ARIA: Toward Human-Centered Embodied AI Instruction in Real-Time Augmented Reality

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

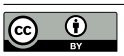
Advances in artificial intelligence and embodied interaction in augmented reality (AR) are creating new opportunities for intelligent instructional systems that dynamically adapt to individual learners. However, sustaining real-time responsiveness while preserving natural, socially meaningful interaction remains a persistent challenge. This work introduces ARIA (Augmented Reality Instructional Agent), a real-time software architecture for embodied AI instruction in augmented reality. ARIA leverages large language models (LLMs) with adaptive prompt engineering to tailor dialogue style, instructional strategy, and persona expression to each user. Its modular pipeline is optimized for robust, low-latency performance, with benchmarks reported for responsiveness and system stability. To complement the technical evaluation, user experience was assessed through standardized questionnaires, offering insights into perceived personalization, trust, and interaction quality. Quantitative and qualitative results demonstrate that ARIA achieves sub-second responsiveness, high pragmatic and hedonic usability, and a strong sense of co-presence and instructional trust. This work contributes a unified framework and reference architecture for developing adaptive embodied agents that combine technical efficiency with human-centered design, highlighting how real-time responsiveness can serve as the foundation for relational engagement in embodied AI instruction.

CCS Concepts

• **Human-centered computing** → **Interaction paradigms**; **Interaction design**; *User interface management systems*; • **Computing methodologies** → *Intelligent agents.*

Keywords

Augmented Reality, Embodied AI, Intelligent User Interfaces, Prompt Personalization, Large Language Models (LLMs), Adaptive Instruction, Human-Centered AI, Multimodal Interaction, User Experience



IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/2026/03
https://doi.org/10.1145/3742413.3789163

ACM Reference Format:
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), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3742413.3789163

1 Introduction

Advancements in the development of large language models (LLMs) and multi-modal AI systems are enabling opportunities in the field of virtual instruction. As models become more robust, embodied instructors powered by these agents are more capable of exhibiting natural, embodied, and context-aware behaviors [5]. In particular, these instructional agents are gaining traction as an interactive solution to current issues within the field of educational technology.

However, a significant gap remains in reconciling high-fidelity embodiment with the stringent latency demands of natural conversation in augmented reality (AR) [27, 39]. As detailed in Section 2, cumulative delays across multimodal pipelines—integrating audio-visual processing and synchronized avatar embodiment—often exceed perceptually acceptable turn-taking thresholds. Furthermore, long, multi-turn instruction risks behavioral drift and reduced affective consistency, often described as "emotion dilution". Prior systems frequently balance real-time responsiveness against social realism; ARIA addresses this by providing a modular architecture designed to preserve both temporal realism and identity continuity.

To address these challenges, we present the Augmented Reality Instructional Agent (ARIA) architecture, a real-time modular software pipeline designed for implementing high-fidelity, low-latency human-agent interaction in AR spaces.

Our system is designed with two primary motives: establishing a high-performance testbed for studying human-AI instructional interaction and proposing a generalized framework designed to support controlled alterations to embodied instructors such as personality, voice, and gender. The ARIA architecture provides five modular layers, defined in detail within Section 3, optimized for extensibility and minimal latency. A Python controller manages these layers, handling data flow in real-time to exchange information between the multimodal inputs and the virtual instructor.

Beyond the introduction of this framework, we provide an evaluation of ARIA through quantitative benchmarking and user assessment through a user study. We also include analysis of key metrics

such as Time to First Audio, LLM Inference Time, and Streaming Overlap in both verbal-only and multi-modal configurations of ARIA.

To evaluate the effectiveness of the ARIA architecture as a platform for real-time embodied instruction, this work addresses the following research questions:

- **RQ1**: To what extent can a modular AR architecture utilize selective multimodal triggering and perceptual masking to achieve the sub-second responsiveness necessary for natural conversational flow?
- **RQ2**: How does the integration of stable, persona-driven prompting influence a user's pragmatic success and their relational trust in an automated instructor?
- **RQ3**: How does the combination of high-fidelity visual embodiment and task-aligned audio-visual cues foster a sense of shared physical-digital presence in AR-based instruction?

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

2 Related Work

2.1 Non-Embodied Virtual Instructors

State–of–the–art LLMs are powering chatbot based applications to provide real-time feedback, mentorship, and personalized instruction. Studies on the effectiveness of these instructors have revealed that LLM-powered chatbot tutors allow students to learn a greater amount of information in significantly faster time frames by reducing the cognitive effort of learners [27]. Ongoing research suggests that text-based instructor interfaces allow for efficiency beyond time to mastery, allowing for high scalability due to the large corpora of knowledge accessible by modern LLM systems [58]. Where these systems in many cases outperform the current state–of–the–art embodied instructional agents in latency, they lack the ability to engage the sense of social presence, a key factor in educational performance [6, 11]. These issues have resulted in systems accepting trade-offs of slower agent response timing for establishing agents which convey crucial social cues such as empathy and encouragement.

2.2 Embodied Virtual Instructors

As new research paves the way for higher fidelity models, embodied agents have become increasingly implemented due to unique advantages in fostering engagement, comprehension, and social presence. In particular, the integration of multimodal LLMs for achieving dynamic and context-aware instruction is becoming increasingly explored. Multimodal designs which leverage embodied GPT based models have demonstrated synchronized verbal and nonverbal cues to significantly increase the sense of trust established between user and instructor [26, 41]. Embodied instruction systems which offer visual content improve student capabilities to acquire knowledge and levels of engagement [53]. Adding speech recognition, perception, and expressive animation in real-time delivers experiences which surpass verbal or text-based feedback [26]. Visual cues beyond those related to the educational content are also a key factor for embodied instructors in AR environments, utilizing nonverbal communication to foster co-presence which enhances motivation and attention in learners [28]. In addition, frameworks that implement embodiment in an AR space have shown to be an improvement to traditional full virtual reality immersion in regards to situational awareness and physical engagement [51]. While virtual instructors can create high quality educational dialogue at scale, systems which implement embodiment facilitate social presence, conversational perception, and empathy, driving realistic instructional communication [15]. This is supported by Mohammed et al., who found that when designing an embodied virtual instructor, key factors in social settings, such as gender and personality, directly influence the impact on user engagement [36–38].

