attribution to faces. *British Journal of Psychology*, 98(1):111–126, 2007. 9

[24] F. Mairesse and M. A. Walker. Can conversational agents express big five personality traits through language?: Evaluating a psychologically-informed language generator. Technical report, Cambridge University, Cambridge & Sheffield, UK, 2009. 3

[25] 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(6):1494, 1985. 9

[26] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In *Proceedings of the IEEE International Conference on Data Mining*, ICDM '20, pp. 1184–1189. IEEE, Piscataway, NJ, USA, 2020. 3

[27] F. Miao, I. V. Kozlenkova, H. Wang, T. Xie, and R. W. Palmatier. An emerging theory of avatar marketing. *Journal of Marketing*, 86(1):67–90, 2022. 1

[28] I. Miguel-Alonso, D. Checa, H. Guillen-Sanz, and A. Bustillo. Evaluation of the novelty effect in immersive virtual reality learning experiences. *Virtual Reality*, 28(1):27, 2024. 8

[29] J. Musek. A general factor of personality: Evidence for the big one in the five-factor model. *Journal of Research in Personality*, 41(6):1213–1233, 2007. 1

[30] M. Neff, Y. Wang, R. Abbott, and M. Walker. Evaluating the effect of gesture and language on personality perception in conversational agents. In *Proceedings of the 10th International Conference on Intelligent Virtual Agents, IVA '10*, vol. 6356 of *Lecture Notes in Computer Science*, pp. 222–235. Springer, Berlin, Heidelberg, Germany, 2010. 3

[31] M. Newman, B. Gatersleben, K. Wyles, and E. Ratcliffe. The use of virtual reality in environment experiences and the importance of realism. *Journal of Environmental Psychology*, 79:101733, 2022. 9

[32] N. Norouzi, K. Kim, G. Bruder, A. Erickson, Z. Choudhary, Y. Li, and G. Welch. A systematic literature review of embodied augmented reality agents in head-mounted display environments. In *Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments*, 2020. 9

[33] T. Numata, H. Sato, Y. Asa, T. Koike, K. Miyata, E. Nakagawa, M. Sumiya, and N. Sadato. Achieving affective human–virtual agent communication by enabling virtual agents to imitate positive expressions. *Scientific Reports*, 10(1):5977, 2020. 3

[34] S. Nyatsanga, T. Kucherenko, C. Ahuja, G. E. Henter, and M. Neff. A comprehensive review of data-driven co-speech gesture generation. *Computer Graphics Forum*, 42(2):569–596, 2023. 3

[35] S. Park, S. Park, and M. Whang. Empathic responses of behavioral-synchronization in human-agent interaction. *Computers, Materials & Continua*, 71(2), 2022. 9

[36] 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*, pp. 347–359. Springer, 2017. 3

[37] G. B. Petersen, A. Mottelson, and G. Makransky. Pedagogical agents in educational VR: An in the wild study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, New York, NY, USA, 2021. 1, 2, 3

[38] M. Rashik, M. Jasim, K. Kucher, A. Sarvghad, and N. Mahyar. Beyond text and speech in conversational agents: Mapping the design space of avatars. In *Proceedings of the ACM Designing Interactive Systems Conference*, pp. 1875–1894. ACM, New York, NY, USA, 2024. 1

[39] M. Saberi, S. DiPaola, and U. Bernardet. Expressing personality through non-verbal behaviour in real-time interaction. *Frontiers in Psychology*, 12, Article no. 660895, 19 pages, 2021. 3

[40] N. L. Schroeder and O. O. Adesope. A systematic review of pedagogical agents' persona, motivation, and cognitive load implications for learners. *Journal of Research on Technology in Education*, 46(3):229–251, 2014. 9

[41] R. M. Schuetzler, G. M. Grimes, and J. S. Giboney. An investigation of conversational agent relevance, presence, and engagement. In *Proceedings of the Twenty-fourth Americas Conference on Information Systems*, AMCIS '18. Association for Information Systems, Atlanta, GA, USA, 2018. 3, 8

