The Personalization Paradox: Trade-offs Between Social Presence and Task Efficiency in Virtual Instructors

Abdul Mannan Mohammed
abdulmannan.mohammed@ucf.edu
University of Central Florida
Orlando, Florida, USA

Martin McCarthy
martin.mccarthy@ucf.edu
University of Central Florida
Orlando, Florida, USA

Carsten Neumann
carsten.neumann@ucf.edu
University of Central Florida
Orlando, Florida, USA

Gerd Bruder
bruder@ucf.edu
University of Central Florida
Orlando, Florida, USA

Dirk Reiners
dirk.reiners@ucf.edu
University of Central Florida
Orlando, Florida, USA

Carolina Cruz-Neira
carolina@ucf.edu
University of Central Florida
Orlando, Florida, USA

Abstract

Large language models continue to become utilized in training situations to power embodied virtual instructors in Mixed Reality (MR). As these models increase in sophistication, a key question emerges: does designing an agent with similarity to the instructee improve outcomes? We present a user study with four guided assembly conditions: a non-matched instructor employing real-life instructor's attributes, a personality-matched instructor, a gender- and voice-matched instructor, and a fully matched instructor reflecting the user's Big Five personality, cloned voice, and gender. Participants completed an ordered assembly task and reported on instructional quality. Results show that fully matched instructors were overwhelmingly preferred and significantly enhanced social presence and user experience. However, these subjective benefits did not translate into faster task completion, revealing a trade-off between engagement and efficiency. These findings offer critical guidance for designing future embodied virtual instructors and highlight the nuanced role of personalization in human–AI interaction.

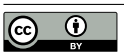
CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Mixed Reality (MR), Intelligent Virtual Agents, Personalization, Social Presence, Large Language Models (LLMs), Personality, Voice, Gender, User Study

ACM Reference Format:
Abdul Mannan Mohammed, Martin McCarthy, Carsten Neumann, Gerd Bruder, Dirk Reiners, and Carolina Cruz-Neira. 2026. The Personalization Paradox: Trade-offs Between Social Presence and Task Efficiency in Virtual Instructors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3772318.3791902



CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791902

1 Introduction

Virtual instructors are increasingly being integrated into mixed reality (MR) environments, offering immersive, interactive experiences for training and instructional tasks. Advances in conversational artificial intelligence (AI), particularly the development of large language models (LLMs), have significantly enhanced these agents' capabilities to generate dynamic, human-like dialogue [8, 9, 59]. As a result, virtual instructors are emerging as powerful tools for delivering engaging and personalized task guidance in fields ranging from manufacturing to education.

While the consistency of generative personas in LLMs has been a topic of discussion, recent advances and specific prompting methodologies have demonstrated a high degree of reliability in maintaining character traits, making such personalized studies viable [9, 43, 45, 76]. Despite these technological advances, a critical question remains: How does aligning a virtual instructor's attributes with individual user characteristics influence user experience, engagement, and task performance?

Prior work has explored the role of virtual instructor personalities [34, 46, 47], but there has been limited investigation into how deeper personalization—such as matching a virtual instructor's personality, voice, and gender to that of the user—affects social presence and task performance in MR. Research in human-computer interaction suggests that perceived similarity between users and virtual agents can foster social presence and user engagement [32, 40, 42, 60, 72]. Studies have shown that agents perceived as sharing a user's personality traits or communication style are often viewed as more likable and effective [46, 47, 56, 60]. Similarly, gender-matched agents may promote rapport and comfort, while voice matching has the potential to enhance familiarity and reduce social distance [4, 42, 69].

However, these effects remain largely underexplored in the specific context of MR-based virtual instruction. The unique combination of immersive presence afforded by MR and the newfound human-like consistency of modern LLMs [9, 43] presents a critical open question: Will the established benefits of personalization be amplified, or will they be mitigated by other factors in these rich, interactive settings?

Specifically, we examine four conditions that systematically vary the degree of similarity between the virtual instructor and the user across three dimensions: personality, voice, and gender. In our most personalized condition, the virtual instructor reflects the

user's Big Five personality profile [17], utilizes a cloned version of the user's own voice, and matches the user's gender. We compare this user-matched instructor against conditions with partial matching (gender only, gender and voice) and a non-matched control condition employing default attributes derived from a real-life instructor. Understanding these trade-offs is crucial for designing future embodied agents that are not only effective but also socially and personally resonant.

Based on these aims, our study addresses the following research questions:

RQ1: Does matching the virtual instructor's personality to the user's personality improve instructional efficacy and performance?

RQ2: Does matching the virtual instructor's voice to the user's own voice increase social presence and rapport?

RQ3: Does matching a virtual instructor's gender to the user improve social presence, user experience, and task performance?

RQ4: Does a fully user-matched instructor outperform partial or no matching in terms of social presence and preferences?

In the remainder of this paper, we first review related work in Section 2. We then describe the design of our virtual instructor testbed, experiment design, and conditions in Section 3. The results, including both objective and subjective data, are presented in Section 4. We discuss these findings in Section 5. We conclude the paper in Section 6 and outline future work.

2 Related Work

LLMs have transformed text-based interactions, producing contextually relevant and human-like language across diverse applications such as education, healthcare, and customer service [8, 59]. Text-based interfaces are valued for their scalability and efficiency, but they lack the rich multimodal signals—tone, emotion, gaze—that are essential for effective social and instructional interactions [10, 65]. Without embodiment or voice-based communication, purely text-driven AI systems may fall short in creating social presence or fostering rapport, particularly in tasks that benefit from human-like communication cues, even as these models demonstrate high emotional intelligence [61, 75]. In contrast, virtual agents that integrate visual embodiment and synthetic voices can enhance the sense of social presence, engagement, and perceived competence [4, 34]. Speech conveys crucial prosodic information—intonation, rhythm, and emphasis—that can clarify meaning and convey affective states [42]. Embodied conversational agents have been shown to improve user engagement and learning outcomes by leveraging both verbal and nonverbal signals [3, 40]. Such multimodal systems are increasingly relevant for instructional applications, where conveying empathy, encouragement, and social connection is critical [34, 71].

2.1 Virtual Instructors

Virtual instructors have gained prominence as scalable, interactive solutions in educational technology, offering guidance across domains from science learning to practical assembly tasks [25, 31]. Early pedagogical agents focused on delivering scripted guidance and answering questions, often with simple visual avatars [54]. Studies have consistently shown that virtual instructors can reduce cognitive load, improve procedural learning, and foster positive attitudes toward instruction [1, 2]. Advancements in mixed and augmented reality (MR/AR) and conversational AI have enabled virtual instructors to deliver real-time, context-aware support, bridging physical and digital worlds [34, 66]. Recent advances have shown that real-time embodied AI architectures combining multimodal sensing, low-latency interaction, and controllable LLM-based persona modeling can support socially present, adaptive, and highly engaging virtual instructors in immersive instructional environments [51–53]. LLM-powered instructors can provide detailed, coherent feedback that sometimes surpasses human teachers, such as generating more fluent and detailed performance summaries than instructors or rephrasing incorrect trainee responses into correct ones with near-human accuracy [14, 45]. Research has demonstrated that embodied virtual instructors can improve task performance and social presence compared to voice-only or purely text-based systems [34, 74]. This has created a critical gap, however, between the rapidly advancing technical fidelity of these agents and the nuanced understanding of their social effectiveness. The field has confirmed that embodied instructors are beneficial, but the fundamental question of who this instructor should be for each individual user remains largely unanswered [39, 66].

2.2 Instructor Characteristics

Beyond an instructor's technical capabilities, their social and relational effectiveness is governed by key characteristics such as personality, voice, and gender. Drawing from the foundational principle of similarity-attraction [42, 56], a compelling hypothesis emerges: aligning these traits with the user may significantly enhance social presence and instructional outcomes. However, the efficacy of this approach—and how it manifests across these different dimensions—remains largely untested within immersive MR environments.