2.3 Latency and Conversational Realism in Embodied AI

A key challenge in embodied AI system design is achieving low-latency interaction loops. The sequential pipeline of speech recognition, LLM inference, and speech synthesis introduces cumulative delays that can disrupt the rapid, 200–300 ms turn-taking rhythm characteristic of natural human conversation [35, 50], although short delays (approximately 500–1000 ms) remain perceptually acceptable in continuous dialogue [25]. This constraint places strong real-time demands on both model inference and audio streaming subsystems. In multimodal configurations that incorporate vision, even optimized pipelines can incur additional perceptible latency due to frame capture and encoding overheads.

Beyond raw timing, latency directly affects user trust and perceived system competence. Empirical studies in AR and embodied agent contexts show that response delays increase cognitive load and degrade engagement, particularly in time-sensitive or instructional settings [13, 39]. To mitigate this, recent work has explored perceptual masking strategies that reduce the *felt* latency without reducing actual processing time. Behavioral fillers, such as brief gaze aversions, subtle breathing animations, or short "thinking" gestures, can occupy the interval between user speech and agent response. Similarly, verbal fillers like "hmm..." can be inserted at natural interruption points, a time-management strategy that masks the 1.5 to 2 seconds often required for the full response to be prepared [21]. These techniques help preserve conversational flow and maintain a sense of continuous presence [18, 29, 34, 49]. Such mechanisms are now considered essential elements of low-latency embodied systems, functioning as perceptual buffers that complement hardware and model-level optimization.

2.4 LLM and Persona Consistency

While latency defines temporal realism [21, 50], persona stability determines the continuity of an agent's identity over time [7, 52]. For embodied AI instructors, maintaining a coherent personality and behavioral tone across interactions is a key system-level requirement. Recent advances in multimodal LLMs, such as GPT-4o, demonstrate emerging capabilities for personality emulation and context retention, enabling role-specific discourse that resembles

human-like instructional behavior [30, 48]. From a systems perspective, these advances motivate architectures that treat persona definition, memory, and state tracking as core design components rather than surface-level prompt configurations.

However, ensuring consistent persona expression across multi-turn interactions remains a non-trivial problem. As conversational depth increases, models can exhibit behavioral drift, where tone and character deviate from the intended profile [52]. This challenge is amplified in multimodal agents, where a lack of affect consistency across modalities like voice, face, and gesture can lead to "emotion dilution," significantly disrupting the user's perception of the agent's emotional state [7]. To address this, researchers have proposed structured personality conditioning, in which psychological frameworks such as the Big Five [33] are operationalized into stable, machine-readable constraints [23], while other work has utilized sociological frameworks like Affect Control Theory (ACT) to moderate LLM responses for emotional congruence [31]. These methods serve as system control mechanisms that guide model behavior and support reproducible persona expression across sessions [17]. In educational or interactive contexts, this stability is as critical as low latency, since it governs not only how an agent responds but also how it sustains a coherent and trustworthy instructional identity.

2.5 Synthesis and Research Gap

The literature reveals a recurring systems trade-off: instructional agents achieve low latency with limited embodiment, or sustain rich persona expression at the cost of real-time responsiveness. ARIA is proposed as an architectural baseline to reconcile these constraints by (1) modular pipeline orchestration for parallelized multimodal perception and streaming dialogue, (2) structured persona conditioning to reduce behavioral drift across turns, and (3) event-driven vision capture to avoid continuous video encoding overhead. Prior work establishes that trust and presence are strengthened through consistent persona sources and latency-mitigation cues; ARIA situates these mechanisms at the architecture level rather than surface prompt design, providing a reference testbed for future AR/VR and 2D embodied-agent comparisons.

3 Methods

3.1 System Architecture Overview

ARIA is designed as a modular, real-time pipeline to enable natural, embodied, and personalized interactions between a human user and an AI instructor in AR. The system's design prioritizes low latency, high fidelity, and the flexibility required for systematic HCI research. The architecture is composed of five distinct layers, managed by a central Python controller that orchestrates the flow of data from multimodal user input to synchronized avatar response. Figure 1(a) illustrates the complete data flow through these layers.

3.1.1 User Interaction Layer. The foundation of ARIA is the physical environment where the user interacts with the virtual instructor. This layer encompasses the hardware that captures user actions and presents the agent's embodiment.

- **Physical Setup**: The user and virtual instructor are positioned on opposite sides of a large-format (1.04 m × 2.05 m) transparent screen. The screen's nano-optic film overlay allows an ultra-short-throw Optoma EH340UST projector to render a high-contrast, holographic-style image of the instructor while maintaining the user's view of the physical world. This setup provides a strong sense of co-presence, as the agent appears to occupy the space behind the screen.
- **Multimodal Input**: User input is captured through two primary channels. A HyperX QuadCast S microphone captures high-quality audio for speech input, while a top-mounted 4K webcam (Depstech DW50) provides a clear view of the user's workspace, enabling visual task assessment.
- **Embodied Output**: The system's response is delivered multimodally. The virtual instructor's speech is rendered through Logitech Z150 speakers, while their visual form is rendered as a high-fidelity MetaHuman character in Unreal Engine 5.5 [14].

3.1.2 Input Processing Layer. This layer is responsible for converting raw sensor data into a structured format suitable for the AI's cognitive processing. A key design principle here is efficiency, minimizing unnecessary data processing to maintain low latency.

- **Real-time Speech Transcription**: User speech is continuously streamed to Azure's Speech-to-Text API [10], which provides a low-latency transcription.
- **Selective Visual Processing**: Instead of processing a continuous video stream, which is computationally expensive and slow, our system uses a conversationally-grounded trigger for image capture [9]. The transcribed text is parsed for phrases indicative of a user seeking verification (e.g., "verify," "check," "is this right?"). Only upon detecting such a cue does the system use OpenCV [43] to capture a single, high-resolution frame from the webcam. This event-driven approach dramatically reduces processing overhead and aligns with natural human-human communication patterns where visual checks are performed on request.