[42] M. Seymour, L. I. Yuan, A. Dennis, K. Riemer, et al. Have we crossed the uncanny valley? understanding affinity, trustworthiness, and preference for realistic digital humans in immersive environments. *Journal of the Association for Information Systems*, 22(3):9, 2021. 9

[43] S. Sonlu, U. Güdükbay, and F. Durupinar. A conversational agent framework with multi-modal personality expression. *ACM Transactions on Graphics*, 40(1), 2021. Article no. 7, 16 pages. 1, 3, 9

[44] C. J. Soto and O. P. John. 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, 2017. 5

[45] P. M. Strojny, N. Dużmańska-Misiarczyk, N. Lipp, and A. Strojny. Moderators of social facilitation effect in virtual reality: Co-presence and realism of virtual agents. *Frontiers in Psychology*, 11:1252, 2020. 3

[46] W. Swartout, R. Artstein, E. Forbell, S. Foutz, H. C. Lane, B. Lange, J. F. Morie, A. S. Rizzo, and D. Traum. Virtual humans for learning. *AI Magazine*, 34(4):13–30, 2013. 1

[47] S. Tan, A. Mansi, and A. Furnham. Students' preferences for lecturers' personalities. *Journal of Further and Higher Education*, 42(3):429–438, 2018. 9

[48] S. Thomas, Y. Ferstl, R. McDonnell, and C. Ennis. Investigating how speech and animation realism influence the perceived personality of virtual characters and agents. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces*, VR '22, pp. 11–20. IEEE, Piscataway, NJ, USA, 2022. doi: 10.1109/VR51125.2022.00018 9

[49] O. Topsakal and E. Topsakal. Framework for a foreign language teaching software for children utilizing AR, Voicebots and ChatGPT (Large Language Models). *The Journal of Cognitive Systems*, 7(2):33–38, 2022. 3

[50] S. T. Völkel and L. Kaya. Examining user preference for agreeableness in chatbots. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21. ACM, New York, NY, USA, 2021. Article no. 38, 6 pages. 2

[51] D.-Y. Wu, J.-H. T. Lin, and N. D. Bowman. Watching VR advertising together: How 3D animated agents influence audience responses and enjoyment to VR advertising. *Computers in Human Behavior*, 133:107255, 2022. 1

[52] F.-C. Yang, K. Duque, and C. Mousas. The effects of depth of knowledge of a virtual agent. *IEEE Transactions on Visualization and Computer Graphics*, 30(11):7140–7151, 2024. 3

[53] R. Yang, T. F. Tan, W. Lu, A. J. Thirunavukarasu, D. S. W. Ting, and N. Liu. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263, 2023. 3

[54] E. Zell, K. Zibrek, and R. McDonnell. Perception of virtual characters. In *ACM Siggraph 2019 Courses*. ACM, New York, NY, USA, 2019. 3

[55] H. Zhang, C. Wu, J. Xie, Y. Lyu, J. Cai, and J. M. Carroll. Redefining qualitative analysis in the AI era: Utilizing ChatGPT for efficient thematic analysis. *arXiv preprint arXiv:2309.10771*, 2023. 3

[56] M. X. Zhou, G. Mark, J. Li, and H. Yang. Trusting virtual agents: The effect of personality. *ACM Transactions on Interactive Intelligent Systems*, 9(2-3):1–36, 2019. 3

[57] S. Zhu, W. Hu, W. Li, and Y. Dong. Virtual agents in immersive virtual reality environments: Impact of humanoid avatars and output modalities on shopping experience. *International Journal of Human–Computer Interaction*, 40(19):5771–5793, 2024. 3

[58] K. Zibrek, E. Kokkinara, and R. McDonnell. The effect of realistic appearance of virtual characters in immersive environments-does the character's personality play a role? *IEEE Transactions on Visualization and Computer Graphics*, 24(4):1681–1690, 2018. 3

[59] K. Zibrek, S. Martin, and R. McDonnell. Is photorealism important for perception of expressive virtual humans in virtual reality? *ACM Transactions on Applied Perception*, 16(3), 2019. Article no. 14, 19 pages. 1