2.2.1 Instructor Personality. Instructor personality is a key factor in shaping learning outcomes, user engagement, and perceptions of effectiveness [38, 62]. Studies using the Big Five framework have shown that instructors high in openness, conscientiousness, extraversion, and agreeableness tend to be perceived as more effective, supportive, and engaging [33, 41]. These traits correlate with innovative teaching methods, classroom management, and students' performance and satisfaction [37, 48]. Recent work has confirmed that modern LLMs can consistently maintain distinct personas based on the Big Five model [17]. Beyond ideal instructor profiles, research suggests that similarity in personality between user and agent may foster smoother communication, increased trust, and enhanced social presence [40, 56]. Users tend to prefer conversational partners, human or artificial, who exhibit similar personality traits, a phenomenon observed across both human-human and human-agent interactions [32, 60, 65, 72]. This presents a fundamental tension for instructional design: should a virtual instructor embody a universally 'ideal' personality (e.g., high in conscientiousness), or should it instead mirror the user's own personality to leverage the benefits of similarity-attraction? This question becomes particularly critical in immersive MR, where heightened social presence could amplify the effects of either similarity or dissonance. Yet, empirical evidence to guide this crucial design choice is conspicuously absent [39, 66].

2.2.2 Instructor Voice. Voice characteristics are a powerful social cue, shaping perceptions of trust, competence, and emotional warmth [55, 69]. Advances in neural text-to-speech have enabled high-fidelity, personalized voice synthesis capable of replicating specific speakers with remarkable accuracy [29, 64]. Research shows that users quickly form social judgments based on an agent's voice, including impressions of likability, expertise, and emotional connection [4, 5, 69]. A particularly novel research frontier involves matching an agent's voice to that of the user. While few studies have directly investigated hearing one's own cloned voice, research on similarity-attraction suggests that hearing a voice that resembles one's own may enhance feelings of familiarity, reduce social distance, and foster stronger social connection [32, 40, 72]. This positions two powerful psychological forces in direct opposition: the predicted comfort of similarity-attraction versus the potential unease of a vocal uncanny valley, a phenomenon where a voice that is almost, but not quite, human is perceived as unpleasant or "eerie," an active area of research in speech synthesis [67]. Will hearing one's own voice from an embodied, co-present MR instructor be a powerful tool for building rapport, or will it be an unsettling distraction that shatters social presence? While voice cloning technology is now mature [29, 64], the answer to this critical question remains unknown.

2.2.3 Instructor Gender. Gender similarity between users and virtual agents has been explored as a potential moderator of trust, engagement, and social comfort [42, 55]. Research has shown that users often report higher levels of trust and rapport when interacting with virtual agents of the same gender, particularly in instructional or collaborative tasks [54, 57]. Gender-congruent instructors may facilitate greater social presence, perceived empathy, and user comfort [40, 42]. However, findings are mixed, suggesting that gender preferences can depend on cultural norms, the instructional domain, and task context [3, 40]. This sets up a direct conflict between the social comfort derived from gender similarity and the powerful influence of social stereotypes regarding task competence [55, 57]. The embodied, socially present nature of an MR virtual instructor provides a critical and unexamined context to investigate this trade-off. It remains an open question whether the immersive experience will strengthen the rapport of a gender-matched agent or amplify the perceived authority of a stereotype-congruent one, making this a crucial area for investigation [66].

3 Experiment

In this section we describe the user study that we conducted to examine how matching virtual instructor attributes to individual users impacts engagement, social presence, and task performance. The following sections describe the participants, material, system architecture, and methods. The user study was approved by the University of Central Florida Institutional Review Board (IRB).

3.1 Participants

Following a power analysis with G*Power on the basis of anticipated strong effects [22], we recruited 25 participants from our university community, including both students and non-student members who responded to open calls for participation. Participants ranged in age between 19 and 60 years, with a mean age of 25.7 (SD = 8.5). Our participant sample included 10 female and 15 male individuals. All of them had normal or corrected-to-normal vision.

We collected participants' Big Five personality profiles using the Ten-Item Personality Inventory (TIPI) [24] for use in the experimental conditions. The personalities of the participant population (on a 0-to-7 range) indicated on average an *Openness* of M = 5.73 (SD = 0.86), *Extraversion* of M = 4.84 (SD = 1.05), *Agreeableness* of M = 5.36 (SD = 1.07), *Conscientiousness* of M = 5.84 (SD = 0.65), and *Neuroticism* of M = 3.14 (SD = 1.12).

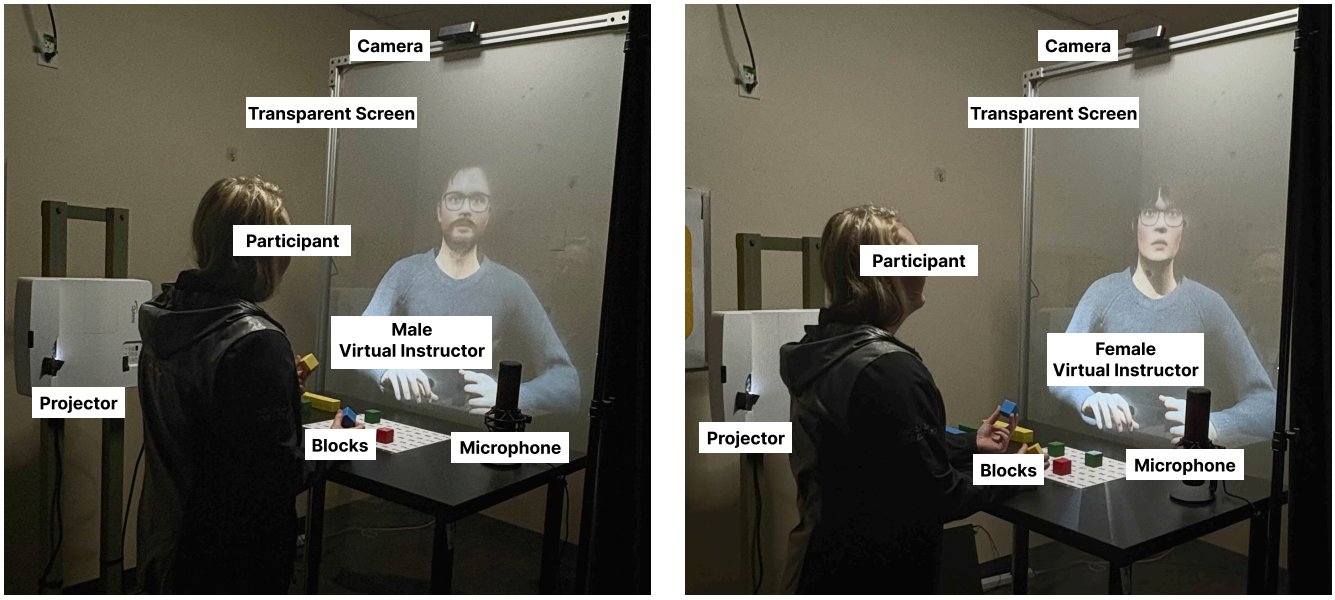
3.2 Apparatus and Materials

As illustrated in Figure 1, the experimental setup featured a rectangular table (0.99 m × 0.63 m) with a holographic-effect display [77] placed between the participant and the virtual instructor. We classify this configuration as a MR environment, consistent with the taxonomy of optical see-through displays defined by Milgram and Kishino [50]. By establishing a condition of "virtual corporeal presence" within the physical domain, this setup effectively merges the virtual agent into the real-world task space without requiring the isolation or encumbrance of Head-Mounted Displays (HMDs) [23, 27, 36]. To reinforce the sense of a shared environment, the virtual scene was aligned to visually extend the physical table into the digital space, creating a continuous "shared workbench." This spatial continuity leveraged the transparent display to produce the illusion that the virtual instructor was standing directly across the table in the physical room, fostering a sense of natural face-to-face collaboration. Crucially, to establish bidirectional co-presence, the webcam (detailed below) functioned as the virtual agent's "visual system," enabling it to verify physical assembly states in real-time. This alignment created a shared perceptual space where the agent appeared to occupy the physical room and "see" the same objects as the user, satisfying the core criteria of MR interaction without the encumbrance of wearable hardware while leveraging a display medium demonstrated to enhance learning and social presence in instruction-based scenarios [15, 23, 36]. The screen combined a rigid Plexiglass substrate and a nano-optic film[1], enabling high-quality projection while maintaining see-through visibility. It measured 1.04 m in width and 2.05 m in height, with the nano-optic layer directly adhered to the Plexiglass surface.

Visual output was provided by an Optoma EH340UST ultra-short-throw projector (1920 × 1080 resolution, 22,000:1 contrast ratio, 4000 ANSI lumens), positioned 0.35 m from the screen and offset laterally by 0.36 m to stay outside the participant's direct line of sight. Projection calibration was conducted in Unreal Engine 5.5 (UE 5.5) [21] using in-house tools, with consistent ambient lighting maintained via two low-intensity ring lights flanking the screen.