3.1.3 Prompt Management Layer (AIManager). The AIManager is the core intelligent component of ARIA's architecture. It is a custom Python module responsible for dynamically constructing a context-rich, multimodal prompt for the LLM on each conversational turn. This layer is what enables the system's deep personalization capabilities. The final prompt sent to GPT-4o [42] integrates several streams of information:

- **Task Instructions:** A base prompt containing the rules and steps for the assembly task.
- **Conversation History:** A rolling window of the last several conversational turns to maintain context.
- **Current User Input:** The latest transcribed utterance from the user.
- **Visual Data:** The base64-encoded image, if one was captured during the input processing stage.
- **Personality and Persona Configuration:** This is the most important aspect of the AIManager.

To ensure consistent and believable persona embodiment for the virtual instructor, we implement a multi-part prompt engineering strategy:

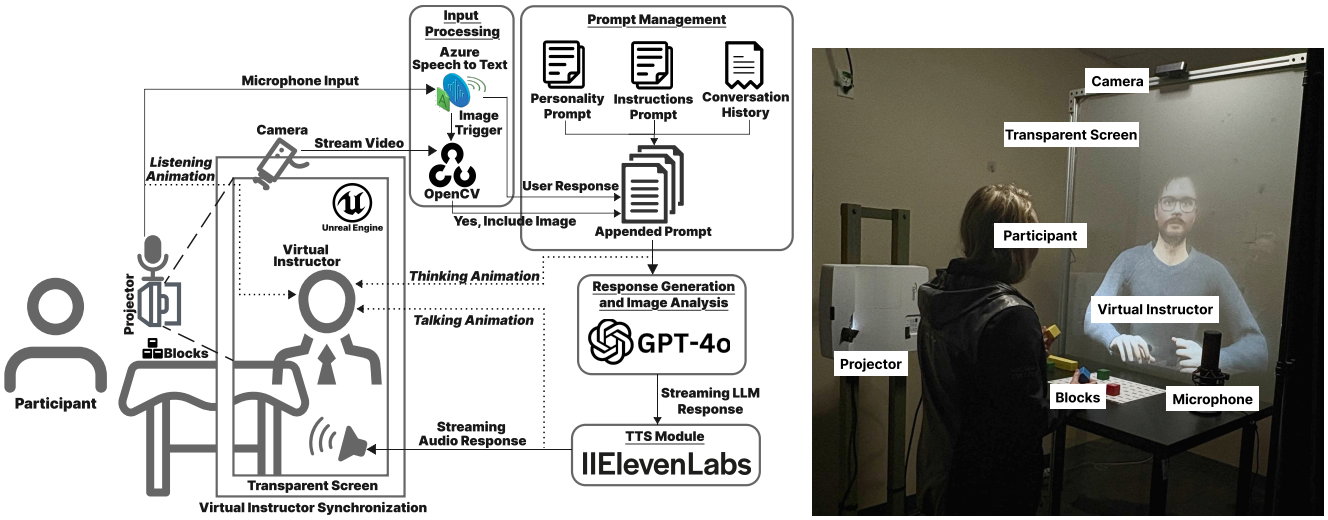


Figure 1: (a) System architecture of the ARIA platform. The diagram presents a modular pipeline spanning user interaction, multimodal input processing, prompt construction, AI-driven response generation, and synchronized virtual instructor feedback. This design enables context-aware, personality-driven guidance using visual and verbal cues. (b) Annotated photo showing the system setup.

– **Trait Definition:** The prompt defines the target Big Five personality using both its quantitative score (e.g., "Conscientiousness: 104/120") and qualitative descriptors (e.g., "highly organized, diligent, and self-disciplined") [24, 32]. This is informed by the PERSONALITY PROMPTING (P^2) method [23]. To achieve conversational mirroring, we employed a few-shot prompting technique [4, 48], using a short speech transcription to learn and adopt the user's specific filler words and speech patterns.
– **Behavioral Constraints (Reverse Role Prompting):** To prevent persona drift, the model is explicitly instructed on behaviors to avoid [8]. For example, an agent with "Low Neuroticism" is instructed not to use language that is anxious, insecure, or overly emotional.
– **Deterministic Output:** We set the LLM's temperature parameter to 0 to minimize creative variance and ensure the agent's responses are consistently aligned with its defined persona, a standard technique for personality evaluation in LLMs [22, 52].

To validate the persona's fidelity, we re-administered the IPIP-NEO-120 instrument to ARIA across five separate sessions. The persona implementation was considered stable and accurate if ARIA's scores consistently remained within ±5 points of the desired personality profile on each of the Big Five domains.

3.1.4 Response Generation and Image Analysis Layer. This layer offloads the core cognitive and perceptual tasks to powerful, pre-trained models.

• **Multimodal Understanding:** The fully constructed prompt is sent to the GPT-4o API [42]. The model's advanced multimodal capabilities allow it to concurrently process the user's language, analyze the visual state of the assembly task from the provided image, and consult the agent's defined persona to formulate a context-aware, personality-driven response.

• **Voice Synthesis:** The textual response generated by GPT-4o is immediately streamed to ElevenLabs' Flash v2.5 TTS API [12]. This service synthesizes the text into natural-sounding speech, using a voice profile that was pre-cloned for a real-world instructor.

3.1.5 Virtual Instructor Synchronization Layer. The final layer is responsible for bringing the agent to life by synchronizing its visual embodiment with its synthesized voice.

• **Backend-Frontend Communication:** The central Python controller communicates with the Unreal Engine 5.5 frontend via a persistent WebSocket connection. This allows for low-latency, bidirectional messaging.
• **Animation State Machine:** The controller commands a simple state machine within Unreal Engine to trigger the appropriate animations on the MetaHuman avatar. When the system detects user speech, it sends a "listen" command. While waiting for the LLM and TTS responses, a "thinking" animation is played to mask the system's processing time and signal cognitive effort. If the measured *Time to First Audio* (TTFA) exceeds the typical human turn-taking threshold of roughly 0.5–1 seconds [25, 49], a short "thinking" filler sound is also played to maintain conversational flow and manage user expectations. Finally, as the TTS audio stream is played, a "speak" command triggers a synchronized talking animation. Upon completion, the avatar returns to an "idle" state. This synchronization creates a believable and continuous presence, making the interaction feel natural rather than turn-based.

