It's All in the Personality: A Comparative Study of Real, Ideal, and Customized Virtual Instructors for AR Assembly Tasks

Abdul Mannan Mohammed, Martin McCarthy, Carsten Neumann, Gerd Bruder, Dirk Reiners, and Carolina Cruz-Neira

Abstract—While embodied conversational agents driven by Large Language Models (LLMs) are emerging as valuable tools for instruction in Augmented Reality (AR), a key challenge lies in crafting their personalities to optimize both instructional efficacy and user engagement. To address this, we present findings from a within-subjects experiment that compared task performance and user experience with a LEGO assembly task. Participants received guidance from a real human instructor and three virtual counterparts, whose Big Five personality profiles were designed to be: (1) a direct replica of the real human, (2) an "ideal" profile based on pedagogical research, or (3) customized by the participant. Our results reveal a critical trade-off: instruction from the real expert resulted in superior task efficiency and clarity; however, among the virtual conditions, instructors with idealized or user-customized personalities fostered significantly higher levels of user engagement and social presence compared to the virtual replica. Crucially, allowing users to customize their instructor's persona led to the strongest preference for future interaction. These findings underscore that personality is a fundamental component in the design of AI-driven instructors, providing empirical evidence for navigating the balance between task-oriented guidance and personalized, socially resonant user experiences.

Index Terms—Intelligent Virtual Agents, Instructor, Personality, Assembly Task, Augmented Reality, Artificial Intelligence.

1 Introduction

Instruction for assembly tasks is a critical challenge faced by both academia and industry, where guiding users through the process requires experienced and accessible faculty [14, 28, 54]. Collaborative efforts between artificial intelligence (AI) based agents and humans have been explored as a method for resolving these issues, where recent advancements in Large Language Models (LLMs) have given way to agents capable of solving novel and difficult tasks at a near human-level across diverse domains [3, 5, 63]. Augmented Reality (AR) based solutions that integrate the embodiment of these agents to engage and instruct users in a physical environment have further improved the instruction guidance process, establishing greater trust with users [35, 66]. While embodiment is a highly explored area for the enhancement of interaction with instructor agents, the role of instructor personality remains less understood. This is a timely area of inquiry, as recent work has shown that modern LLMs possess a remarkable ability to role-play and consistently emulate specific human personality traits [68], as well as understand human motives and emotions [3].

To address this gap in research, we explore the impacts of personality design for virtual agents in an assembly task scenario. Grounded in the widely-accepted Big Five personality model [9, 26], we designed and evaluated three distinct virtual instructor personalities:

- A personality matched to that of a real-world instructor.
- A personality based on literature recommendations for ideal instructor traits.
- A personality individually customized by each participant.

To evaluate these personalities, we conducted a user study assessing their subjective and objective performance when embodied in a virtual agent for an assembly task. We also included a condition with the real-world instructor to serve as a baseline for comparison. To guide our research, we formulate the following research questions:

- **RQ1**: Does instruction provided by a real human instructor contribute to an overall higher perceived efficacy in comparison to virtual instructors?
- **RQ2**: Does an enhanced personality based on the current body of literature outperform a virtual instructor matched to the personality of a real instructor?
- **RQ3**: Does allowing participants to customize their virtual instructor's personality lead to an improved perception of instructional efficacy?

In the remainder of this paper, we first provide an overview of background information in Section 2. We then describe the experiment we performed in Section 3. The results are presented in Section 4 and discussed in Section 5. We conclude the paper in Section 6.

2 Background

This section provides an overview of the research established on interactive virtual agents for instruction-based task scenarios with different instructor personalities.

2.1 Effects of Instructor Personalities

Personality traits are generally understood as consistent patterns in an individual's thoughts, emotions, and behaviors across time and situations. Among the various frameworks proposed to study these traits, the Big Five model has emerged as one of the most widely accepted and empirically supported frameworks [9]. This model stems from two major research traditions: psycholexical and questionnaire-based [12, 32]. In this context, the psycholexical tradition refers to the hypothesis that the most socially significant traits are encoded in natural language [6, 25, 26], and identifies core traits through criteria such as synonym frequency and cross-cultural universality. Further, the questionnaire-based tradition refers to the development of questionnaires or inventories containing items designed to tap into different aspects of personality, which then allows factor analysis to distill these descriptors into five broad dimensions based on the correlations between these items [62].

The Big Five model's robust framework has proven useful across numerous research domains, including educational and instructional contexts, where it helps explain individual differences in learning and behavior [46]. A substantial body of research has demonstrated that instructors' personality traits—particularly those captured by the Big

• *The authors are with the VARLab, University of Central Florida, Orlando, Florida, USA, 32816. E-mail: {abdulmannan.mohammed, martin.mccarthy, carsten.neumann, gerd.bruder, dirk.reiners, carolina}@ucf.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

Five framework—significantly influence instructional effectiveness and student outcomes.

Personality traits have consistently been shown to influence instructional effectiveness. In particular, *openness* is associated with innovative teaching practices, creative pedagogies, and adaptability to diverse learning needs [39, 41, 56], while *conscientiousness* promotes effective classroom management, careful planning, and reliable instruction through traits such as orderliness and self-discipline [34, 56]. *Extraversion*, defined by sociability and enthusiasm, contributes to active student engagement and dynamic classroom environments, and *agreeableness*—encompassing empathy, cooperation, and trust—enhances instructor-student rapport and emotional support [39]. Conversely, high *neuroticism*, which correlates with emotional instability and elevated stress, tends to impede teaching quality; in contrast, *emotional stability* is linked to reduced burnout and improved student engagement [37]. Together, these insights suggest that instructor profiles emphasizing high *openness*, *conscientiousness*, *extraversion*, and *agreeableness*, while minimizing *neuroticism*, foster more effective, adaptive, and student-centered learning environments, forming the theoretical and empirical foundation for the personality configurations employed in our study. Beyond education, research on user innovation shows that people often prefer and value what they design themselves, even when it closely resembles premade options. This "I designed it myself" effect [22, 23] highlights how co-creation fosters attachment and preference—an idea directly relevant to our user-customized instructor condition (VMY).

The emergence of LLMs capable of consistently and controllably emulating diverse Big Five personality profiles marks a significant technological advance [7, 68], creating a novel opportunity to move beyond theory and empirically investigate how a virtual instructor's personality design directly impacts the user experience in learning scenarios.

2.2 Real and Virtual Task Instructors

AI systems have become progressively implemented as a method of enhancing task guidance and decision-making processes, where research is ongoing on how trust between users and agents impacts the effectiveness of task performance [64]. Functionally, modern LLMs have proven particularly adept in this context; studies show they can generate more detailed and coherent feedback than human instructors and can identify and rephrase incorrect trainee responses with a proficiency rivaling that of human experts [10, 43]. The findings show that a lack of trust between an intelligent agent and the user calls for a "human-in-the-loop" system in which AI provides suggestions rather than commands [4, 21]. To further research towards better AI trust, Vodrahalli et al. [65] provide a comparative analysis of how humans use advice depending on the source, agent or person, using the judge-advisor paradigm. Their findings suggest that the belief of a person in AI has a significant impact on trusting the agent's advice, indicating that perception plays a key role in how agents can influence users in task-based scenarios. This is a common theme in the literature, where contextual factors often play a key role in shaping trust between agent and user. This is indicated by Fahnenstich et al. [19] who observed individuals in high-risk situations with AI decision-support agents in comparison to a human counterpart, finding that risk levels provide a significant influence in trust dynamics in the way humans interact with AI-based instructors. While the currently explored interactions with AI agents in task guidance scenarios is a heavily addressed topic, there is little exploration on how user interpretation of an instructor's personality may enhance their trust. Addressing these gaps could enhance the design of AI systems for fostering trust across a variety of users for task guidance and decision-making scenarios.