The virtual instructor's 3D model was created by scanning the real instructor using Polycam [63], refining the mesh in Blender [7], and generating a MetaHuman character [20]. This custom pipeline was chosen over standardized avatar systems to leverage the state-of-the-art photorealism and micro-detail capabilities of the MetaHuman framework. This approach allowed us to create an accurate visual replica of the real-world instructor, ensuring that the avatar's visual likeness aligned with the specific real-world voice and personality profile used in the non-matched control conditions. The

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



(a) Male Virtual Instructor

(b) Female Virtual Instructor

Figure 1: Annotated photos showing a participant performing the assembly task in the experiment with the (a) male instructor and (b) female instructor. The setup consisted of the table space in front of the divider, the transparent screen, and the virtual instructor rendered behind the screen.

avatar was imported into UE 5.5 and animated with four basic states (idle, listening, thinking, speaking) to ensure uniform nonverbal cues. All virtual instructor conditions used this same model, differing only in personality, voice, or gender as described in Section 3.4. Rendering ran on a Windows 11 Pro workstation with an Intel Core i7-8750 CPU, 32GB RAM, and an NVIDIA RTX 4070 GPU.

Speech synthesis was performed via the ElevenLabs TTS API [19]. For conditions requiring the real instructor's voice, speech was generated from recorded samples to maintain consistency across virtual appearances. For user-specific voice matching, participants' voices were cloned using the same TTS technology. Participants spoke to the system via a HyperX QuadCast S microphone, with audio output delivered through Logitech Multimedia Z150 speakers.

Assembly progress was monitored with a Depstech DW50 4K webcam (3840 × 2160 at 30 fps, 90° field of view) mounted above the screen, capturing the workspace where participants assembled magnetic block structures. Four unique block figures were used, each with similar complexity and consisting of 28 pieces (see Figure 2). During the task, participants placed blocks onto a predefined, labeled 7x9 grid following step-by-step instructions from the virtual instructor, who specified precise grid positions for each piece. Since the task only required placing uniformly sized, square-shaped blocks onto specific grid positions regardless of their color, the task did not require color vision.

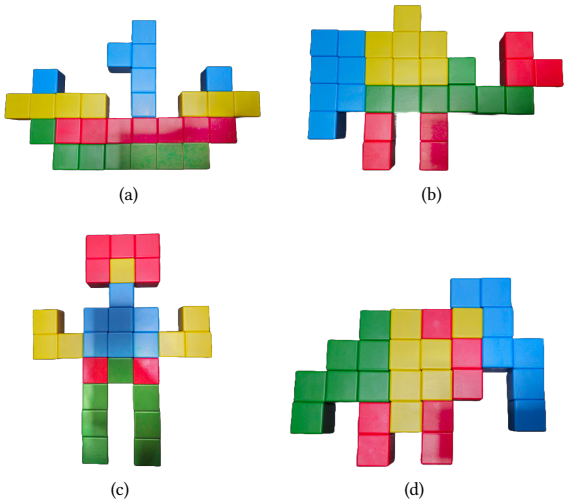


Figure 2: The four magnetic block structures (a)–(d) with comparable difficulties that participants assembled on the table in the experiment. The tasks were randomized for each participant between the conditions in the experiment.

3.3 System Architecture

The system architecture enabled our AI-driven virtual instruction by integrating principles from real-time embodied environments [16] and multimodal LLM-driven agent systems [18]. Rather than benchmarking various models, our design priority was to select a state-of-the-art model with proven instruction-following and persona-adherence capabilities to serve as a stable foundation for our personalization variables. We specifically chose GPT-4o [58] as it has demonstrated remarkable performance on a wide range of tasks [9], shows a high degree of reliability in emulating personality traits when using principled prompting and temperature controls [76], and exhibits predictable, humanlike consistency

under psychological frameworks, making it a robust model for personalization studies [43]. The architecture, managed by a central Python controller, consists of the following components:

- **User Interaction Layer:** Captures natural speech input from the participant via a microphone and monitors their assembly progress with an overhead RGB camera.
- **Speech and Input Processing:** The controller sends captured audio to Azure's Speech-to-Text API [49] for real-time transcription. It then parses the returned text for explicit confirmation requests (e.g., "verify," "check," "is this right?"). This keyword-based trigger is based on the principle of conversational grounding, where participants in a dialogue use such requests to establish mutual understanding during a collaborative task [13]. Upon detecting a grounding request, the system triggers an image capture via OpenCV to provide visual feedback.
- **Prompt Generation:** The controller dynamically constructs a prompt for GPT-4o, combining system instructions (including the agent's persona), the dialogue history, the user's new utterance, and a base64-encoded image, if captured.
- **Response Synthesis:** The controller sends the prompt to the GPT-4o API and receives a textual response. This text is then immediately sent via another API call to ElevenLabs to be synthesized into speech.
- **Virtual Instructor Rendering and Synchronization:** The controller receives the audio stream from ElevenLabs. It simultaneously sends a WebSocket command for MetaHuman's animations in Unreal Engine through a simple state machine. When the controller received an audio stream from ElevenLabs, it would simultaneously send a command to play a general"speaking" animation loop while the audio played, achieving a basic synchronization between mouth movement and speech. Once the audio finished, a command was sent to transition the instructor back to an"idle" animation. When the system detected that the participant was speaking, it triggered a"listening" animation. After the participant finished, the instructor would briefly enter a"thinking" animation as a fallback to signify processing, though this state was often momentary due to the system's fast response time.

For verbal-only interactions, which involved only text-based processing, the end-to-end response time averaged M = 0.80 s (SD = 0.19 s). The overall system latency across all turns, which includes the additional processing time for multimodal interactions involving image analysis, averaged M = 1.38 s (SD = 0.25 s), supporting a natural, low-latency interaction.

3.4 Experiment Design and Conditions

We employed a within-subjects design to address our research questions. Participants experienced four experimental conditions in a randomized order: VA (all attributes matched), VG (gender-matched only), VGV (gender and voice matched), and VX (no-match control). For each condition, participants were asked to assemble one of four magnetic block figures of comparable difficulty (see Figure 2), with the specific task also being randomized.

In the following, we present additional details about these conditions:

- **VA:**

In this condition, the virtual instructor was matched to the individual participant across three key dimensions:

– **Personality Matching**: Participants completed the Ten-Item Personality Inventory (TIPI) [24] to assess their Big Five personality traits. To translate these scores into a conversational style, our approach was inspired by the PERSONAGE system [46, 47], which links personality traits to linguistic variables like verbosity, polarity, and syntactic complexity.
  Since PERSONAGE does not provide fixed numeric cutoffs, we first discretized participants' TIPI scores into qualitative categories (e.g., Extraversion 1.0–2.4 → reserved, low-verbosity style; Extraversion 5.5–7.0 → sociable, high-verbosity style). To ensure the virtual instructor reliably embodied these categories, we implemented several principled prompt engineering strategies using GPT-4o, a model noted for its improved capacity for consistent persona simulation. Specifically, our system prompt first defined the target personality using the five dimensions with their corresponding numeric scores, a technique similar to the PERSONALITY PROMPTING (P^2) method [30]. This was followed by Reverse Role Prompting, where the model was instructed on behaviors to avoid, helping it maintain the assigned traits more effectively [11]. 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 [28, 76].
  Finally, to create a greater sense of conversational mirroring, we employed a few-shot prompting technique [8, 70]. The system was provided with examples from a three-minute transcribed speech sample—in which participants introduced themselves and shared their views on AI—to allow the virtual instructor to adopt the user's characteristic filler words and idiosyncratic speech patterns into its dialogue.
– **Voice Matching**: Participants' voices were cloned using ElevenLabs TTS voice cloning features. The virtual instructor delivered all speech using a synthesized voice modeled on the participant's own.
– **Gender Matching**: The virtual instructor's gender was matched to that of the participant.

The VA condition was designed to explore the maximal effect of user similarity in shaping user experience and engagement.

- **VG:**

In this condition, the virtual instructor exhibited the following characteristics:

– **Gender Matching**: The virtual instructor's gender matched the participant's gender.
– **Voice**: The virtual instructor's voice was unrelated to the participant's voice. The instructor spoke using the voice of a real person—an instructor of the laboratory where this experiment was conducted, who agreed to be voice-cloned for the purpose of this experiment. The instructor spoke with a pitch-adjusted version of the real instructor's cloned voice to match the avatar's gender using ElevenLabs TTS.
– **Personality**: The virtual instructor's personality profile was based on the same real instructor mentioned above, who agreed to have their personality captured using the same TIPI questionnaire

and integrated into the virtual instructor's system prompt using the same method as for user personality mapping.