The virtual instructor's persona and embodiment were modeled after the available instructor for this setup to ensure maximum realism by matching the avatar's traits exactly with the real-world source's voice and personality profile. This approach prioritized "persona-voice-visual" congruence, reducing the risk of identity drift or "emotion dilution" during system validation. While gender

can influence user presence in XR [54], our choice focused on providing a consistent, realistic instructional source for this baseline architectural testbed.

3.2 System Evaluation Procedure

To evaluate the performance and user perception of our ARIA, we employed a dual-pronged strategy: (1) direct benchmarking of system responsiveness under controlled conditions, and (2) user evaluation in which we deployed the system in instructional scenarios.

3.2.1 Benchmarking Procedure. To evaluate system responsiveness, we conducted controlled benchmarking tests under two experimental configurations: (1) a verbal-only condition using the speech pipeline and (2) a multimodal condition that additionally included camera-based visual processing. Benchmarks were conducted locally on a workstation running a Windows 11 Pro machine (Intel Core i7-8750 CPU, 32 GB RAM, NVIDIA RTX 4070 GPU) under a stable network connection.

We measured the end-to-end response delay from the completion of user speech input to the onset of synthesized agent audio, defined as *Time to First Audio* (TTFA). Supporting metrics included the following:

- **Time to First Audio (TTFA):** latency from the end of user speech to the onset of synthesized agent audio.
- **Time to First Token (TTFT):** latency from the end of user speech to the generation of the first output token by the LLM.
- **LLM Inference Time:** total duration of the model's response generation, from prompt submission to the last token produced.
- **Prompt Processing Time:** time required by the controller to assemble the multimodal prompt and dispatch it to the LLM API.
- **Streaming Overlap:** degree of parallelization between text generation and text-to-speech; negative values indicate audio synthesis begins before text generation completes.
- **Camera Latency:** time from detection of a visual-check trigger to the availability of the captured frame to the prompt builder (capture + transfer + encoding).
- **Tokens per Second (TPS):** sustained token-generation throughput during streaming, computed as total tokens divided by elapsed time from first to last token.
- **Animation Latency:** delay between the initiation of a speech or behavioral command and the corresponding avatar animation onset in Unreal Engine; this reflects synchronization precision between backend and visual embodiment through websocket connection.

Each condition was measured across 20 complete interaction cycles, and the best runs per configuration were averaged for analysis. This benchmarking procedure enabled quantitative comparison of responsiveness across configurations and informed optimization of multimodal latency.

The primary contribution of this benchmarking procedure is the establishment of a **performance baseline** for LLM-driven AR agents. By quantifying these sub-second delays, we demonstrate that ARIA's modular architecture successfully remains within the 1.5-second human-perceptual window required for natural turn-taking [21]. This proves that high-fidelity embodiment and real-time responsiveness are technically compatible, providing a reference for future developers to evaluate the efficiency.

3.2.2 User Evaluation. We evaluated the ARIA system with a focus on the virtual instructor's effectiveness and interaction quality.

Participants. We recruited 51 participants (19 female, 31 male, 1 non-binary) 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 25.6. All of them had normal or corrected-to-normal vision.

Task Procedure. Participants were engaged in interactive sessions with ARIA, assembling magnetic block structures while conversing with the virtual instructor. The goal was interaction quality rather than task completion, so participants were encouraged to engage in brief casual dialogue or small talk with ARIA between steps. Each build lasted on average *7–10 minutes*, a duration consistent with prior work on social presence and trust formation in short human–agent interactions [1–3, 40]. Four distinct target figures were used across trials, providing multiple opportunities for users to interact with the system and become familiar with its conversational behavior. Each figure comprised 28 uniformly sized square magnetic blocks of comparable spatial complexity, see Figure 2. During each session, ARIA verbally guided participants through precise block placements on a predefined grid. Crucially, ARIA's instructional style was grounded in the personality of a real-world instructor. We used the IPIP-NEO-120 instrument [24] to create a Big Five trait profile, which revealed Low Neuroticism (69/120), Moderate Conscientiousness (73/120) and Extraversion (81/120), and High Openness (96/120) and Agreeableness (103/120). This detailed profile was then embedded into the prompt structure discussed in Section 3.1.3 to align the agent's dialogue style with that of a human persona.

User Experience Questionnaire (UEQ). Participants completed the short version of the UEQ [47] after interacting with the instructor to assess hedonic and pragmatic qualities of the interaction (7-point scale from -3 to 3; higher is better).

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

The single-item probes in Table 1 were selected to target well-isolated perceptual dimensions while minimizing participant fatigue—a critical consideration in hands-busy AR tasks. This follows established psychometric evidence that single-item measures can be as effective as long-form scales in educational and psychological research [16, 46]. Furthermore, our use of 7-point Likert probes (see Table 1) aligns with current standardized practices for assessing real-time perceptions in AR-based systems testbeds [19].