The recent integration of LLMs with digitally generated 3D characters has allowed virtual instructors to integrate into task-based scenarios [44, 55]. These agents leverage conversational capabilities with human-form embodiment, especially when combined with AR-based systems, and allow for increased realism and enhanced user experiences [24, 66]. Psychologically, this increased realism stems not just from visual appearance but from the advanced emotional emulation capabilities of modern LLMs; for instance, a state-of-the-art model achieved an emotional quotient (EQ) score surpassing 89% of human participants and has even exhibited predictable, humanlike patterns of cognitive consistency that suggest a functional analog of selfhood [42, 67]. The incorporation of embodied instructors in AR is a commonly explored research topic [40, 51, 59]. More recently, systems have further explored embodied and personalized AI instructors, examining how adaptive personas, multimodal interaction, and real-time system design influence user experience, social presence, and engagement in task-based scenarios [48–50]. These works build upon a broader foundation of empirical studies examining how embodiment and visual realism shape user perception and trust in virtual instructors. For instance, Kim et al. [35, 36] performed human-subject studies that evaluated the need for a visual embodiment of virtual agents within AR environments. Their findings suggest a greater amount of confidence in a virtual agent's ability to comprehend the real world when it is visually embodied in comparison to voice-only agents, specifically in a simple task scenario. They also identify a need to avoid overly exaggerated elements of the design process to avoid uncanniness, a phenomenon where something appears unnatural or uneasy [61]. In addition to this, users found embodied agents within AR experiences to be more reliable and functional. However, they indicate that they do not factor multiple demeanors or personalities into the impact of how likely they impact user opinion on trust in the agent. This is supported by Reinhardt et al. [58], who demonstrate that AR-based systems are preferred for task-based scenarios due to the ability to establish a social connection via cues such as eye contact. This body of literature suggests a need for high-fidelity humanoid representations for communication-based tasks. Likewise, they suggest that improvement upon the current state of research includes analyzing how embodied agents are improved through personality factors.

3 Experiment

In this section, we describe the experiment in which we evaluated and compared the effects of real and virtual instructors (see Figure 1). Our experimental protocol was approved by the University of Central Florida Institutional Review Board (IRB) under Protocol No. SBE-15-11405.

3.1 Participants

Following a power analysis with G*Power on the basis of anticipated strong effects [20], we recruited 26 participants from our university community, 9 female, 16 male and 1 non-binary, ages between 19 and 45, $M = 25.5$, $SD = 6.0$. The participants were either students or non-student members of our university community who responded to open calls for participation. Participants had no experience with the human instructor (described in detail in Section 3.3.1) in prior instructional situations. Out of our participant sample, eleven wore glasses and two wore contact lenses during the experiment. We also asked our participants to indicate their experience with AI technologies. All of our participants indicated that they had some experience, neither being completely inexperienced nor experts. Participants received monetary compensation for their participation in the experiment in the form of a USD $15 Amazon gift card.

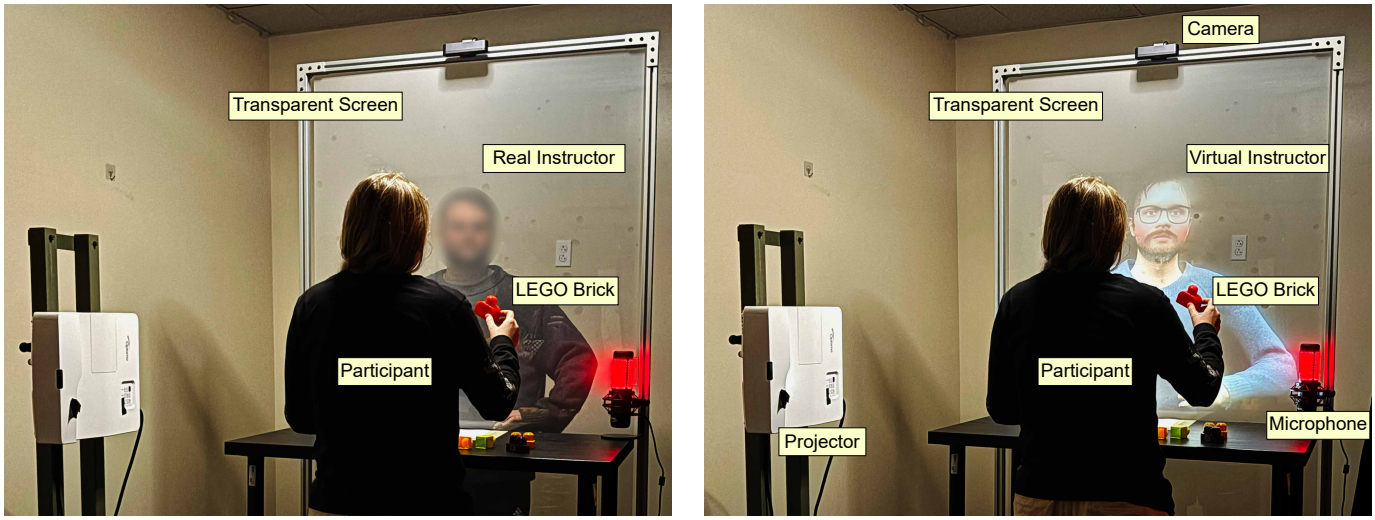
3.2 Material

In this section, we describe the experiment setup.

3.2.1 Apparatus

As shown in Figure 1, the experimental setup included a table (0.99 m wide, 0.63 m deep), with the instructor and participant positioned on opposite sides of a transparent screen (1.04 m wide, 2.05 m high). The screen combined a rigid Plexiglass base with a nano-optic overlay[1], enabling projection while maintaining transparency. While our study does not evaluate the impact of the display system on the experience,

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



(a) (b)

Fig. 1: Annotated photos showing a participant performing the assembly task in the experiment with (a) the real instructor and (b) a virtual instructor. The setup was divided into the portion of the table in front of the divider, the transparent screen, and the real or virtual instructor on the other side of the table behind the screen. We used the transparent screen to closely control and match the appearance of the real and virtual instructors in the experiment. We intentionally blurred the face of the real instructor in the image to protect their identity.

there is an active body of research to suggest holographic-based AR displays over traditional methods for interaction [1, 29].

An Optoma EH340UST projector (1920×1080 resolution, 22,000:1 contrast ratio, 4000 ANSI lumens) was placed 0.35 m in front and 0.36 m to the left of the screen, outside participants' direct view. Projection calibration was conducted in Unreal Engine 5.5 (UE 5.5) [18] using in-house tools. Ambient lighting was balanced using two low-brightness selfie ring lights positioned on either side of the screen.

To replicate the real instructor's appearance, we captured a 3D scan using Polycam [57], refined it in Blender [2], and generated a MetaHuman agent [17], imported into UE 5.5 with a minimal animation set (idle, listening, thinking, speaking) to standardize nonverbal behavior. Rendering was performed on a Windows 11 Pro machine (Intel Core i7-8750, 32 GB RAM, NVIDIA RTX 4070 GPU). All virtual instructors shared this appearance, differing only in their configured personality traits (see Section 3.3). The real instructor's voice was cloned using ElevenLabs' TTS API [16], enabling consistent voice output across virtual conditions. Participants interacted with the system using a HyperX QuadCast S microphone (input) and Logitech Multimedia Z150 speakers (output).

A Depstech DW50 Webcam 4K (3840×2160 at 30 fps, 90° field of view) mounted above the screen captured the participant workspace to track LEGO assembly progress. We used four LEGO figures for this experiment, which were designed to be of comparable difficulty and consisted of 12 LEGO pieces (see Figure 2). To ensure accessibility for participants with color-blindness, the LEGO pieces were labeled with unique letter identifiers.

3.2.2 System Architecture Overview

We developed a modular system based on the literature [11, 15] and refined through iterative prototyping to enable AI-driven virtual instruction for assessing user engagement and task efficacy. Crucially for this study, research has shown that that OpenAI's GPT-4o can reliably be steered to emulate diverse personality traits [68]. Furthermore, GPT-4o [52] in particular exhibits predictable, humanlike patterns of psychological consistency, making it a robust choice for studies involving agent personalization [3, 42]. The final architecture comprises five primary layers:

User Interaction Layer Participants interacted naturally via voice (microphone) and visual input (overhead RGB camera), enabling blended verbal and visual communication for task guidance and verification.

Input Processing Layer Real-time speech transcription was handled using Azure's Speech-to-Text API [47], with keyword detection (e.g., "verify", "check") triggering conditional image capture via OpenCV [53]. This avoided unnecessary visual processing unless explicitly required, a trigger based on the principle of conversational grounding where users seek to establish mutual understanding [8].

Prompt Management Layer A custom *AIManager* module managed dialogue history and dynamically constructed GPT-4o prompts. Prompts integrated the instructor's Big Five personality profile, task instructions, user history, and, when applicable, base64-encoded images, ensuring personalized and context-aware responses. The prompt for each instructor included instruction for analysis of user emotion via the produced STT output. When keyword activation occurred, the image input of current task status would then be additionally attached to the STT, this allowed for easy change of the AIManager module's state of user progress on the task. Through these methods, each prompt uniquely produced output based on the personality of the agent, either more direct or more encouraging.

Response Generation and Image Analysis GPT-4o generated real-time, personality-driven responses, which were synthesized using ElevenLabs' TTS API [16] for natural voice output. GPT-4o's image analysis capabilities enabled visual assessment of user progress for context-relevant feedback.