The VG condition isolated the impact of gender congruence without voice or personality personalization.

• **VGV:**

In this condition, the virtual instructor exhibited the following characteristics:

– **Gender Matching**: The virtual instructor's gender matched the participant's gender.
– **Voice Matching**: The virtual instructor's voice was cloned from the participant's own voice, as in the VA condition.
– **Personality**: The virtual instructor's personality was unrelated to the participant's personality. The virtual instructor retained the same real instructor's personality profile as detailed for VG.

The VGV condition examined whether voice matching combined with gender matching alone would influence user experience and performance, independent of personality alignment.

• **VX:**

This condition served as a control where the virtual instructor's attributes were based on a single, real instructor but were intentionally not matched to the participant.

– **Gender**: The virtual instructor's gender was set opposite to the participant's. We created this avatar by generating a "gender-bent" version of the real-life instructor using MetaHumans. This approach was chosen over utilizing a different real-life instructor for two reasons. First, it ensured internal validity by holding the instructor's personality profile and linguistic style constant; a different human actor would have introduced confounding variations in teaching style. Second, MetaHuman Creator provided industry-standard tools for realistic gender transformation while preserving the original avatar's rigging and animation fidelity, ensuring that the control condition maintained the same level of visual realism as the matched conditions.
– **Voice**: The virtual instructor's voice being pitch-adjusted to match the opposite-gender avatar.
– **Personality**: The virtual instructor used the real-life instructor's Big Five personality profile.

The VX condition was designed to serve as a high-fidelity, ecologically valid baseline of a non-matched instructor.

3.4.1 Task and Procedure. Upon providing informed consent, participants were introduced to the study goals and the experimental setup. They then completed two preparatory tasks: (1) a 30-second voice recording for voice cloning via ElevenLabs TTS, and (2) the Ten-Item Personality Inventory to assess their Big Five personality traits [24]. To capture idiosyncratic speech patterns for the fully matched condition (VA), participants also provided a three-minute verbal response giving introduction and their opinion about AI.

Following these steps, each participant then experienced all four virtual instructor conditions (VA, VG, VGV, VX) in counterbalanced order, with each condition lasting approximately six minutes. Each condition used a different randomly assigned magnetic block assembly task of comparable difficulty (see Figure 2).

After interacting with all four instructors, participants completed a final set of questionnaires, including the User Experience Questionnaire (UEQ), the Social Presence Assessment, and a demographics survey (see Section 3.4.2). The entire session lasted approximately 60 minutes.

3.4.2 Feedback Measures. We collected the following objective and subjective measures to evaluate each condition:

(1) **Task Performance Metrics**:

• *Task Duration: Total time taken by participants to complete each assembly task, measured through automated timestamp logging.*
• *Instructor Errors: Instances where the virtual instructor provided incorrect or misleading instructions, recorded via live observation.*
• *Instructor Clarifications: Occasions when the virtual instructor was called upon for further clarifications of the step.*
• *Corrections: Occasions when the virtual instructor corrected prior instructions, recorded via live observation.*

(2) **User Experience Questionnaire (UEQ)**: We employed the short version of the UEQ [68] to assess hedonic and pragmatic qualities as well as overall user experience with each instructor.

• *Participants responded to nine bipolar adjective pairs on a 7-point semantic differential scale.*
• *Responses form three dimensions: Pragmatic Quality, Hedonic Quality, and Overall.*

(3) **Social Presence Assessment**: We used the 36-item Harms-Biocca Social Presence questionnaire [26] to measure participants' perceptions of social presence across conditions.

• *Participants rated each item on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).*
• *Scores were averaged within six subscales and combined into an overall social presence score.*

(4) **Instructor Rankings**: After completing all conditions, participants ranked the four virtual instructors in order of overall preference (1st choice to 4th choice).

3.4.3 Hypotheses. The following hypotheses were derived to address the research questions (RQ1–RQ4) presented in Section 1. These hypotheses reflect our expectations regarding how different levels of user-matching in the virtual instructor's attributes may influence user experience and task outcomes.

H1: A fully user-matched virtual instructor (Condition VA) will result in the highest performance, social presence, user experience, and preference among the conditions.

H2: Gender matching alone (Condition VG) will improve performance, social presence, user experience, and preference over no matching (Condition VX).

H3: Voice matching with gender matching (Condition VGV) will produce higher performance, social presence, user experience, and preference than gender matching alone (Condition VG).

H4: No matching (Condition VX) will result in the lowest performance, social presence, user experience, and preference among the conditions.

As discussed in Section 2, these hypotheses are grounded in prior research indicating that perceived similarity between users and virtual agents—whether in personality, voice, or gender—can enhance social presence, usability, and user preference [40, 42, 46].

However, much of this evidence stems from studies with simpler agents, chatbots, or non-immersive settings. Our study investigates whether these benefits of user-agent similarity extend to virtual AI instructors in AR environments, where advanced conversational abilities and lifelike avatars may amplify or alter the impact of user-agent similarity on user experience and task performance.

4 Results

In this section, we present the results of our statistical analysis of the participant responses.

4.1 Objective Data

The objective results are reported in Figure 3. We had to exclude three data sets from this analysis due to a malfunction in the recording equipment.

We report the results of a one-way repeated-measures ANOVA, comparing the four instructor conditions. We verified the assumptions of the statistical test [73]. This includes the *scale data type* of the objective data, the *normality* of the data, which we confirmed with Shapiro-Wilk normality tests (all indicating that the hypothesis of non-normality was not supported, $p > 0.05$), and *sphericity* of the data, which we tested for with Mauchly's test of sphericity. Mauchly's test indicated that the hypothesis of non-sphericity was not supported for the measures, $p > 0.05$, except for the *Instructor Corrections* measure. For this measure, we used Greenhouse-Geisser adjustments to correct for sphericity. For the multiple comparisons, we used paired samples post-hoc t-tests with Bonferroni correction.

Task Duration. We found a significant main effect of the instructor conditions on the duration of the assembly tasks, $F(3, 63) = 3.70$, $\boldsymbol{p = 0.016}$, $\eta_p^2 = 0.15$. Our post-hoc tests showed that the tasks were completed significantly slower for VX compared to VG ($\boldsymbol{p < 0.05}$). This effect supports in particular our Hypothesis **H2**.

Instructor Errors, Clarifications, Corrections. We found no significant main effect for the occurrences of the instructor making errors, $F(3, 63) = 2.10$, $p = 0.11$, $\eta_p^2 = 0.09$, clarifications, $F(3, 63) = 1.61$, $p = 0.20$, $\eta_p^2 = 0.07$, or corrections, $F(3, 63) = 1.00$, $p = 0.14$, $\eta_p^2 = 0.09$.

4.2 Subjective Data

We present the results for the rating scales and rankings in the different questionnaires for the four instructor conditions. As these subjective data are based on ordinal data scales, we analyzed them with non-parametric Friedman tests and pairwise Wilcoxon Signed Rank tests with Bonferroni correction for the post-hoc comparisons.

4.2.1 User Experience Questionnaire.

Figure 4 shows the results for the UEQ.

We found a significant main effect for *Overall User Experience*, $\chi^2 = 8.04$, $\boldsymbol{p = 0.045}$, Kendall's $W = 0.11$. Our post-hoc tests showed that the scores for VA were significantly higher than those of VG and VX (both $\boldsymbol{p < 0.05}$), which supports in particular Hypothesis **H1**.

We also further found a significant main effect for *Hedonic Quality*, $\chi^2 = 8.55$, $\boldsymbol{p = 0.036}$, Kendall's $W = 0.11$. Our post-hoc tests showed that the scores for VA were significantly higher than those of VG and VX (both $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**.

We found no significant main effect for *Pragmatic Quality*, $\chi^2 = 1.65$, $p = 0.65$, Kendall's $W = 0.02$.

Overall, these results support our Hypothesis **H1**.

4.2.2 Harms-Biocca Social Presence questionnaire.

Figure 5 shows the results for the Harms-Biocca Social Presence questionnaire.

We found a significant main effect for *Co-Presence*, $\chi^2 = 12.17$, $\boldsymbol{p = 0.007}$, Kendall's $W = 0.16$. Our post-hoc tests showed that the scores for VA were significantly higher than those of VG, VGV, and VX (all $\boldsymbol{p < 0.05}$), supporting Hypothesis **H1**.