Harms-Biocca Social Presence Questionnaire. To assess participants' perception of social presence with the instructor, we employed the 36-item version of the Harms-Biocca Social Presence

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

Question	Scale
How much do you TRUST the instructions provided?	1 (Not Trustworthy at All) .. 7 (Very Trustworthy)
How much did you RELY on the instructions provided?	1 (Didn't Rely) .. 7 (Fully Relied)
How CONFIDENT were you in the accuracy and usefulness of the instructions provided?	1 (No Confidence at All) .. 7 (Extremely Confident)
How ADVANTAGEOUS do you believe the instructions provided were compared to what you could have achieved on your own?	1 (Disadvantageous) .. 7 (Advantageous)
How SATISFIED are you with the outcome of the task?	1 (Dissatisfied) .. 7 (Satisfied)
Rate the CLARITY of the instructions provided	1 (Unclear) .. 7 (Clear)
How ENJOYABLE were the instructions provided?	1 (Not Enjoyable at All) .. 7 (Extremely Enjoyable)
How EXCITING was the experience following the instructions?	1 (Not Exciting at All) .. 7 (Very Exciting)
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 QUICKLY do you think most people would learn to use this system?	1 (Very Slowly) .. 7 (Very Quickly)

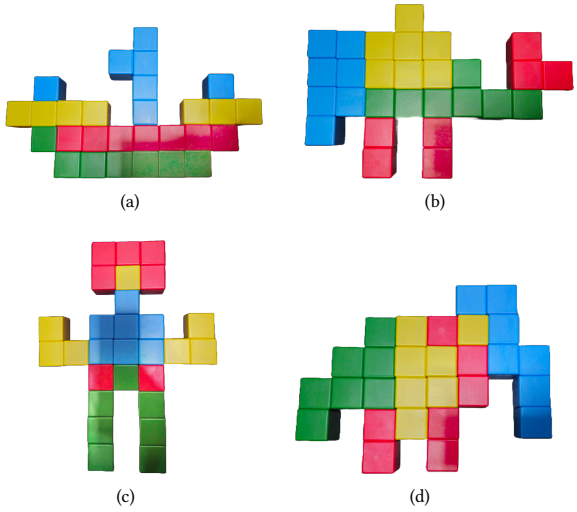


Figure 2: The four magnetic block structures (a)–(d) that participants assembled during the study. Each figure featured comparable spatial complexity and served to facilitate natural interaction with the virtual instructor.

questionnaire [20]. 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 interaction quality with each instructor.

Qualitative Feedback. In addition to these questionnaires, participants were invited to provide open-ended verbal feedback on their experience with ARIA, including comments on timing, clarity, and the instructor's demeanor. These qualitative responses were recorded and later analyzed to complement the quantitative findings.

4 Results

4.1 Benchmarking for Real-Time Responsiveness and Latency

Following the benchmarking procedure described in Section 3.2.1, we evaluated system performance under two configurations: a verbal-only condition and a multimodal condition that included camera-based visual analysis. As shown in Table 2, the verbal-only setup achieved a mean Time to First Audio (TTFA) of 0.84 s (SD = 0.21 s), delivering sub-second responses that preserve natural conversational flow. When multimodal processing was enabled, the mean TTFA increased by 0.58 s to 1.42 s (SD,=,0.29 s). This additional latency was primarily driven by the multimodal model's visual analysis..

Supporting metrics showed similar efficiency across the processing pipeline. The Time to First Token (TTFT) averaged 0.35 s (SD = 0.08 s) for verbal and 0.95 s (SD = 0.22 s) for multimodal exchanges, while LLM inference times averaged 0.75 s (SD = 0.15 s) and 1.6 s (SD = 0.38 s), respectively. The system sustained high token-generation throughput (68 vs. 37 tokens per second) and began audio synthesis before completion of text generation (negative streaming overlap), further reducing perceived delay. Animation latency was minimal, averaging only 0.012 s (SD = 0.004 s), which is effectively instantaneous from the user's perspective. The very low variance further indicates that the WebSocket-based synchronization between the backend and Unreal Engine frontend was highly stable and reliable.

Even with the inclusion of vision, total latencies remained within conversationally acceptable limits. While human turn-taking averages 200–300 ms [35, 50], prior work shows that delays up to 1–1.5 s remain conversationally natural aided by its use of both nonverbal "thinking" animations and verbal fillers to mitigate perceived latency [21, 29, 34, 49]. Our observed TTFA of 0.84 s (verbal) and 1.42 s (multimodal) thus aligns closely with this perceptual window, maintaining engagement through dynamic "thinking" and "listening" animations.

Table 2: Comparison of Real-Time Latency Metrics With and Without Camera Input

Metric	Without Camera		With Camera	
	Mean (s)	SD (s)	Mean (s)	SD (s)
TTFA (Time-to-first-audio)	0.84	0.21	1.42	0.29
TTFT (Time-to-first-token)	0.35	0.08	0.95	0.22
LLM Inference	0.75	0.15	1.60	0.38
Prompt Processing	0.09	0.02	0.12	0.03
Streaming Overlap	−0.21	0.12	−0.30	0.17
Camera Latency	—	—	0.03	0.01
TPS (Tokens-per-second)	68.0	9.5	37.0	8.5
Animation Latency	0.012 (SD = 0.004) s			

Note. TTFA and TTFT measure perceived responsiveness; negative overlap values indicate parallel speech synthesis during token streaming. All values represent averages of the best runs per configuration.

4.2 User Evaluation Results

Using the procedure described in Section 3.2.2, we assessed participants' perception of the ARIA system, virtual instructor effectiveness, and overall interaction quality.

User Experience Questionnaire. Figure 3(a) shows our results for the UEQ. The two sub-scales of the UEQ indicate the Pragmatic Quality (M = 1.90, SD = 1.06) and Hedonic Quality (M = 1.52, SD = 1.24) of participants' experience with the system. Following the recommended interpretation of the results according to the Data Analysis Tool (UEQ-S) [1], these values indicate an *excellent* pragmatic quality and a *good* hedonic quality. The pragmatic quality in this context indicates that the system was perceived as highly usable and functional, and the hedonic quality indicates that it further had a high emotional and experiential appeal.

Single-Item Questionnaires. Further, Figure 3(b) shows our results for the Single-Item Questionnaires, which captured participants' impressions of the system's usability, see Table 1.

Participants consistently rated the system highly in terms of Trust and Reliance, indicating that participants felt like they could place trust in the system's guidance and relied on it during the task. The high scores for Confidence and Advantageous further indicate that participants felt confident in the accuracy and usefulness of the instruction, meaning that they felt that the system added value and supported their success. The results for Satisfaction and Clarity further indicate that participants were pleased with their results and found the instructions easy to follow. Moreover, the results for Enjoyable, Exciting, and Personable are in line with the hedonic qualities as seen for the UEQ results above. Last but not least, participants rated the system as highly Engaging and believed that others could Quickly learn the system, indicating that it is intuitive and easy to adopt.