Virtual Instructor Synchronization The MetaHuman avatar was rendered in Unreal Engine's MetaHuman framework [**?**], with its animation states managed by a simple state machine comprising four primary states: **Idle**, **Listening**, **Thinking**, and **Speaking**. This system triggered animations in response to user speech and synchronized talking animations with the system's audio output. Transitions were driven by the Input Processing Layer; voice activity detection moved the agent to **Listening**, while speech completion triggered the **Thinking** state during GPT-4o processing. The instructor returned to an idle state after interactions, unless an API timeout or generation failure triggered a fail-safe **Error State**, which defaulted the system back to **Idle** to prevent non-responsive animation loops. The final avatar was projected onto the transparent display to maintain visual consistency.
We measured the end-to-end system latency, which included all processing from speech transcription and image analysis to response synthesis.

The system showed a minimal mean delay of M = 1.58s (SD = 0.32s) between participant actions and the virtual instructor's reaction. For interactions without visual input, this latency was even lower, averaging M = 0.91s (SD = 0.27s). This system latency naturally aligned with the time required by the real instructor for comparable actions. For instance, when a participant requested a check of their build, the real instructor's process of visually inspecting the work and formulating a response resulted in a natural delay. This human response time was consistent with the virtual instructor's technical latency for analyzing the camera feed and generating a reply, ensuring the timing of interactions was comparable across conditions.

3.2.3 Assembly Task

Participants were instructed to assemble one of four LEGO figures that were at comparable difficulty, as described in Section 3.2.1 and shown in Figure 2. As discussed in Section 3.2.2, progress in the LEGO assembly task was automatically tracked using the top-mounted camera. The camera feed was streamed directly into GPT-4o. Responses from the virtual instructors were generated by GPT-4o through processing both the transcribed speech from the participant and the visual context provided by the live camera feed.

3.2.4 Instructor Personalities

As described in Section 3.2.1, all virtual instructors' *appearances* and *voices* were matched to that of the real instructor. However, their *personalities* were configured differently across conditions. This influenced primarily what the instructors verbally communicated to the participants, and at which times during the assembly tasks. We assessed personality using the IPIP-NEO-120 instrument [33], administered via an open-access online platform[2]. This 120-item inventory measures the Five Factor Model—*Openness*, *Conscientiousness*, *Extraversion*, *Agreeableness*, and *Neuroticism*—with each of the 30 facets evaluated by four Likert-scale items, yielding domain scores ranging from 0 to 120. To support meaningful interpretation of these scores, we operationalized trait ranges to reflect increasing intensities of each personality dimension: *Very Low* (24–47), *Low* (48–71), *Moderate* (72–95), *High* (96–108), and *Very High* (109–120). These descriptive ranges enabled mapping participant preferences to behavioral characteristics grounded in validated trait descriptions [33].

To ensure the virtual instructor reliably embodied these profiles, we implemented several principled prompt engineering strategies using GPT-4o. Specifically, our system prompt first defined the target personality using both its numeric scores across the five dimensions and the descriptive text summaries generated by the test website, an approach informed by the PERSONALITY PROMPTING (P²) method [31]. This was followed by Reverse Role Prompting, where the model was instructed on behaviors to avoid, helping it maintain the assigned traits more effectively [7]. We also controlled the model's temperature setting to 0 to regulate creative variance and promote behavioral consistency, a standard technique for personality evaluation in LLMs [30, 68].

3.3 Methods

3.3.1 Experiment Conditions

In order to address our research questions (see Section 1), we implemented a within-subjects experimental design with four conditions:

- **Real Instructor (R):** Interaction with a real human instructor.
- **Personality-Matched Virtual Instructor (VA):** The virtual instructor is configured to match the real instructor's personality, based on their Big Five personality matrix.
- **Literature-Based Virtual Instructor (VB):** The virtual instructor's personality traits are "idealized"—informed by empirical research on instructor performance in the literature.
- **User-Customized Virtual Instructor (VMY):** Participants customize their optimal virtual instructor by specifying their preferred levels of the Big Five personality dimensions.

[2] https://bigfive-test.com

Participants interacted with each instructor for approx. 5 min while completing different basic LEGO assembly task of matched difficulty. All tasks and conditions were randomized in the experiment.

Real Instructor (R) The real instructor (male; 23 years; Caucasian; 1.78 m tall) was a teaching assistant in an undergraduate engineering lab course, where he regularly guided students through hands-on assembly and troubleshooting tasks. This role provided him with practical experience in delivering clear, task-based instructions in real-time settings, making him well-suited for this study. R's personality was assessed using the IPIP-NEO-120 instrument [33], revealing *Low* Neuroticism (69/120), *Moderate* Conscientiousness (73/120) and Extraversion (81/120), and *High* Openness (96/120) and Agreeableness (103/120). These assessment results—including the numeric scores were combined with an analysis of his natural instructional style and delivery patterns to inform the design of the personality-matched virtual instructor (VA).

Personality-Matched Virtual Instructor (VA) The virtual instructor was a high-fidelity replica of the real instructor (R). Its system prompt for GPT-4o was constructed by embedding the complete results from R's IPIP-NEO-120 assessment—including numeric scores and validated descriptive summaries [33]—together with an analysis of his frequent phrases and characteristic wording patterns.

To ensure behavioral fidelity, we used a consistent procedure across all virtual instructor conditions: Johnson's IPIP-NEO categories provided the numeric ranges and trait labels, while Mairesse and Walker's method [45] informed concise linguistic exemplars that anchored each trait in naturalistic phrasing. For example, R's Extraversion score of 81 fell within Johnson's "Moderate" range (72–95) and was paired with descriptors such as "sociable but not dominant," whereas his Neuroticism score of 69 placed him in the "Low" range (48–71) and was anchored with descriptors like "calm" and "emotionally stable.

We validated this persona by re-administering the IPIP-NEO-120 instrument to the VA agent; across five separate sessions, the agent's scores consistently fell within ±5 points of the real instructor's profile on each of the Big Five domains. This demonstrated that the personality-matched virtual instructor was both stable across runs and faithfully captured R's original assessment.

Literature-Based Virtual Instructor (VB) This instructor embodied an empirically-supported "ideal" personality, derived from Kim and MacCann's survey of 378 students on their preferred instructor traits [38]. The reported means on their original 8–72 scale were: Openness ($M = 56.6$), Conscientiousness ($M = 61.2$), Extraversion ($M = 53.5$), Agreeableness ($M = 60.8$), and Emotional Stability ($M = 52.8$). To align these values with our instrumentation, we applied a linear transformation to the 24–120 IPIP-NEO scale, a standard approach for mapping between validated instruments:

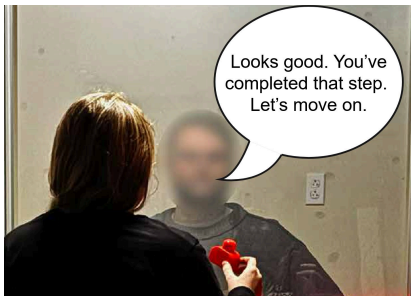
$$\text{New Score} = 24 + \frac{(\text{Original Mean} - 8)}{72 - 8} \times (120 - 24).$$

This yielded the following target profile: *High Openness* (97), *High Conscientiousness* (104), *Moderate Extraversion* (92), *High Agreeableness* (103), and *Low Neuroticism* (53). As in VA, these numeric values were embedded in GPT-4o's system prompt together with their validated labels and behavioral descriptors. To construct these descriptors, we followed the same procedure as in the other conditions, drawing on Johnson's IPIP-NEO categories [33] for range labels and on Mairesse and Walker's method [45] to anchor each trait with concise linguistic exemplars (e.g., "warm, sociable" for high Extraversion).

We validated this persona by administering the IPIP-NEO-120 instrument to the VB agent across five separate sessions; the agent's scores consistently fell within ±5 points of the transformed Kim and MacCann profile on each domain. This demonstrated that the "ideal instructor" persona was both stable across runs and faithful to the literature-derived target profile.

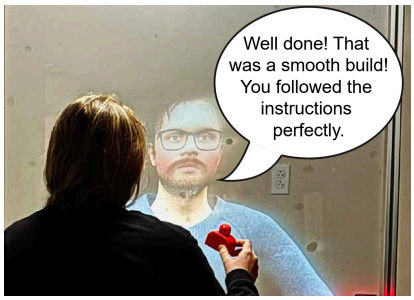
User-Customized Virtual Instructor (VMY) Participants configured their "ideal" instructor using five 5-point sliders (levels 0–4), each directly mapped to the validated IPIP-NEO 24–120 scale [33]:

(a) Assembly Task

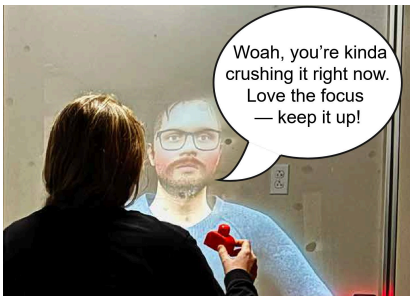


(b) Condition R

(c) Condition VA



(d) Condition VB



(e) Condition VMY

Fig. 2: (Left) Material used in the experiment: four LEGO figures with comparable difficulties that participants assembled on the table in the experiment. The tasks were randomized for each participant between the conditions in the experiment. (Right) Example responses from all four instructor conditions when prompted to check the current progress of assembly. Left to right: R, VA, VB, VMY. We blurred the real instructor in condition R to protect their privacy. Due to their same personality, R and VA both verified participants' progress in a direct manner and using brief, neutral feedback. In contrast, VB and VMY responded in a more engaging tone, offering praise.