We further found a significant main effect for *Perceived Behavioral Interdependence*, $\chi^2 = 14.66$, $\boldsymbol{p = 0.002}$, Kendall's $W = 0.20$. Our post-hoc tests showed that the scores for VX were significantly lower than those of VA, VG, and VGV (all $\boldsymbol{p < 0.05}$). In other words, all other instructors were rated better than VX, supporting Hypothesis **H4**.

We found no significant main effect for *Attentional Allocation*, $\chi^2 = 3.26$, $p = 0.35$, Kendall's $W = 0.04$, *Perceived Message Understanding*, $\chi^2 = 1.61$, $p = 0.66$, Kendall's $W = 0.02$, *Perceived Affective Understanding*, $\chi^2 = 2.21$, $p = 0.53$, Kendall's $W = 0.03$, *Perceived Emotional Interdependence*, $\chi^2 = 6.02$, $p = 0.11$, Kendall's $W = 0.08$, and the *Overall* scale, $\chi^2 = 4.94$, $p = 0.18$, Kendall's $W = 0.07$.

Overall, these results support our Hypotheses **H1** and **H4**.

4.2.3 Instructor Rankings.

We present the results of the instructor rankings in Figure 6.

We found a significant main effect for *Preference*, $\chi^2 = 14.66$, $\boldsymbol{p = 0.002}$, Kendall's $W = 0.20$. Our post-hoc tests showed that the rankings for VX were significantly worse than those of VA, VG, and VGV (all $\boldsymbol{p < 0.05}$), supporting Hypothesis **H4**.

4.2.4 Lexical Diversity.

To verify that the deterministic temperature setting (T = 0) preserved generative quality, we analyzed the lexical diversity of the virtual instructor's dialogue. We computed Distinct-2 (bigram) scores, a standard metric for assessing response variety in dialogue-generation systems [44].

The Fully Matched (VA) condition achieved a Distinct-2 score of 0.41, slightly higher than the VGV (0.39) and VX (0.37) conditions. These values exceed those commonly reported for standard sequence-to-sequence dialogue baselines (often <0.20) and fall above the range typically observed in personalized dialogue systems (0.10–0.30) [44, 78]. This level of lexical diversity suggests that the virtual instructor did not rely on repetitive templates but instead generated a varied set of utterances across participants.

The comparable spread of scores across conditions also aligns with recent findings on LLM behavioral stability and steerability [76]. In particular, the VA condition—whose prompt incorporated the most detailed persona specification—appears to have leveraged this steerability to produce the richest conversational output. Taken together, these results indicate that the system maintained high generative diversity despite deterministic sampling, supporting its role as a dynamic conversational agent rather than a static script-driven system.

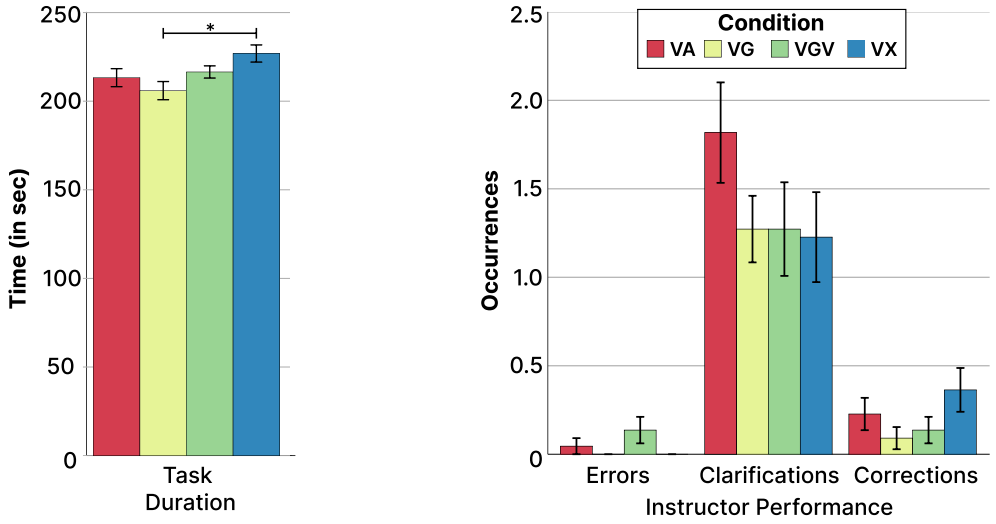


Figure 3: Objective results indicating the task performance of the participants and instructors. The bar charts shows the results for (a) task duration and (b) instructor errors, clarifications, and corrections for the four instructor conditions (VA, VG, VGV, VX). Lower is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level.

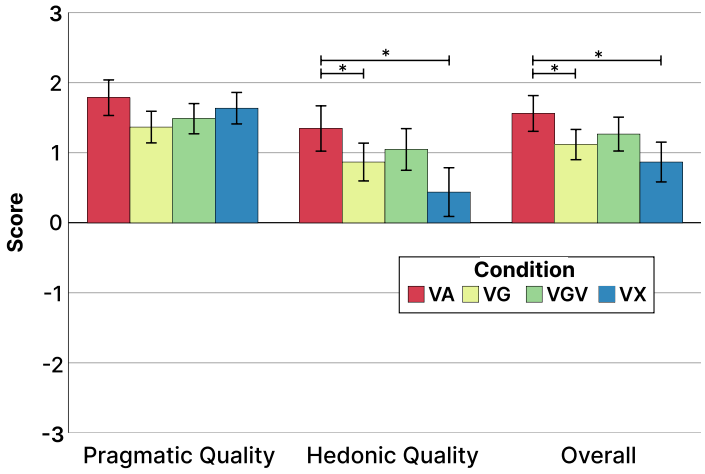


Figure 4: User experience scores assessed by the UEQ in the experiment. The bar chart shows the results for the three subscales pragmatic quality, hedonic quality, and overall for the four instructor conditions (VA, VG, VGV, VX). Higher is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level.

5 Discussion

Our results reveal a central tension in the design of embodied virtual instructors: the Personalization Paradox. We found that while participants overwhelmingly preferred a fully personalized instructor, these significant subjective benefits did not translate into greater task efficiency. This divergence between subjective experience and objective performance frames our discussion of the specific findings related to each dimension of personalization.

5.1 Importance of Close User Matching (VA)

Supporting **H1**, the virtual instructor with matched personality, voice, and gender (VA) outperformed other conditions on multiple subjective metrics. Higher ratings for hedonic quality and co-presence suggest that combining personality, voice, and gender matching creates a more engaging and immersive instructional experience. These findings are consistent with theories of social identity and similarity-attraction, which propose that perceived

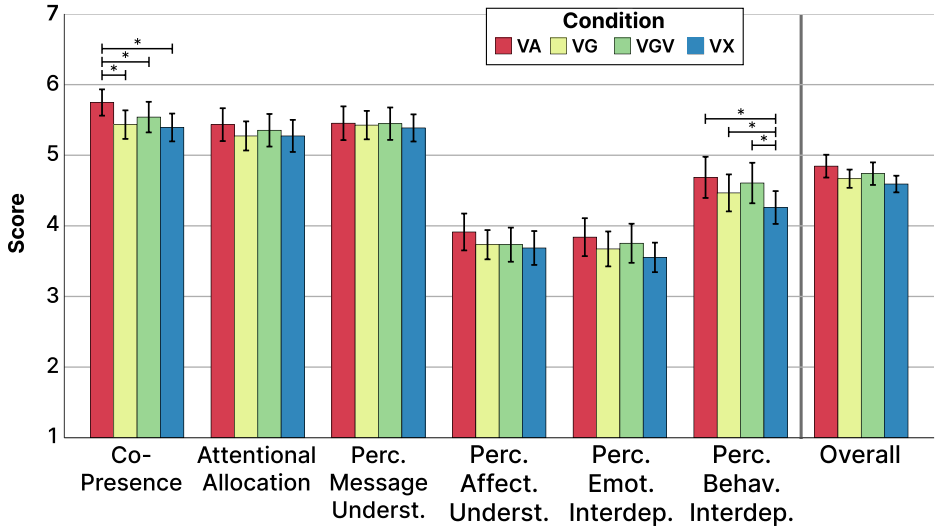


Figure 5: Results for the Harms-Biocca Social Presence questionnaire. Bar chart showing the results for the six subscales Co-Presence, Attentional Allocation, Perceived Message Understanding, Perceived Affective Understanding, Perceived Emotional Interdependence, and Perceived Behavioral Interdependence, as well as the Overall score, for the four instructor conditions (VA, VG, VGV, VX). Higher is better. The vertical error bars show the standard error. The horizontal bars indicate pairwise significance at the 5% significance level.