Harms-Biocca Social Presence Questionnaire. Our results for the Harms-Biocca Social Presence Questionnaire [20] are shown in Figure 4. The sub-scales indicate very high results for Co-Presence (M = 5.9, SD = 0.9), Attentional Allocation (M = 5.7, SD = 1.0), and Perceived Message Understanding (M = 5.8, SD = 1.0). These scores suggest that participants felt strongly aware of the virtual instructor's presence, remained focused during the interaction, and perceived the communication as clear and comprehensible. High Co-Presence reflects a sense of "being together" with the agent in the shared AR space, while elevated Attentional Allocation indicates that participants were cognitively engaged and not distracted. The strong score for Perceived Message Understanding further confirms that the system's verbal and visual cues were effective in conveying instructional content.

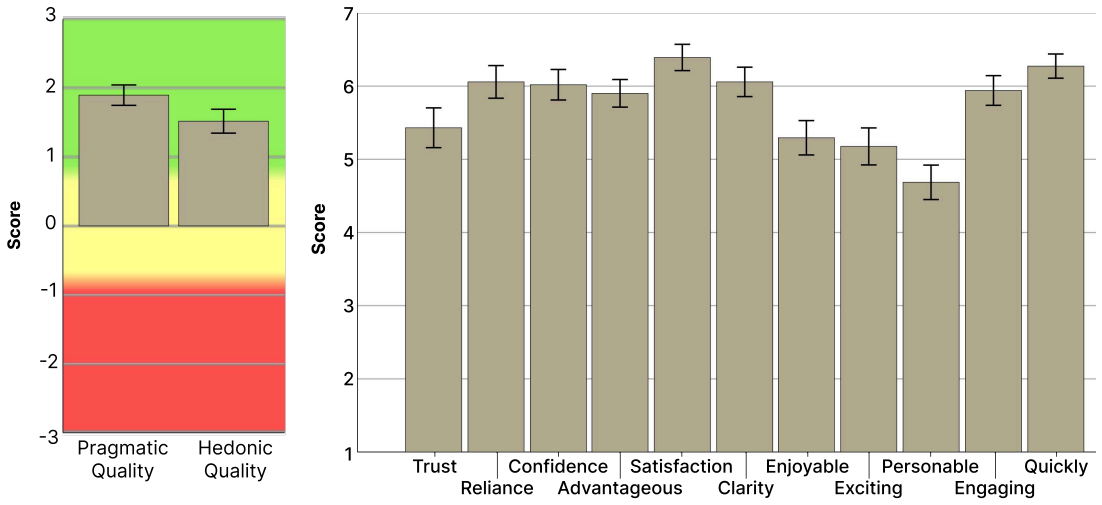
The sub-scales further indicate lower but still reasonable results for Perceived Affective Understanding (M = 4.5, SD = 1.3), Perceived Emotional Interdependence (M = 4.3, SD = 1.4), and Perceived Behavioral Interdependence (M = 4.9, SD = 1.3). These dimensions reflect more subtle aspects of social presence, such as the agent's ability to recognize and respond to emotional states, and the extent to which participants felt their actions influenced the agent's behavior. While these scores are lower, they remain above the midpoint of the scale, indicating that the system conveyed a moderate level of emotional and behavioral responsiveness. This may be attributed to the prioritization of clarity and consistency in this system over nuanced emotional expression.

The overall scores indicate high social presence (M = 5.2, SD = 0.8), suggesting that the system successfully created a believable and engaging interaction experience. Taken together, these results demonstrate that ARIA was perceived as socially present and communicatively competent, with particularly strong performance in task-relevant dimensions of presence and engagement.

5 Discussion

Our findings indicate that ARIA achieves a critical balance for embodied AI: the technical architecture delivers real-time responsiveness that, in turn, fosters a sense of social presence and instructional trust, resulting in an experience users found to be both highly functional and enjoyable. This is evidenced by two converging streams of data: strong objective performance benchmarks that place ARIA within the perceptual window of human dialogue, and very positive subjective user ratings across multiple perceptual scales. This section deconstructs these results, exploring how the system's sub-second latency provided the foundation for a positive

[1] https://www.ueq-online.org



(a) (b)

Figure 3: (a) Bar chart showing the results for the User Experience Questionnaire with the two scales Pragmatic Quality and Hedonic Quality. (b) Results for the Single-Item Questionnaires, with the eleven scales (see Table 1). Higher is better. The vertical error bars show the standard error. Responses were screened for outliers and inconsistent submissions (e.g., straight-lining or rapid completion); no exclusions were required, confirming the high internal consistency of the participant responses.

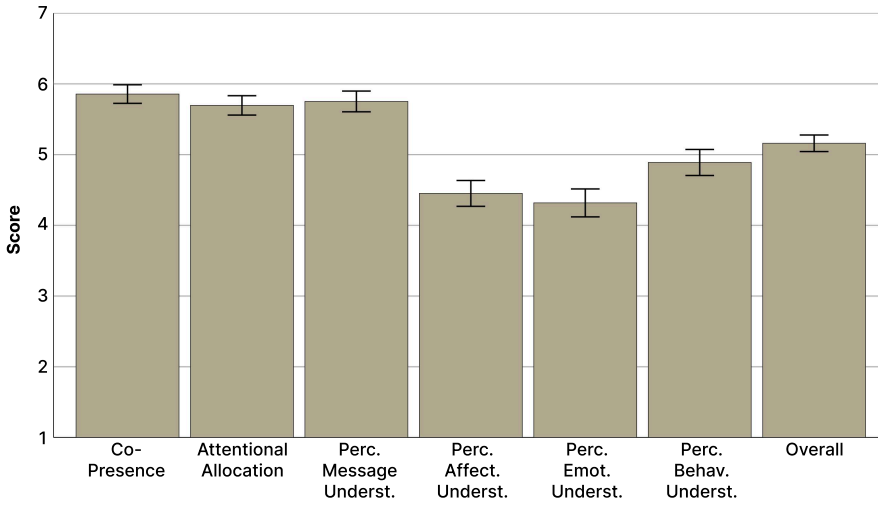


Figure 4: Results for the Harms-Biocca Social Presence questionnaire. Bar chart showing the results for the six sub-scales Co-Presence, Attentional Allocation, Perceived Message Understanding, Perceived Affective Understanding, Perceived Emotional Interdependence, Perceived Behavioral Interdependence, and Overall score. Higher is better. The vertical error bars show the standard error.

user experience, how its persona consistency built relational trust, and how its embodiment in AR created a sense of a shared space with the user.