0 = 24–47 (*Very Low*), 1 = 48–71 (*Low*), 2 = 72–95 (*Moderate*), 3 = 96–108 (*High*), 4 = 109–120 (*Very High*). This design did not introduce a new measurement scheme but exposed the existing scale in a participant-friendly interface. To aid interpretation, slider endpoints were accompanied by concise behavioral anchors (e.g., "quiet, reserved" vs. "warm, sociable" for Extraversion), drawing on personality–language generation work [45], while Johnson's categories defined the labels and cut-points.

When a slider was selected, the system generated the corresponding interval, its validated label, and a short descriptor, ensuring that—as in VA and VB—profiles were internally represented by both numeric scores and interpretable descriptors. For analysis, each interval was reduced to its midpoint (e.g., 96–108 $\rightarrow$ 102), enabling aggregation across participants. This parallels standard practice in personality research, where ordinal responses are converted to representative numeric values and averaged. For instance, Kim and MacCann [38] analyzed students' "ideal instructor" traits by averaging Likert-scale ratings into mean domain scores. Our approach follows the same principle, with the distinction that the slider categories were defined by validated IPIP-NEO intervals rather than arbitrary scale points.

On average, participants selected profiles characterized by *High Openness* ($M = 97.85$, $SD = 5.2$), *Very High Conscientiousness* ($M = 104.31$, $SD = 4.6$), *High Extraversion* ($M = 95.08$, $SD = 6.1$), *High Agreeableness* ($M = 92.31$, $SD = 5.7$), and *Low Neuroticism* ($M = 44.31$, $SD = 7.3$). Like VA and VB, the numeric values were embedded to GPT-4o's system prompt together with their validated labels and behavioral descriptors.

Across VA, VB, and VMY, the system prompt structure was identical to that described earlier; only the target personality profiles differed. To preserve fidelity, we also applied Reverse Role Prompting [7], where each instructor was instructed not to use behaviors inconsistent with its traits—for example, a "Low Neuroticism" agent was told not to respond with worry or defensiveness, and a "High Agreeableness" agent was told not to use dismissive or confrontational phrasing.

3.3.2 Procedure

Participants were welcomed into the study environment and provided with an overview of the experiment. After providing informed consent, each participant completed four assembly tasks, one with each instructor condition (R, VA, VB, VMY). Both the order of instructor conditions and the specific LEGO build used in each session were randomized using a wheel-based assignment tool. Because the four LEGO builds were distinct but comparable in difficulty (Figure 2), this randomization was sufficient to balance exposure, making counterbalancing unnecessary and thus minimizing potential learning effects across trials.

In each trial, participants received real-time instructions from the assigned instructor while completing the corresponding LEGO build. After completing all four sessions, participants filled out post-study questionnaires assessing their experiences, perceptions of each instructor, and overall task engagement. The full study procedure took approximately one hour per participant.

3.3.3 Measures

Task Performance We measured performance through a combination of automated logging and live observation. Python scripts recorded time-stamped events throughout each session to calculate *task duration*, reflecting the total time taken by participants to complete an assembly task. Additionally, we recorded *instructor errors* (incorrect or misleading instructions) and *instructor corrections* (instances where the instructor clarified or corrected their own instructions) based on live observations during the study.

Harms-Biocca Social Presence questionnaire To assess participants' perception of social presence across instructor conditions, we employed the 36-item version of the Harms-Biocca Social Presence questionnaire [27]. This widely-used and validated instrument captures six core dimensions of social presence as well as an overall score. Participants rated each item on a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Responses were averaged within each dimension to generate subscale scores, providing a multidimensional profile of perceived interpersonal connection and

Table 1: Single-Item Questionnaires we used in the experiment.

Question	Scale
How much do you TRUST the instructions provided?	1 (Not Trustworthy at All) .. 7 (Very Trustworthy)
To what extent did the task instructions contribute to your FRUSTRATION?	1 (Extremely Frustrating) .. 7 (Not Frustrating at All)
How PERSONABLE did the instructions feel?	1 (Very Impersonal) .. 7 (Very Personal)
How ENGAGED were you while following the instructions?	1 (Not Engaged at All) .. 7 (Fully Engaged)
How FREQUENTLY do you think you would use this system?	1 (Not Frequently at All) .. 7 (Very Frequently)

interaction quality with each instructor.

Single-Item Questionnaires To capture participants' subjective impressions of each instructor, each participant responded to five custom single-item questions. The questions and 7-point rating scales are shown in Table 1. These items were included to capture participant attitudes toward the instructional experience beyond formal social presence dimensions.

User Experience Questionnaire (UEQ) To evaluate the overall user experience with each instructor, we employed the short form of the UEQ [60], which is designed to capture participants' subjective impressions across a range of experience-related qualities. Participants responded to nine bipolar adjective pairs, each presented as a 7-point semantic differential scale. These nine items are then combined to form the three main dimensions: *Attractiveness*, *Pragmatic Quality*, and *Hedonic Quality*. The UEQ dimensions indicate how each instructor was perceived in terms of their usability.

Instructor Rankings After completing all conditions, we further asked the participants to rank the four instructors in terms of their preferences from their 1st choice to their 4th choice.

3.3.4 Hypotheses

Based on the literature in this field as well as our research questions **RQ1** to **RQ3** (see Section 1), we formulated the following three hypotheses for our experiment, where we define efficacy to be the ability to produce a desired or intended result with respect to our experiment measures:

- **H1** Higher efficacy of a real instructor compared to virtual instructors (R > VA, VB, VMY).
- **H2** Higher efficacy of a literature-based virtual instructor compared to a personality-matched virtual instructor (VB > VA).
- **H3** Higher efficacy of a user-customized virtual instructor compared to personality-matched or literature-based virtual instructors (VMY > VA, VB).

Specifically, the first hypothesis is based on our experience and prior results from the literature suggesting some persistent differences between real and virtual instructors due to limitations of the underlying AI models, rendering, and displays, just to name a few. The second hypothesis directly stems from the literature, as discussed in Section 2, which identified idealized personality traits among instructors that are unlikely to reflect the reality for any particular real instructor. The third hypothesis is based on support from the literature that individuals often feel more comfortable with and invested in systems that they had a hand in shaping, which likely extends to user-customized instructors.

4 Results

In this section we present the results of our statistical analysis. We analyzed the data with non-parametric Friedman tests and pairwise Wilcoxon Signed Rank tests with Bonferroni correction for the post-hoc comparisons.

4.1 Objective Data

The objective results for the assembly tasks are shown in Figure 3.

Task Duration We found a significant main effect for *Task Duration*, $\chi^2 = 20.68$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the task durations for R were significantly lower than those for VA, VB, and VMY (all $\boldsymbol{p < 0.05}$), which supports Hypothesis **H1**. In other words, our results show that the tasks were completed significantly faster with the real instructor compared to all virtual instructors.

Instructor Task Errors We found no significant main effect for *Instructor Errors*, $\chi^2 = 5.00$, $p = 0.172$. It is worth noting that errors occurred only with virtual instructors but not with the real instructor.

Instructor Corrections We found no significant main effect for *Instructor Corrections*, $\chi^2 = 3.40$, $p = 0.335$.

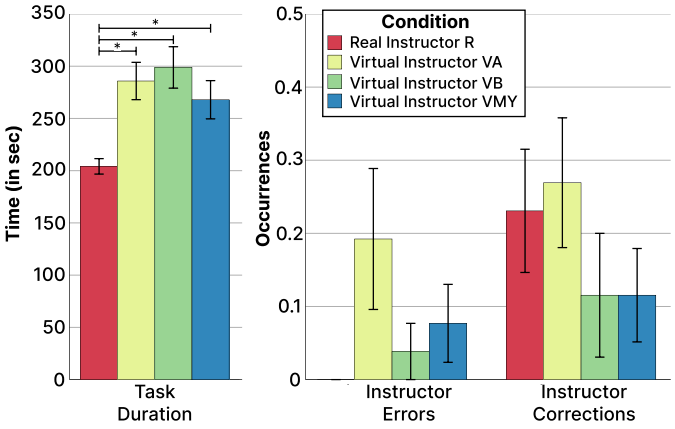


Fig. 3: Objective results indicating the task performance and occurrences of errors and corrections with the instructors. Lower is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level.