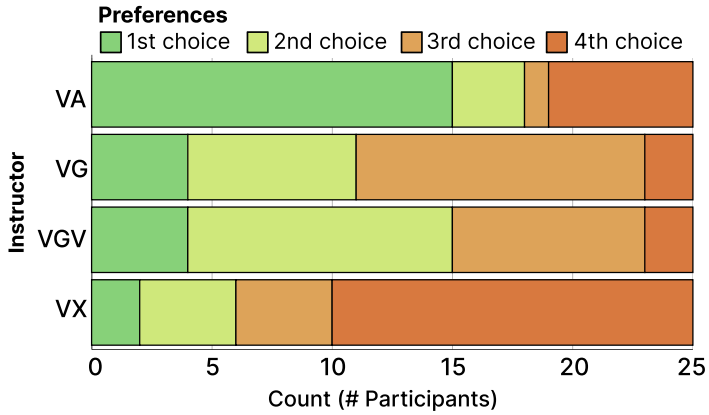


Figure 6: Subjective results for the Instructor Rankings. The stacked bar chart shows the numbers of participants who ranked the four instructor conditions (VA, VG, VGV, VX) in first to last place.

similarity enhances interpersonal trust and communication effectiveness [42, 46]. Because decoding settings and persona prompts were held constant across conditions (temperature=0; identical scaffolds), these differences reflect our intended manipulations rather than stochastic response drift by the underlying model [9, 43].

Interestingly, the subjective gains observed for VA did not translate into significantly faster task performance. This divergence highlights the core of the Personalization Paradox. An analysis of the conversation logs reveals that the richer, personalized dialogue from the VA instructor often led to more conversational elaboration. For example, when one participant made a mistake, the non-matched instructor responded with: *"Almost there! For this step, the correct spot is G4. Let's place it there and then we can move on to the next one."*

In contrast, an instructor embodying a participant's highly agreeable personality from a different session handled the same situation with albeit longer, interaction: *"Umm, okay, let's see... Ah, it looks like that last one is just a little off. No worries, that's an easy fix. How about we nudge that block over to G4 to get it perfect? Once that's locked in, we'll be golden for the next part."*

This richer interaction style, an emergent property of the personality simulation, likely explains both the higher subjective ratings and the longer task times for the VA condition. While full personalization can improve users' enjoyment, designers must balance this

personality-driven interaction with concise, goal-oriented instruction, especially in time-sensitive tasks.

5.2 Role of Gender Matching (VG)

Consistent with **H2**, participants completed tasks more quickly under the gender-matched condition (VG) than the non-matched condition (VX). This suggests that even minimal personalization—such as matching the instructor's gender—can reduce social distance and facilitate smoother communication.

However, VG did not significantly improve subjective measures of social presence or user experience. This may indicate that gender alone is a relatively weak cue in the absence of richer personalized signals like voice or personality traits. Users may require multiple overlapping signals of similarity to form strong social bonds with virtual agents.

This finding has practical implications: while gender matching is easy to implement, it may not suffice to create truly engaging virtual instruction experiences if used in isolation.

5.3 Limited Benefits of Voice Matching (VGV)

Contrary to H3, combining voice matching with gender matching (VGV) did not produce significant improvements over gender matching alone. This finding suggests a nuanced psychological trade-off, where the potential benefits of voice similarity [69] were likely counteracted by a vocal uncanny valley effect.

Even when a voice is cloned from a participant's sample, subtle artifacts in pitch, intonation, or cadence may make synthesized speech feel uncanny or slightly artificial. As the study by Ross et al. on synthesized voices found, listeners can perceive a voice as "eerie" even when it is not rated as fully human-like, suggesting a negative reaction can occur without perfect realism [67]. Participants might recognize their own voice yet simultaneously perceive it as unnatural, dampening the intended social presence benefits. Our open-ended feedback supports this interpretation: participants frequently described the cloned voice as both recognizable and unsettling. For instance, one participant remarked that "it freaked me out that only thirty seconds of audio was enough to copy my voice… I don't like listening to my own voice, so I found it a bit creepy," while another commented that "hearing my voice come out of the instructor was unsettling." Others described the experience as simultaneously familiar and artificial: "it sounded like me, but also… not fully me; there was something slightly off that made it feel uncanny." They further noted: "it was close, but the tone and pacing weren't exactly how I normally talk." These reactions illustrate how partial mismatches in prosody or emotion can trigger an uncanny response even when the overall voice identity is accurate.

Additionally, participants may not attach strong social significance to voice similarity unless it is paired with other consistent behavioral cues, like matching speech style, emotional tone, or conversational pacing. Further, consistency between vocal and facial features may contribute to these effects [12]. This highlights the need for future work to explore how multiple layers of personalization can synergize to strengthen the perception of agent authenticity.

5.4 Drawbacks of Non-Matched Instructors (VX)

Supporting **H4**, the non-matched instructor (VX) consistently performed worst across subjective ratings and user preferences. Participants rated VX significantly lower in perceived behavioral interdependence, social presence, and overall user experience. VX was also ranked lowest in preference rankings and was associated with the longest task durations.

This underscores the risk of deploying virtual instructors with generic, non-personalized designs, as such agents may be perceived as impersonal or detached. Users may interpret mismatches in gender, voice, or conversational style as signals of social distance, reducing trust and engagement [4].

Interestingly, VX's negative impact was evident not only in subjective measures but also in task performance, suggesting that the lack of social connection may translate into tangible performance costs in instructional contexts. This aligns with research indicating that affective and relational cues can influence cognitive effort and task focus [32].

5.5 Interplay Between User-Matching Dimensions

An important insight from our findings is that effects of user-matching are not merely additive. While personality matching alone drove significant subjective gains, voice matching did not enhance outcomes unless combined with other cues. This suggests that successful virtual instructor design may depend on achieving coherence across multiple layers of user-matching—voice, language style, and nonverbal behavior—to establish a cohesive social identity for an effective instructor.

Moreover, our results indicate that different user-matching dimensions serve different functions:

- Gender matching offers basic social comfort, potentially reducing barriers to task engagement.
- Personality matching drives perceived social connection and engagement but was also associated with longer task completion times, likely due to the richer, more elaborate dialogue it produced.
- Voice matching remains promising but requires further technical advances to achieve naturalness and emotional nuance.

These findings support a layered approach to virtual instructor design, where designers selectively deploy personalization features based on task context and desired outcomes.

5.6 On LLM Consistency and Reproducibility

A common threat to validity in LLM-mediated systems is output variability across sessions or turns. In our setup, this threat is minimized: We used GPT-4o with temperature=0, a fixed system/persona prompt with reverse-role constraints, and few-shot style anchors; the only prompt differences across conditions instantiated the intended manipulations (personality, voice, gender). These methodological choices, together with recent reports on stable expression of personality-linked linguistic markers and improved reproducibility under controlled decoding [9, 17, 43, 76], support the interpretation that our observed effects arise from attribute matching rather than model drift. Thus, we are confident that LLM inconsistency is not a confounding factor in our results.

5.7 Ethics of Embodied Agents

As LLM based technologies become further integrated into embodied forms, a heightened awareness is necessary when creating systems designed to establish trust. Prior works suggest over-reliance on LLM-based systems which offer guidance, neglecting errors or uncertainty in model responses. Likewise, research findings show that the robustness of this models is a direct contribution to this level of trust, an especially concerning area of interest in the field of embodied agent design, where physical presence, social cues, and human-like interactivity can amplify perceived competence of the system. Emerging works show the capabilities of system engineering to reduce errors in response generation through techniques such as retrieval-augmented generation, allowing for better fencing of agent answers, which can foster a relationship where false trust is a non-factor.

5.8 Future Directions: Spatial Integration and Physical Agency

Our findings suggest that the effects of personalization observed here are likely to generalize across the full spectrum of Mixed Reality technologies, including both immersive AR and VR headsets and spatial, non-wearable devices such as projection mapping and volumetric displays [6, 74]. While headsets offer ego-centric immersion, spatial displays provide distinct advantages for long-term instruction by eliminating physical encumbrance [23]. Regardless of the display medium, as virtual instructors evolve from passive observation to active physical agency (e.g., highlighting objects with spotlights or actuating smart tools), the coherence of their personality and voice will become increasingly critical for establishing user trust [34, 35].