5.1 The Primacy of Speed: How Real-Time Responsiveness Fosters Social Connection

In response to **RQ1**, our findings confirm that a cornerstone of effective human-agent interaction is the synergy between low latency and perceived social presence, and our results indicate ARIA successfully achieves this. In its verbal-only configuration, the system delivered sub-second responses, achieving a mean Time to First Audio (TTFA) of just 0.84 s. This level of responsiveness is highly conducive to preserving a natural conversational flow. Even when multimodal visual processing was enabled, the mean TTFA of 1.42 s remained squarely within the 1.5-second window that prior work [21] has established as conversationally natural when mitigated by responsive conversational cues. To further minimize latency, ARIA's visual pipeline uses phrase spotting, capturing images only

when users explicitly seek verification (e.g., "is this right?" or "check this"). This selective triggering avoids continuous video streaming, reducing computational load while keeping the interaction tightly coupled to conversational intent.

This perceived responsiveness is not merely a product of raw processing speed but is strategically enhanced by key architectural choices designed to preserve conversational rhythm. Crucially, any processing delay was perceptually masked by the agent's seamless "thinking" and "listening" animations, a form of the behavioral fillers and perceptual masking strategies proposed in prior work to preserve conversational flow [18, 29, 34]. This was a recurring theme in user feedback, with one participant noting, "It was impressive how quickly it responded. The 'thinking' animation made the pause feel natural." This effect was further amplified by the use of negative streaming overlap, which reduces felt latency by beginning audio synthesis before the full response is generated. This synergy of technical optimizations translated directly into a fluid interaction. Multiple users commented on the *"fast response time,"* with their remarks indicating they perceived the system not as a series of queries, but as a *"continuous and smooth conversational exchange."*

5.2 From Functional Success to Relational Trust

When addressing **RQ2**, we find that ARIA's responsiveness provided the foundation for an experience that users found both effective and enjoyable. The "excellent" Pragmatic Quality (M=1.90) from the User Experience Questionnaire confirms the ARIA's functional success as a tool, while its "good" Hedonic Quality (M=1.52) shows that this efficiency was also perceived as engaging.

The single-item questionnaires provide granular insight into this positive perception. The system's pragmatic success is underscored by high scores for Clarity and Satisfaction, reinforced by strong ratings for Confidence and how Advantageous the system was. This suggests users felt the agent added significant value, with many sharing sentiments like, *"The instructions were super clear, and I felt confident I was doing it right."*

Our evaluation focuses on trust, confidence, and reliance as indicators of interaction acceptability rather than relative instructional improvement. Prior HCI work shows that perceived trust in embodied conversational agents is shaped by explainability and transparency of guidance [45], pedagogical framing [57], and trustworthy embodiment in task-based settings [44]. Additional research highlights that agent familiarity, perceived knowledgeability, and depth of knowledge are critical determinants of perceived competence and instructional reliability [55, 56]. We therefore interpret ARIA's persona-stability design as aligning with these established factors for instructor-identity continuity, motivating future work to evaluate these architectural choices against explicit comparative baselines.

Beyond mere functionality, the equally high scores for Trust and Reliance point to a deeper relational dynamic, reflecting a common sentiment that *"you could really trust its instructions."* This trust appears to emerge from the agent's predictability. The AIManager's use of persona-driven prompting ensured a stable personality, addressing the common challenge of behavioral drift in long-form conversations [52] and allowing users to form a reliable mental model of the agent. The positive impact of this was a recurring theme in user feedback. As one participant shared, *"I was surprised that the praise from the agent actually made me feel more confident and engaged. When it was positive and supportive, it really helped my mood."* Another highlighted how the agent's specific traits directly impacted their experience, stating, *"The personality of the agent definitely made a difference. When the instructor was more upbeat and supportive, I found it easier to focus and the task felt less like a chore."* These comments suggest that a consistent and well-defined persona is a critical component for building instructional trust.

These comments suggest that a consistent persona is a critical component for building instructional trust, which in turn fostered high Engagement. Finally, participants believed others could Quickly learn the system, confirming its intuitive design and ease of adoption.

5.3 Embodied Co-Presence: The Experience of "Being There" Together

Beyond usability and trust, the ultimate goal of an embodied agent is to create a sense of shared presence, a quality that is critical for educational performance and a key advantage over non-embodied tutors [6, 11]. On this basis, we find that the Harms-Biocca results show ARIA was highly successful in this regard, addressing **RQ3**. The exceptionally high scores for Co-Presence (M=5.9), Attentional Allocation (M=5.7), and Perceived Message Understanding (M=5.8) suggest that users perceived the agent not merely as an image on a screen, but as a true partner occupying their physical space. The agent was seen as "always attentive," creating a powerful sense of being seen and heard. This was achieved not just through visual embodiment, but through subtle audio cues that reinforced its attentiveness. As one participant explained, *"The little sounds made a big difference. When I was talking, it would give a little 'mhmm,' and before it answered, there was sometimes a soft 'hmm' sound. It made it feel like it was actively listening and then actually considering my question, not just processing data."*

For some, this presence offered unique advantages over a human instructor, with one explaining, *"It was nice having an instructor that was always patient. You don't have to worry about being judged or asking a dumb question."* This highlights the potential for embodied agents to create psychologically safe learning environments.