4.2 Subjective Data

We present the results for the rating scales and rankings in the different questionnaires for the four instructor conditions.

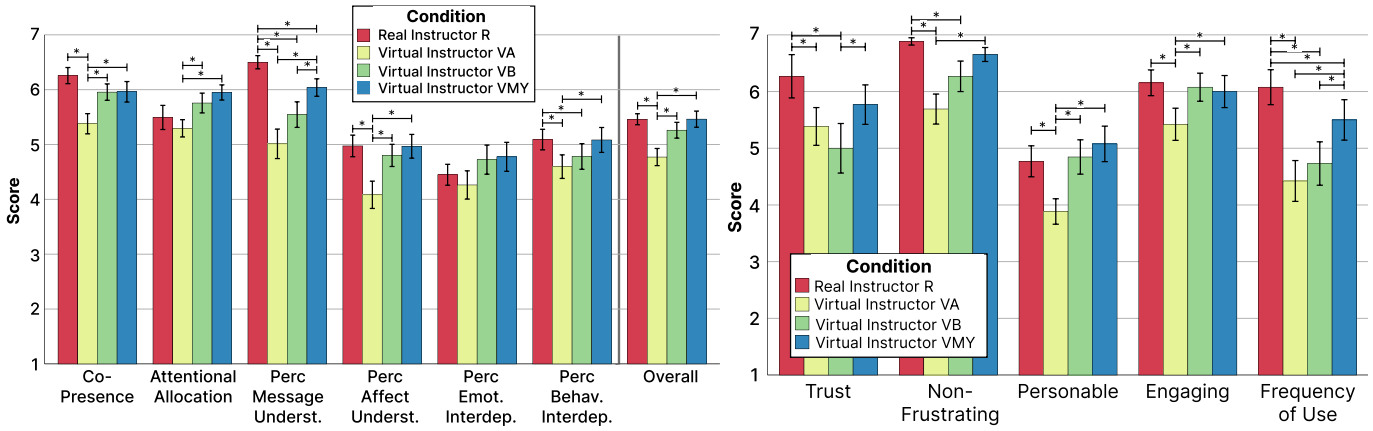
Harms-Biocca Social Presence Questionnaire The results for this questionnaire are shown in Figure 4(a).

We found a significant main effect for *Co-Presence*, $\chi^2 = 22.48$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. This confirms that the real instructor was more effective at producing co-presence than the virtual replica. The scores for VB were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Further, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

We also found a significant main effect for *Attentional Allocation*, $\chi^2 = 11.12$, $\boldsymbol{p = 0.011}$. Our post-hoc tests showed that the scores for VB were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Further, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

We found a significant main effect for *Perceived Message Understanding*, $\chi^2 = 27.15$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA, VB, and VMY (all $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VMY were also significantly higher than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

We found a significant main effect for *Perceived Affective Understanding*, $\chi^2 = 18.55$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VB were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Also, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.



(a) Harms-Biocca Social Presence

(b) Single-Item Questionnaires

Fig. 4: Subjective results: (a) **Harms-Biocca Social Presence questionnaire** with subscales co-presence, attentional allocation, perceived message understanding, perceived affective understanding, perceived emotional interdependence, and perceived behavioral interdependence, as well as the overall score. (b) **Single-Item Questionnaires** with scales trust, non-frustrating, personable, engaging, and frequency of use. Higher is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level.

We found no significant main effect for *Perceived Emotional Interdependence*, $\chi^2 = 4.34$, $p = 0.227$.

We found a significant main effect for *Perceived Behavioral Interdependence*, $\chi^2 = 14.42$, $\boldsymbol{p = 0.002}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VMY were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

And, finally, we found a significant main effect for the *Overall* subscale, $\chi^2 = 22.98$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VB were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Also, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

Overall, these results show strong support for all three hypotheses.

Single-Item Questionnaires

The results for the Single-Item Questionnaires are reported in Figure 4(b).

We found a significant main effect for *Trust* (how trustworthy the virtual instructor appeared), $\chi^2 = 17.96$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. Further, the scores for VMY were significantly higher than those of VB ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

We found a significant main effect for *Non-Frustrating* (with high scores indicating less frustration as indicated in Table 1), $\chi^2 = 26.82$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. Further, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

Moreover, we found a significant main effect for *Personable* (how personable the virtual instructor appeared), $\chi^2 = 17.96$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VB were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Further, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

Further, we found a significant main effect for *Engaging* (how engaging the virtual instructor was), $\chi^2 = 15.55$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. Also, the scores for VB were also significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H2**. Moreover, the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), which supports our Hypothesis **H3**.

Last but not least, we found a significant main effect for *Frequency of Use* (how frequently they would use this system), $\chi^2 = 34.59$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA, VB, and VMY (all $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. The scores for VMY were also significantly higher than those of VA and VB ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**.

Overall, these results support all three hypotheses.

User Experience Questionnaire

The subjective results for the UEQ are reported in Figure 5(a).

We found a significant main effect for *Attractiveness*, $\chi^2 = 21.06$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), which supports Hypothesis **H1**. Further, the scores for VMY were significantly higher than those of VA and VB ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**. The scores for VMY were even significantly higher than those for R ($\boldsymbol{p < 0.05}$), which we did not anticipate.

We also found a significant main effect for *Pragmatic Quality*, $\chi^2 = 21.24$, $\boldsymbol{p < 0.001}$. Our post-hoc tests further showed that the scores for R were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), which supports Hypothesis **H1**. Moreover, the scores for VMY were significantly higher than those of VA and VB ($\boldsymbol{p < 0.05}$), which is in support of Hypothesis **H3**.

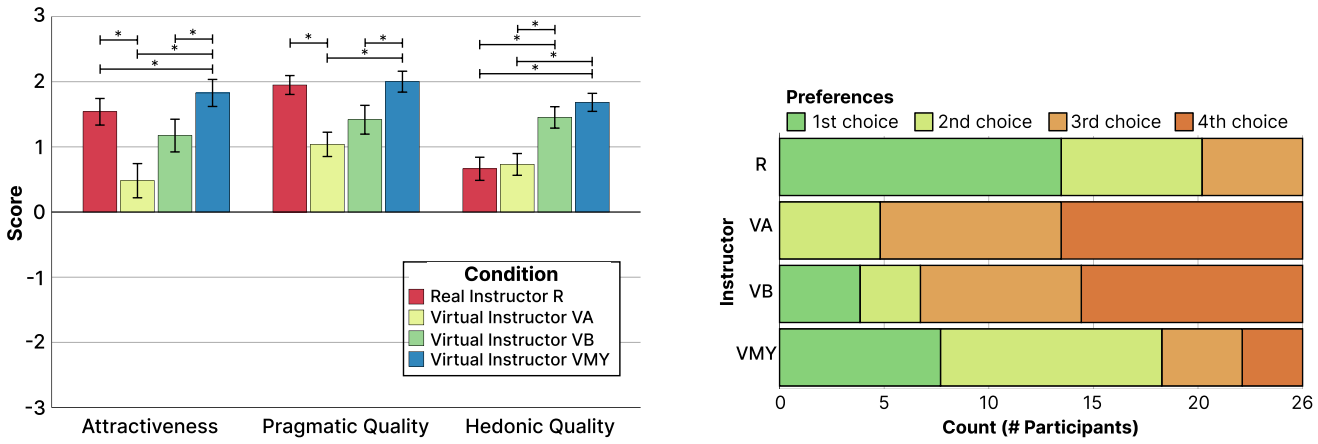
Last but not least, we also found a significant main effect for *Hedonic Quality*, $\chi^2 = 22.60$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the scores for VB were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), which supports Hypothesis **H2**. Our results further show that the scores for VMY were significantly higher than those of VA ($\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**. Additionally, we found that the scores for VMY were significantly higher than those of R and that the scores for VB were significantly higher than those for R (both $\boldsymbol{p < 0.05}$).

Overall, these results support our three hypotheses. However, while R was rated significantly worse than VB and VMY on some subscales, we found no evidence for a difference from VA. The personalities of the enhanced virtual instructors seemed to elicit a higher hedonic quality than the personality of the real or matched virtual instructors, supporting the basic notion that started this line of research.

Instructor Rankings

The results for the Instructor Rankings are reported in Figure 5(b).