5.9 Limitations

A limitation of our work is that it was conducted in a controlled laboratory environment using a single, relatively simple assembly task. While this controlled setup allowed us to isolate and examine specific factors, it limits the external validity of our findings. We position this study as an initial step toward understanding the effects of matched personalities, voice, and gender in AR-based instruction. Future research should extend these insights by evaluating a broader range of LLM architectures and applying them to more complex, diverse tasks and ecological settings. Longitudinal studies, where users interact with virtual instructors repeatedly over extended periods, would also provide additional evidence and help validate the generalizability of these findings in real-world scenarios. Finally, future iterations should prioritize a larger, balanced gender demographic to further generalize the congruence effects observed in our current sample.

6 Conclusion

In this paper, we showed that user-agent similarity with respect to embodied virtual instructors, especially through matched personalities, voice, and gender, can significantly enhance social presence, user experience, preference, and performance in MR-based instruction.

Our findings inform the design of virtual instructors through an understanding of how user interaction with these respective agents is influenced by matching their characteristics to individual users. By analyzing objective measures of performance, we showed benefits of gender-matched virtual instructors on task completion times. Further, subjective measures revealed that the more an instructor is matched to the individual user, such as through voice and personality, the more effective they are at increasing user perception of agent-based instruction systems.

However, these subjective gains do not automatically translate into improved task performance, highlighting a complex interplay between relational and cognitive dimensions of user-agent interaction. Minimal personalization, such as gender matching, provides modest benefits for task efficiency but limited subjective impact, while voice matching alone appears insufficient to drive significant gains. Our findings emphasize that effective virtual instructor design may require coherent, multi-layered personalization strategies and further technological improvements in emotional expressiveness and voice naturalness. As virtual instructors evolve from passive observation to active physical agency, the "Personalization Paradox" identified here offers a critical guideline: when agents gain the power to manipulate the physical world, their social coherence will be paramount in establishing the necessary trust and psychological safety.

Acknowledgments

This research was supported by the Office of Naval Research under Award No. N000142512245 (Dr. Peter Squire, Code 34), and partially by the U.S. Department of the Army (Ground Vehicles Systems Center) under Award No. 2670-201-2016671. We are also grateful for Dallas Kirkland's technical expertise in rigging and animating the virtual instructor character. 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] Robert K. Atkinson, Richard E. Mayer, and Mary Margaret Merrill. 2005. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemporary Educational Psychology* 30, 1 (Jan. 2005), 117–139. doi:10.1016/j.cedpsych.2004.07.001

[2] Amy L. Baylor. 2011. The design of motivational agents and avatars. *Educational Technology Research and Development* 59, 2 (April 2011), 291–300. doi:10.1007/s11423-011-9196-3

[3] Russell Beale and Chris Creed. 2009. Affective interaction: How emotional agents affect users. *International Journal of Human-Computer Studies* 67, 9 (Sept. 2009), 755–776. doi:10.1016/j.ijhcs.2009.05.001

[4] Timothy Bickmore and Justine Cassell. 2005. Social Dialongue with Embodied Conversational Agents. In *Advances in Natural Multimodal Dialogue Systems*, Jan C. J. van Kuppevelt, Laila Dybkjær, and Niels Ole Bernsen (Eds.). Springer Netherlands, Dordrecht, 23–54. doi:10.1007/1-4020-3933-6_2

[5] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2010. Maintaining Engagement in Long-Term Interventions with Relational Agents. *Applied Artificial Intelligence* 24, 6 (July 2010), 648–666. doi:10.1080/08839514.2010.492259

[6] Oliver Bimber and Ramesh Raskar. 2005. *Spatial augmented reality: merging real and virtual worlds.* CRC Press. doi:10.1201/b10624

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

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901.

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial

General Intelligence: Early experiments with GPT-4. doi:10.48550/arXiv.2303.12712 arXiv:2303.12712 [cs].

[10] Justine Cassell. 2001. Embodied Conversational Agents: Representation and Intelligence in User Interfaces. *AI Magazine* 22, 4 (Dec. 2001), 67. doi:10.1609/aimag.v22i4.1593

[11] Siyuan Chen, Pittawat Taveekitworachai, Yi Xia, Xiaoxu Li, Mustafa Can Gursesli, Antonio Lanata, Andrea Guazzini, and Ruck Thawonmas. 2025. Don't Do That! Reverse Role Prompting Helps Large Language Models Stay in Personality Traits. In *Interactive Storytelling*, John T. Murray and María Cecilia Reyes (Eds.). Springer Nature Switzerland, Cham, 101–114. doi:10.1007/978-3-031-78453-8_7

[12] Zubin Choudhary, Nahal Norouzi, Austin Erickson, Ryan Schubert, Gerd Bruder, and Gregory F. Welch. 2023. Exploring the Social Influence of Virtual Humans Unintentionally Conveying Conflicting Emotions. In *Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*. 1–10. doi:10.1109/VR55154.2023.00072

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

[14] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*. 323–325. doi:10.1109/ICALT58122.2023.00100 ISSN: 2161-377X.

[15] Daniela De Angeli, Fotos Frangoudes, Savvas Avraam, Kleanthis Neokleous, and Eamonn O'Neill. 2022. Towards Engaging Intangible Holographic Public Displays. In *2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies (IMET)*. 01–08. doi:10.1109/IMET54801.2022.9929788

[16] Fernanda De La Torre, Cathy Mengying Fang, Han Huang, Andrzej Banburski-Fahey, Judith Amores Fernandez, and Jaron Lanier. 2024. LLMR: Real-time Prompting of Interactive Worlds using Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 600, 22 pages. doi:10.1145/3613904.3642579

[17] Joost C. F. de Winter, Tom Driessen, and Dimitra Dodou. 2024. The use of ChatGPT for personality research: Administering questionnaires using generated personas. *Personality and Individual Differences* 228 (Oct. 2024), 112729. doi:10.1016/j.paid.2024.112729

[18] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. 2023. PaLM-E: An Embodied Multimodal Language Model. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. Article 340, 20 pages.

[19] ElevenLabs. 2025. ElevenLabs Text-to-Speech AI. https://elevenlabs.io Accessed: Jul. 10, 2025.

[20] Epic Games. 2025. MetaHuman Creator. https://www.unrealengine.com/en-US/metahuman Accessed: Jul. 10, 2025.

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

[22] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* 41 (2009), 1149–1160. doi:10.3758/BRM.41.4.1149

[23] Markus Funk, Thomas Kosch, and Albrecht Schmidt. 2016. Interactive worker assistance: comparing the effects of in-situ projection, head-mounted displays, tablet, and paper instructions. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*. 934–939. doi:10.1145/2971648.2971706

[24] Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the Big-Five personality domains. *Journal of Research in Personality* 37, 6 (Dec. 2003), 504–528. doi:10.1016/S0092-6566(03)00046-1

[25] Arthur C Graesser, Natalie Person, and Derek Harter. 2000. Teaching Tactics in AutoTutor. *International Journal of Artificial Intelligence in Education* 11 (2000), 1020–1029.

[26] Chad Harms and Frank Biocca. 2004. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. In *Annual International Presence Workshop*.

[27] Thomas Holz, Abraham G Campbell, Gregory MP O'Hare, John W Stafford, Alan Martin, and Mauro Dragone. 2011. Mira—mixed reality agents. *International journal of human-computer studies* 69, 4 (2011), 251–268. doi:10.1016/j.ijhcs.2010.10.001

[28] Yongyi Ji, Zhisheng Tang, and Mayank Kejriwal. 2024. Is persona enough for personality? Using ChatGPT to reconstruct an agent's latent personality from simple descriptions. In *ICML 2024 Workshop on LLMs and Cognition Poster*.

[29] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, and Yonghui Wu. 2018. Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. In *Advances in Neural Information Processing Systems*, Vol. 31. Curran Associates, Inc.

[30] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems* (2023), 10622–10643.

[31] W Lewis Johnson, Jeff W Rickel, and James C Lester. 2000. Animated Pedagogical Agents: Face-to-Face Interaction in Interactive Learning Environments. *International Journal of Artificial Intelligence in Education* 11 (2000), 47–78.