We found a significant main effect for *Preference*, $\chi^2 = 27.09$, $\boldsymbol{p < 0.001}$. Our post-hoc tests showed that the rankings for R were significantly better than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**. Also, the rankings for VMY were significantly better than those of VA and VB (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H3**. We found no evidence for a difference in preferences between R and VMY.



(a) User Experience Questionnaire

(b) Instructor Rankings

Fig. 5: (a) Results for the **User Experience Questionnaire (UEQ)** with subscales attractiveness, pragmatic quality, and hedonic quality. Higher is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level. (b) **Instructor Rankings**: stacked bar chart showing the numbers of participants who ranked the four instructor conditions (R, VA, VB, VMY) in first to last place.

5 Discussion

In this section we discuss our main findings.

5.1 Differences Between Real and Virtual Instructors

5.1.1 Higher efficacy of real instructor compared to personality-matched virtual instructor (R > VA)

Supporting Hypothesis **H1**, our findings reveal clear advantages of the real instructor (R) over the personality-matched virtual instructor (VA), despite both sharing identical personality profiles. Participants rated R significantly higher than VA in attractiveness, pragmatic quality, and hedonic quality on the UEQ, underscoring the added impact of human delivery on perceived effectiveness and appeal. Notably, while VA closely mirrored R's personality and language patterns, participants still experienced meaningful differences, highlighting unique human qualities difficult to replicate virtually. R also outperformed VA across all Harms-Biocca Social Presence dimensions—suggesting stronger connection, clarity, and adaptability in human-led interactions. Single-item measures further showed R as less frustrating, more personable, and more engaging than VA, with participants preferring repeated interactions with R. Importantly, this subjective advantage translated to objective gains, with participants completing tasks 40% slower with VA than under R's guidance. Overall, R performed better in dynamic communication and real-time adaptability compared to VA. This gap reinforces prior work showing that trust and engagement depend not only on personality traits but also on the source of delivery [65]. For design, this suggests that replicating personality alone is insufficient—virtual instructors may need adaptive responsiveness and richer nonverbal behaviors to approximate human immediacy.

5.1.2 Higher efficacy of real instructor compared to enhanced virtual instructors (R > VB,VMY)

Further supporting Hypothesis **H1**, R also outperformed the enhanced virtual instructors—the Literature-Based Virtual Instructor (VB) and User-Customized Virtual Instructor (VMY)—particularly in task efficiency, clarity, and trust. Participants completed tasks 46% slower with VB than R and 31% slower with VMY than R. Based on our observations and informal comments from our participants, we believe that these differences stem mainly from the outgoing personalities of VB and VMY, which encouraged playful exchanges and exploratory questions that extended task duration. R was also rated higher in perceived message understanding and trustworthiness, suggesting clearer communication and greater credibility in human-delivered instruction. Additionally, participants preferred repeated interactions with R over both virtual instructors, emphasizing clarity and efficiency as key factors. While VB and VMY fostered more engaging and socially rich conversations, R remained especially effective for goal-oriented instructional tasks requiring clarity, speed, and trust. This aligns with prior findings that human instructors maintain an advantage in high-stakes or efficiency-driven settings where trust and precision are critical [19]. For design, this suggests virtual instructors may need explicit guardrails to balance sociability with task efficiency.

5.1.3 Higher efficacy of enhanced virtual instructors compared to real instructor in hedonic qualities (VB,VMY > R,VA)

Although the real instructor (R) excelled in efficiency and clarity, the enhanced virtual instructors (VB and VMY) surpassed R in hedonic quality on the UEQ. Participants rated VB and VMY significantly higher in the underlying aspects excitement, stimulation, and novelty. In contrast, both R and the personality-matched virtual instructor (VA) received lower hedonic ratings, suggesting that these impressions depended less on human embodiment and more on virtual personality design. These results underscore the potential of carefully crafted virtual instructor personalities to deliver lively, stimulating, and enjoyable interactions, with VB and VMY's playful language and encouragement of small talk likely driving their higher hedonic appeal. These findings highlight virtual instructors' distinctive strength in fostering socially engaging and entertaining learning experiences through dynamic personality expression. This echoes work showing that virtual agents can exceed humans in affective stimulation when designed for novelty and play [55]. For design, this highlights an opportunity to leverage personality exaggeration to make instruction more engaging, so long as efficiency demands are balanced.

5.2 Differences Between Virtual Instructors

5.2.1 Higher efficacy of user-customized virtual instructor compared to personality-matched virtual instructor (VMY > VA)

The objective of discovering how well a user-customized virtual instructor would perform against the personality-matched instructor matches our prediction in Hypothesis **H3** based on a variety of outcomes. Our results indicate a significant difference in attractiveness, the pragmatic quality, and the hedonic quality based on the UEQ, which suggests that customizing a personality influences how a participant perceives the utility and appeal of the instructor. In addition to this, we find that VMY exceeds VA in co-presence, attentional allocation, perceived message understanding, perceived affective understanding, and overall social presence based on the Harms-Biocca Social Presence questionnaire. This reveals that personality customization may result in a deeper connection between the user and the instructor, allowing for better communication and understanding. This enhancement in social presence has the potential to lead to more engaging virtual learning

experiences. The findings also highlight that VMY is significantly less frustrating, more personable, and would be used more frequently than VA. As these attributes show that interacting with VMY was overall more engaging over a longer period of time and potentially capable of reducing cognitive load of participants, it furthers the idea that customized personalities for virtual instructors improve the overall user experience when compared to a pre-defined personality. This resonates with prior findings that user control and personalization enhance trust and investment in AI systems [4,21,65], and parallels broader evidence that embodied agents foster social connection when tailored to user expectations [58].

5.2.2 Higher efficacy of literature-based virtual instructor compared to personality-matched virtual instructor (VB > VA)

When comparing the literature-based virtual instructor to that of the personality-matched virtual instructor, our findings indicate support for Hypothesis **H2**. For instance, the hedonic quality of VB surpassed that of VA, confirming the literature's suggestion that these personality traits are important for engaging participants in a manner which excites them. When combined with higher co-presence, perceived affective understanding, and overall social presence, it is clear the literature-based personality creates an environment which resonates with users and enhances their overall social connection. This is in line with the significant difference between VB and VA in the personable trait, where VB was consistently rated more approachable by participants. This is consistent with extensive research linking high openness, conscientiousness, and agreeableness to effective, student-centered instruction [39, 41, 56], and supports claims that these Big Five traits drive engagement in instructional contexts [46].

5.2.3 Higher efficacy of user-customized virtual instructor compared to literature-based virtual instructor (VMY > VB)

While evidence for Hypothesis **H3** was limited, one notable finding was that participants expressed a stronger desire to reuse the user-customized instructor (VMY) over the literature-based instructor (VB). This preference is intriguing because the aggregate personality profiles were strikingly similar: both featured high openness, high conscientiousness, high agreeableness, and low neuroticism, with VMY differing only in being less exaggerated. In other words, participants effectively recreated the "ideal" profile from the literature, yet still preferred their own version.

We believe this effect stems less from differences in the personalities themselves and more from the act of co-creation. Having control over sliders likely fostered a greater sense of personal connection and ownership, making VMY feel more engaging and worth returning to. This resonates with broader research on user innovation, which shows that individuals systematically prefer and value self-designed products—even when nearly identical to premade ones. Franke and Piller [22] demonstrated that self-designed goods command a substantial willingness-to-pay premium, while Franke et al. [23] describe this as the "I designed it myself" effect. So, even modest customization can increase attachment and willingness to reuse a system. From a design perspective, these results suggest that hybrid approaches—providing literature-based defaults while allowing user fine-tuning—may combine the strengths of validated models with the motivational benefits of customization.

Importantly, this was not a staged outcome: The similarity between VB and VMY only became apparent during post-hoc analysis, underscoring the robustness of the result. At the same time, it is worth noting that although participants preferred VMY for reuse, we found no significant evidence that VB was superior on any measure, indicating that the two approaches are at least comparable in effectiveness.

5.2.4 LLM stability and consistency

A frequent concern in LLM-mediated systems is variability of outputs across sessions or conversational turns. In our study, this threat was minimized through the use of GPT-4o with deterministic decoding (temperature = 0), fixed system/persona prompts with reverse-role constraints, and consistent few-shot style anchors. Beyond these controls, the stable reproduction of both VB and VMY profiles across multiple re-administrations of the IPIP-NEO further demonstrates GPT-4o's reliability in maintaining assigned personality configurations. Together with recent reports on the reproducibility of personality-linked language patterns under controlled prompting [3, 13, 42, 68], this supports our interpretation that the observed effects stem from the intended personality manipulations rather than model drift.

5.3 Limitations

A limitation of this research is that all instructor conditions were based on a single male instructor, which may limit the generalizability of our findings across instructor gender. Future investigation may examine female instructors or explore the effects of gender-matching between instructors and users. When addressing gender, it is to be noted there is a slight imbalance between the male and female participants in our study, future work should aim to recruit a larger demographic pool to gain a better understanding of the impacts on a wider scale. It is also to be acknowledged that a multitude of other factors can influence perception of instructors, including race, attire, body shape, and others. Future research should explore these factors to better shape how we design virtual instructors for assembly scenarios. We also note that facial expressions and body language were not modeled in the current implementation and are planned for future work. Additionally, the study was conducted within a controlled lab environment using a structured assembly task, which may not fully capture the complexities of real-world instructional scenarios. Expanding this work to diverse task domains and dynamic, real-world settings will provide a more comprehensive understanding of virtual instructor effectiveness.

6 Conclusion

In this paper, we investigated the role of personality design for virtual instructors in assembly-based instructional scenarios. By comparing a real-world instructor, a literature-based virtual instructor, and a user-customized virtual instructor, we examined how different personality configurations affect user experience, social presence, and task performance. While real instructors provided greater clarity, adaptability, and task efficiency, virtual instructors with literature-based enhancements or customized personalities significantly improved engagement, excitement, and user preference for future interactions. These findings highlight personality design as an important factor in developing effective AI-driven instructional systems. Future work may explore scalable personalization techniques, long-term interactions with instructor personalities, and the role of other factors, such as embodiment and communication style, in shaping instructional experiences.

Acknowledgments

This research was supported in part by the Office of Naval Research under Award No. N000142512245 (Dr. Peter Squire, Code 34), and the U.S. Department of the Army (Ground Vehicles Systems Center) under Award No. 2670-201-2016671. We also thank Dallas Kirkland for support with rigging and animating the virtual instructor character and Azhar Ali Mohammad for his valuable contributions to the initial conceptual development of the project. The views and findings expressed in this work are solely those of the authors and do not necessarily represent the official views of the Office of Naval Research or the U.S. Department of the Army.

References

[1] M. Bajwa, M. Morris, W. Ghias, and A. Linzels. Feasibility of holographic team training simulation: An information technology (it) perspective for healthcare and educational institutions. *Cureus*, 16(7):e65380, 2024. doi: 10.7759/cureus.65380 3

[2] Blender Online Community. Blender - Free and Open 3D Creation Software. https://www.blender.org/, 2025. Accessed: Apr. 10, 2025. 3

[3] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023. doi: 10.48550/arXiv.2303.12712 1, 3, 9

[4] Z. Buçinca, M. B. Malaya, and K. Z. Gajos. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–21, 2021. doi: 10.1145/3449287 2, 9

[5] A. Bucher, B. Schenk, M. Dolata, and G. Schwabe. When Generative AI Meets Workplace Learning: Creating A Realistic & Motivating Learning Experience With A Generative PCA. *arXiv, cs.HC, 2405.15561*, 2024. doi: 10.48550/arXiv.2405.15561 1

[6] R. B. Cattell. The description of personality: Basic traits resolved into clusters. *The Journal of Abnormal and Social Psychology*, 38(4):476–506, 1943. doi: doi.org/10.1037/h0054116 1

[7] S. Chen, P. Taveekitworachai, Y. Xia, X. Li, M. C. Gursesli, A. Lanata, A. Guazzini, and R. Thawonmas. Don't Do That! Reverse Role Prompting Helps Large Language Models Stay in Personality Traits. In J. T. Murray and M. C. Reyes, eds., *Interactive Storytelling*, pp. 101–114. Springer Nature Switzerland, Cham, 2025. doi: 10.1007/978-3-031-78453-8_7 2, 4, 5

[8] H. H. Clark and E. F. Schaefer. Contributing to discourse. *Cognitive science*, 13(2):259–294, 1989. doi: 10.1016/0364-0213(89)90008-6 3

[9] P. T. Costa and R. R. McCrae. A five-factor theory of personality. *Handbook of personality: Theory and research*, 2:139–153, 1999. 1

[10] W. Dai, J. Lin, H. Jin, T. Li, Y.-S. Tsai, D. Gašević, and G. Chen. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE international conference on advanced learning technologies (ICALT)*, pp. 323–325. IEEE, 2023. doi: 10.1109/ICALT58122.2023.00100 2

[11] F. De La Torre, C. M. Fang, H. Huang, A. Banburski-Fahey, J. Amores Fernandez, and J. Lanier. Llmr: Real-time prompting of interactive worlds using large language models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613904.3642579 3

[12] B. E. de Raad and M. E. Perugini. *Big five assessment.* Hogrefe & Huber Publishers, 2002. 1

[13] J. C. F. de Winter, T. Driessen, and D. Dodou. The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences*, 228:112729, Oct. 2024. doi: 10.1016/j.paid.2024.112729 9

[14] H. Dorloh, K.-W. Li, and S. Khaday. Presenting Job Instructions Using an Augmented Reality Device, a Printed Manual, and a Video Display for Assembly and Disassembly Tasks: What Are the Differences? *Applied Sciences*, 13(4), 2023. doi: 10.3390/app13042186 1

[15] D. Driess, F. Xia, M. S. M. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, W. Huang, et al. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023. 3

[16] ElevenLabs. ElevenLabs Text-to-Speech AI. https://elevenlabs.io/, 2025. Accessed: Apr. 10, 2025. 3

[17] Epic Games. MetaHuman Creator. https://www.unrealengine.com/en-US/metahuman, 2024. Accessed: Apr. 10, 2025. 3

[18] Epic Games. Unreal Engine 5. https://www.unrealengine.com/, 2025. Version 5.5, Accessed: Apr. 10, 2025. 3

[19] H. Fahnenstich, T. Rieger, and E. Roesler. Trusting Under Risk – Comparing Human to AI Decision Support Agents. *Computers in Human Behavior*, 153:1–8, 2024. doi: 10.1016/j.chb.2023.108107 2, 8

[20] F. Faul, E. Erdfelder, A. Buchner, and A.-G. Lang. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41:1149–1160, 2009. doi: 10.3758/BRM.41.4.1149 2

[21] R. Feldman, E. Aldana, and K. Stein. Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know. *Stanford Law & Policy Review*, 30:399–428, 2019. 2, 9

[22] N. Franke and F. Piller. Value creation by toolkits for user innovation and design: The case of the watch market. *Journal of product innovation management*, 21(6):401–415, 2004. doi: 10.1111/j.0737-6782.2004.00094.x 2, 9

[23] N. Franke, M. Schreier, and U. Kaiser. The "i designed it myself" effect in mass customization. *Management science*, 56(1):125–140, 2010. doi: 10.1287/mnsc.1090.1077 2, 9

[24] Q. Gan, Z. Liu, T. Liu, Y. Zhao, and Y. Chai. Design and user experience analysis of ar intelligent virtual agents on smartphones. *Cognitive Systems Research*, 78:33–47, 2023. doi: 10.1016/j.cogsys.2022.11.007 2

[25] L. R. Goldberg. Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, 2(1):141–165, 1981. 1

[26] L. R. Goldberg. An Alternative "Description of Personality": The Big-Five Factor Structure. 59(6):1216–1229, 1990. doi: 10.1037/0022-3514.59.6.1216 1

[27] C. Harms and F. Biocca. Internal consistency and reliability of the networked minds measure of social presence. In *Annual International Presence Workshop*, 2004. 5

[28] V. Hernandez Moreno, S. Jansing, M. Polikarpov, M. G. Carmichael, and J. Deuse. Obstacles and opportunities for learning from demonstration in practical industrial assembly: A systematic literature review. *Robotics and Computer-Integrated Manufacturing*, 86:102658, 2024. doi: 10.1016/j.rcim.2023.102658 1

[29] D. Iwai. Projection mapping technologies: A review of current trends and future directions. *The Japanese Journal of Applied Physics*, 63(SG):SG0801, 2024. doi: 10.2183/pjab.100.012 3

[30] Y. Ji, Z. Tang, and M. Kejriwal. Is persona enough for personality? Using ChatGPT to reconstruct an agent's latent personality from simple descriptions. In *Proceedings of the 41st International Conference on Machine Learning*, vol. 235. PMLR, Vienna, Austria, 2024. 4

[31] G. Jiang, M. Xu, S.-C. Zhu, W. Han, C. Zhang, and Y. Zhu. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643, 2023. 4

[32] O. P. John and S. Srivastava. The big five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin and O. P. John, eds., *Handbook of Personality: Theory and Research*, pp. 102–138. Guilford Press, 1999. 1

[33] J. A. Johnson. Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality*, 51:78–89, 2014. doi: 10.1016/j.jrp.2014.05.003 4