[32] Maurits Kaptein, Deonne Castaneda, Nicole Fernandez, and Clifford Nass. 2014. Extending the Similarity-Attraction Effect: The Effects of When-Similarity in Computer-Mediated Communication. *Journal of Computer-Mediated Communication* 19, 3 (April 2014), 342–357. doi:10.1111/jcc4.12049

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

[34] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F Welch. 2018. 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 *2018 IEEE international symposium on mixed and augmented reality (ISMAR)*. IEEE, 105–114. doi:10.1109/ISMAR.2018.00039

[35] Kangsoo Kim, Ryan Schubert, Jason Hochreiter, Gerd Bruder, and Gregory Welch. 2019. Blowing in the wind: Increasing social presence with a virtual human via environmental airflow interaction in mixed reality. *Computers & Graphics* 83 (2019), 23–32. doi:10.1016/j.cag.2019.06.006

[36] Kangsoo Kim, Ryan Schubert, and Greg Welch. 2016. Exploring the impact of environmental effects on social presence with a virtual human. In *International Conference on Intelligent Virtual Agents*. Springer, 470–474. doi:10.1007/978-3-319-47665-0_57

[37] Lisa E. Kim, Verena Jörg, and Robert M. Klassen. 2019. A Meta-Analysis of the Effects of Teacher Personality on Teacher Effectiveness and Burnout. *Educational Psychology Review* 31, 1 (March 2019), 163–195. doi:10.1007/s10648-018-9458-2

[38] Lisa E Kim and Carolyn MacCann. 2018. Instructor personality matters for student evaluations: Evidence from two subject areas at university. *British Journal of Educational Psychology* 88, 4 (2018), 584–605. doi:10.1111/bjep.12205

[39] Katerina Koleva, Maurizio Vergari, Tanja Kojic, Sebastian Moeller, and Jan-Niklas Voigt-Antons. 2024. Influence of Personality and Communication Behavior of a Conversational Agent on User Experience and Social Presence in Augmented Reality. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 929–930. doi:10.1109/VRW62533.2024.00261

[40] Nicole Krämer, Stefan Kopp, Christian Becker-Asano, and Nicole Sommer. 2013. Smile and the world will smile with you—The effects of a virtual agent's smile on users' evaluation and behavior. *International Journal of Human-Computer Studies* 71, 3 (2013), 335–349. doi:10.1016/j.ijhcs.2012.09.006

[41] Mareike Kunter, Uta Klusmann, Jürgen Baumert, Dirk Richter, Thamar Voss, and Axinja Hachfeld. 2013. Professional competence of teachers: effects on instructional quality and student development. *Journal of educational psychology* 105, 3 (2013), 805. doi:10.1037/a0032583

[42] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender? An experimental test of gender stereotype. In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*. ACM, 289–290. doi:10.1145/633292.633461

[43] Steven A. Lehr, Ketan S. Saichandran, Eddie Harmon-Jones, Nykko Vitali, and Mahzarin R. Banaji. 2025. Kernels of selfhood: GPT-4o shows humanlike patterns of cognitive dissonance moderated by free choice. *Proceedings of the National Academy of Sciences* 122, 20 (2025), e2501823122. doi:10.1073/pnas.2501823122

[44] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 1192–1202. doi:10.18653/v1/D16-1127

[45] Jionghao Lin, Zifei Han, Danielle R. Thomas, Ashish Gurung, Shivang Gupta, Vincent Aleven, and Kenneth R. Koedinger. 2025. How Can I Get It Right? Using GPT to Rephrase Incorrect Trainee Responses. *International Journal of Artificial Intelligence in Education* 35 (2025), 482–508. doi:10.1007/s40593-024-00408-y

[46] François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality Generation for Dialogue. In *Proceedings of the 45th Annual Meeting*. ACL, Prague, Czech Republic, 496–503.

[47] François Mairesse and Marilyn Walker. 2008. A Personality-based Framework for Utterance Generation in Dialogue Applications.. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior*. 80–87.

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

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

[50] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* E77-D, 12 (1994), 1321–1329.

[51] Abdul Mannan Mohammed, Martin McCarthy, Carsten Neumann, Gerd Bruder, Dirk Reiners, and Carolina Cruz-Neira. 2026. 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. doi:10.1145/3742413.3789163

[52] Abdul Mannan Mohammed, Martin McCarthy, Carsten Neumann, Gerd Bruder, Dirk Reiners, and Carolina Cruz-Neira. 2026. It's All in the Personality: A Comparative Study of Real, Ideal, and Customized Virtual Instructors for AR Assembly Tasks. *IEEE Transactions on Visualization and Computer Graphics* (2026).

[53] Abdul Mannan Mohammed, Azhar Ali Mohammad, Jason Ortiz, Carsten Neumann, Grace Bochenek, Dirk Reiners, and Carolina Cruz-Neira. 2025. A Human Digital Twin Architecture for Knowledge-based Interactions and Context-Aware Conversations. (04 2025). doi:10.48550/arXiv.2504.03147

[54] Roxana Moreno, Richard E. Mayer, Hiller A. Spires, and James C. Lester. 2001. The Case for Social Agency in Computer-Based Teaching. *Cognition and Instruction* 19, 2 (2001), 177–213. doi:10.1207/S1532690XCI1902_02

[55] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship.* MIT Press.

[56] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7, 3 (2001), 171–181. doi:10.1037/1076-898X.7.3.171

[57] Kristine L. Nowak and Christian Rauh. 2005. The influence of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *Journal of Computer-Mediated Communication* 11, 1 (Nov. 2005), 153–178. doi:10.1111/j.1083-6101.2006.tb00308.x

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

[59] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, Vol. 35. Curran Associates, Inc., 27730–27744.

[60] Gina Pancorbo, Mieke Decuyper, Lisa E. Kim, Jacob A. Laros, Loes Abrahams, and Filip De Fruyt. 2022. A teacher like me? Different approaches to examining personality similarity between teachers and students. *European Journal of Personality* 36, 5 (Sept. 2022), 771–786. doi:10.1177/08902070211015583

[61] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3586183.3606763

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

[63] Polycam, Inc. 2025. Polycam Photogrammetry Tool. https://poly.cam Accessed: Jul. 10, 2025.

[64] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. 2019. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP 2019*. 3617–3621. doi:10.1109/ICASSP.2019.8683143

[65] Bryon Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people.* Cambridge University Press.

[66] Jens Reinhardt, Luca Hillen, and Katrin Wolf. 2020. Embedding conversational agents into ar: Invisible or with a realistic human body?. In *Proceedings of the Fourteenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, New York, NY, USA, 299–310. doi:10.1145/3374920.3374956

[67] Alice Ross, Martin Corley, and Catherine Lai. 2024. Is there an uncanny valley for speech?: Investigating listeners' evaluations of realistic synthesised voices. In *Speech Prosody 2024*. International Speech Communication Association (ISCA), 1115–1119. doi:10.21437/SpeechProsody.2024-225

[68] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2014. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In *Design, User Experience, and Usability: Theories, Methods, and Tools for Designing the User Experience*. Springer International Publishing, Cham, 383–392. doi:10.1007/978-3-319-07668-3_37

[69] Juliana Schroeder and Nicholas Epley. 2016. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology. General* 145, 11 (Nov. 2016), 1427–1437. doi:10.1037/xge0000214

[70] Sakib Shahriar, Brady D. Lund, Nishith Reddy Mannuru, Muhammad Arbab Arshad, Kadhim Hayawi, Ravi Varma Kumar Bevara, Aashrith Mannuru, and Laiba Batool. 2024. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. *Applied Sciences* 14, 17 (Jan. 2024), 7782. doi:10.3390/app14177782

[71] Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM, New York, NY, USA, 78–84. doi:10.1145/964442.964458

[72] Betty Tärning, Annika Silvervarg, Agneta Gulz, and Magnus Haake. 2019. Instructing a Teachable Agent with Low or High Self-Efficacy. *International Journal of Artificial Intelligence in Education* 29, 1 (2019), 89–121. doi:10.1007/s40593-018-0167-2

[73] Jai Prakash Verma and Abdel-Salam G. Abdel-Salam. 2019. *Testing Statistical Assumptions in Research*. Wiley. doi:10.1002/9781119528388

[74] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring virtual agents for augmented reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–12. doi:10.1145/3290605.3300511

[75] Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology* 17 (Jan. 2023), 18344909231213958. doi:10.1177/18344909231213958

[76] Yilei Wang, Jiabao Zhao, Deniz S. Ones, Liang He, and Xin Xu. 2025. Evaluating the ability of large language models to emulate personality. *Scientific Reports* 15, 1 (Jan. 2025), 519. doi:10.1038/s41598-024-84109-5

[77] Hawon Yoo, Jaehong Jang, Hyunju Oh, and Innwoo Park. 2022. The potentials and trends of holography in education: A scoping review. *Computers & Education* 186 (2022), 104533. doi:10.1016/j.compedu.2022.104533

[78] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 2204–2213. doi:10.18653/v1/P18-1205