[34] H. J. Kell. Do teachers' personality traits predict their performance? a comprehensive review of the empirical literature from 1990 to 2018. *ETS Research Report Series*, 2019(1):1–27, 2019. doi: 10.1002/ets2.12241 2

[35] 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)*, pp. 105–114, 2018. doi: 10.1109/ISMAR.2018.00039 1, 2

[36] K. Kim, N. Norouzi, T. Losekamp, G. Bruder, M. Anderson, and G. Welch. Effects of Patient Care Assistant Embodiment and Computer Mediation on User Experience. In *Proceedings of the IEEE International Conference on Artificial Intelligence & Virtual Reality (AIVR)*, pp. 17–24, 2019. doi: 10.1109/AIVR46125.2019.00013 2

[37] L. E. Kim, V. Jörg, and R. M. Klassen. A meta-analysis of the effects of teacher personality on teacher effectiveness and burnout. *Educational Psychology Review*, 31:163–195, 2019. doi: 10.1007/s10648-018-9458-2 2

[38] L. E. Kim and C. MacCann. What is students' ideal university instructor personality? an investigation of absolute and relative personality preferences. *Personality and Individual Differences*, 102:190–203, 2016. doi: 10.1016/j.paid.2016.06.068 4, 5

[39] L. E. Kim and C. MacCann. Instructor personality matters for student evaluations: Evidence from two subject areas at university. *British Journal of Educational Psychology*, 88:584–605, 2018. doi: 10.1111/bjep.12205 2, 9

[40] K. Koleva, M. Vergari, T. Kojić, S. Möller, and J.-N. Voigt-Antons. Influence of Personality and Communication Behavior of a Conversational Agent on User Experience and Social Presence in Augmented Reality. *IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 929–930, 2024. doi: 10.1109/VRW62533.2024.00261 2

[41] M. Kunter, U. Klusmann, J. Baumert, D. Richter, T. Voss, and A. Hachfeld. Professional competence of teachers: effects on instructional quality and student development. *Journal of Educational Psychology*, 105(3):805–820, 2013. 2, 9

[42] S. A. Lehr, K. S. Saichandran, E. Harmon-Jones, N. Vitali, and M. R. Banaji. Kernels of selfhood: Gpt-4o shows humanlike patterns of cognitive dissonance moderated by free choice. *Proceedings of the National Academy of Sciences*, 122(20):e2501823122, 2025. doi: 10.1073/pnas.2501823122 2, 3, 9

[43] J. Lin, Z. Han, D. R. Thomas, A. Gurung, S. Gupta, V. Aleven, and K. R.

Koedinger. How can i get it right? using gpt to rephrase incorrect trainee responses. *International Journal of Artificial Intelligence in Education*, pp. 482–508, 2025. doi: 10.1007/s40593-024-00408-y 2

[44] X. Ma, Y. Bhalgat, B. Smart, S. Chen, X. Li, J. Ding, J. Gu, D. Z. Chen, S. Peng, J.-W. Bian, et al. When LLMs Step Into the 3D World: A Survey and Meta-Analysis of 3D Tasks via Multi-Modal Large Language Models. *arXiv preprint*, 2024. doi: 10.48550/arXiv.2405.10255 2

[45] F. Mairesse and M. Walker. Trainable generation of big-five personality styles through data-driven parameter estimation. In *Proceedings of ACL-08: HLT*, pp. 165–173, 2008. 4, 5

[46] S. Mammadov. Big Five personality traits and academic performance: A meta-analysis. *Journal of Personality*, 90(2):222–255, 2022. doi: 10.1111/jopy.12663 1, 9

[47] Microsoft Corporation. Azure AI Speech Service. https://azure.microsoft.com/en-us/products/ai-services/ai-speech, 2025. Accessed: Apr. 10, 2025. 3

[48] A. M. Mohammed, M. McCarthy, C. Neumann, G. Bruder, D. Reiners, and C. Cruz-Neira. ARIA: Toward human-centered embodied AI instruction in real-time augmented reality. In *31st International Conference on Intelligent User Interfaces (IUI '26)*. Association for Computing Machinery, Paphos, Cyprus, 2026. doi: 10.1145/3742413.3789163 2

[49] A. M. Mohammed, M. McCarthy, C. Neumann, G. Bruder, D. Reiners, and C. Cruz-Neira. The personalization paradox: Trade-offs between social presence and task efficiency in embodied ar instructors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, Barcelona, Spain, 2026. doi: 10.1145/3772318.3791902 2

[50] A. M. Mohammed, A. A. Mohammad, J. Ortiz, C. Neumann, G. Bochenek, D. Reiners, and C. Cruz-Neira. A human digital twin architecture for knowledge-based interactions and context-aware conversations. 2025. doi: 10.48550/arXiv.2504.03147 2

[51] 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 F. Argelaguet, R. McMahan, and M. Sugimoto, eds., *ICAT-EGVE 2020 - International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments*. The Eurographics Association, 2020. doi: 10.2312/egve.20201264 2

[52] OpenAI. GPT-4o – Fast, intelligent, flexible GPT model. https://platform.openai.com/docs/models/gpt-4o, 2024. Accessed: Apr. 10, 2025. 3

[53] OpenCV Contributors. OpenCV: Open Source Computer Vision Library. https://opencv.org/, 2025. Accessed: Apr. 10, 2025. 3

[54] R. Palmarini, J. A. Erkoyuncu, R. Roy, and H. Torabmostaedi. A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, 49:215–228, 2018. doi: 10.1016/j.rcim.2017.06.002 1

[55] P. Pataranutaporn, J. Leong, V. Danry, A. P. Lawson, P. Maes, and M. Sra. AI-Generated Virtual Instructors Based on Liked or Admired People Can Improve Motivation and Foster Positive Emotions for Learning. In *Proceedings of IEEE Frontiers in Education Conference (FIE)*, pp. 1–9, 2022. doi: 10.1109/FIE56618.2022.9962478 2, 8

[56] C. L. Patrick. Student evaluations of teaching: effects of the Big Five personality traits, grades and the validity hypothesis. *Assessment & Evaluation in Higher Education*, 36(2):239–249, 2011. doi: 10.1080/02602930903308258 2, 9

[57] Polycam, Inc. Polycam Photogrammetry Tool. https://poly.cam, 2025. Accessed: Apr. 10, 2025. 3

[58] J. Reinhardt, L. Hillen, and K. Wolf. Embedding Conversational Agents into AR: Invisible or with a Realistic Human Body? In *Proceedings of the International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 299–310, 2020. doi: 10.1145/3374920.3374956 2, 9

[59] C. Schindler, D. Mayumi, Y. Matsuda, N. Rach, K. Yasumoto, and W. Minker. ARCADE: An Augmented Reality Display Environment for Multimodal Interaction with Conversational Agents. In *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 85–87, 2024. doi: 10.1145/3686215.3688376 2

[60] M. Schrepp, A. Hinderks, and J. Thomaschewski. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In A. Marcus, ed., *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, pp. 383–392. Springer, Cham, 2014. doi: 10.1007/978-3-319-07668-3_37 6

[61] M. Seymour, L. Yuan, A. Dennis, and K. Riemer. 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), 2021. doi: 10.17705/1jais.00674 2

[62] W. Strus, J. Cieciuch, and T. Rowiński. The Circumplex of Personality Metatraits: A synthesizing model of personality based on the Big Five. *Review of General Psychology*, 18(4):273–286, 2014. doi: 10.1037/gpr0000017 1

[63] D. R. Thomas, J. Lin, E. Gatz, A. Gurung, S. Gupta, K. Norberg, S. E. Fancsali, V. Aleven, L. Branstetter, E. Brunskill, and K. R. Koedinger. Improving student learning with hybrid human-ai tutoring: A three-study quasi-experimental investigation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*, LAK '24, p. 404–415. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3636555.3636896 1

[64] T. Ueno, Y. Sawa, Y. Kim, J. Urakami, H. Oura, and K. Seaborn. Trust in human-ai interaction: Scoping out models, measures, and methods. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3491101.3519772 2

[65] K. Vodrahalli, R. Daneshjou, T. Gerstenberg, and J. Zou. Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, p. 763–777. Association for Computing Machinery, New York, NY, USA, 2022. doi: 10.1145/3514094.3534150 2, 8, 9

[66] I. Wang, J. Smith, and J. Ruiz. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300511 1, 2

[67] X. Wang, X. Li, Z. Yin, Y. Wu, and J. Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17, 2023. doi: 10.1177/18344909231213958 2

[68] Y. Wang, J. Zhao, D. S. Ones, L. He, and X. Xu. Evaluating the ability of large language models to emulate personality. *Scientific reports*, 15:519, 2025. doi: 10.1038/s41598-024-84109-5 1, 2, 3, 4, 9