At the same time, the results also clearly define the current boundaries of ARIA's social capabilities. The more moderate scores for Perceived Affective Understanding (M=4.5) and Perceived Emotional Interdependence (M=4.3) suggest that while users felt the agent was present, they did not feel it was emotionally attuned to them. This distinction was also noted in qualitative feedback, with one user stating, *"The instructor was incredibly helpful and clear, but it didn't feel like it understood me. It was focused on the task, not on how I was feeling about it."* This is likely an outcome of our system's intentional prioritization of instructional clarity and persona consistency over nuanced emotional expression. This highlights a key difference between achieving cognitive presence and affective presence. This gap reflects a well-known challenge in the field, where systems often trade nuanced social cues like empathy for faster performance [15], pointing to a rich area for

future work in developing agents that are not only present but also perceptive.

5.4 Implications for Design and Deployment

The architectural validation of ARIA highlights three takeaways for future embodied instructional agent design in AR/VR/XR environments:

- **Temporal–Social Alignment:** Modular pipelines and lightweight behavioral fillers (e.g., listening or thinking animations) can help preserve perceived conversational timing realism [21]. These mechanisms offer a design lever to align system delays with human expectations without requiring further reductions in model inference time.
- **Instructor Identity Continuity:** Grounding persona stability using a real, verified instructor source supports reproducible instructor expression across interaction turns. This reduces the risks of behavioral drift [52] and ensures the "trustworthy embodiment" necessary for task-based reliability [44].
- **Selective, Efficient Perception:** Event-triggered vision capture based on linguistic verification intent offers a promising direction for lowering compute and power overhead. This approach is particularly relevant for mobile AR headsets with thermal constraints and provides an inherent privacy-by-design framework by limiting continuous sensor access.

ARIA is positioned as a high-performance architectural baseline and testbed, establishing a benchmark for future comparative studies across 2D, VR, and AR embodied instructional interfaces.

5.5 Limitations and Future Work

While this work validates ARIA's core architecture, it is important to acknowledge the boundaries that define its current scope and offer avenues for future research. First, the system was evaluated in a controlled laboratory setting with a structured assembly task, which, while necessary for establishing a baseline, may not fully reflect the complexities of real-world instructional scenarios. Furthermore, our evaluation was based on a single instructor persona; future studies should explore a wider range of personalities to assess how different learners respond to diverse teaching styles.

This study did not include a comparative baseline, such as a non-embodied agent, a 2D interface, or alternative XR modalities like a Virtual Reality (VR) head-mounted display or a non-holographic AR screen. As our goal was to validate ARIA as a high-performance architectural testbed rather than compare instructional interfaces, we acknowledge this bounds interpretation of results to feasibility and pipeline acceptability. This documents a performance benchmark for the community while motivating future work with explicit comparative baselines.

While ARIA was perceived as attentive, the capacity for genuine empathy, conveyed through both dialogue and nuanced expression, remains a key frontier for future work, marking the transition from a proficient instructor to a truly perceptive partner.

6 Conclusion

In this paper, we introduced ARIA (Augmented Reality Instructional Agent), a real-time architecture for embodied AI instruction that combines low-latency processing, adaptive persona prompting, and multimodal interaction within AR. The findings demonstrate that responsiveness is not just a technical performance benchmark, but a crucial social cue that enables participants to experience the agent as competent, trustworthy, and present. User feedback highlights that ARIA is a system capable of dynamic interactions which increase co-presence and establish trust in the embodied agent.

Our evaluation demonstrates the influence of design choice on the level of engagement users have with the system. By contributing a framework, we introduce a method for designing virtual instructors that establish credibility through an enhanced sense of co-presence. Our evaluation advances current research on designing embodied agents, and the limitations of the current architecture that we have established pave the way for future research into engaging virtual instructor scenarios. Likewise, introducing an architecture which handles real-time user to agent communication in a modular fashion proposes the ability for stakeholders to easily design and integrate embodied instructors. ARIA's capability to provide sub-second verbal responsiveness and maintains near-real-time performance even with multimodal data allows for a greater focus on design choices, reducing future development time of AR-based embodied instructor applications, where speed and stability are necessary components to establish credibility with the instructor.

GenAI Usage Disclosure

We used Microsoft Copilot (University-affiliated) to assist with text refinement and language polishing during manuscript preparation. No generative AI tools were used for data collection, analysis, or experimental evaluation described in this work. Artificial intelligence models, including Large Language Models (LLMs), are an integral component of the ARIA system investigated in this work. However, generative AI tools were not used for the primary generation of the system's source code. The authors take full responsibility for the integrity and accuracy of the final content.

Acknowledgments

This material includes work supported in part by the Office of Naval Research under Award No. N000142512245 (Dr. Peter Squire, Code 34). This work was also partially supported by the U.S. Department of the Army (Ground Vehicles Systems Center) under Award No. 2670-201-2016671. We thank Dallas Kirkland for her valuable contributions in rigging and animating the virtual instructor character used in this work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research or the U.S. Department of the Army.

References

[1] Lesley A Albertson. 1980. Review essay: Trying to eat an elephant the social psychology of telecommunications, by John Short, Ederyn Williams, and Bruce Christie. London: John Wiley, 1976. *Communication Research* 7, 3 (1980), 387–400. doi:10.1177/009365028000700307

[2] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 701–706. doi:10.1109/ROMAN.2008.4600749

[3] Frank Biocca, Chad Harms, and Judee K Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators & virtual environments* 12, 5 (2003), 456–480. doi:10.1162/105474603322761270

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. , Article 159 (2020), 25 pages.

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

[6] Daniela Castellanos-Reyes, Jennifer C. Richardson, and Yukiko Maeda. 2024. The evolution of social presence: A longitudinal exploration of the effect of online students' peer-interactions using social network analysis. *The Internet and Higher Education* 61 (2024), 100939. doi:10.1016/j.iheduc.2024.100939

[7] Che-Jui Chang, Samuel S Sohn, Sen Zhang, Rajath Jayashankar, Muhammad Usman, and Mubbasir Kapadia. 2023. The importance of multimodal emotion conditioning and affect consistency for embodied conversational agents. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 790–801. doi:10.1145/3581641.3584045

[8] 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

[9] Herbert H Clark and Edward F Schaefer. 1989. Contributing to discourse. *Cognitive science* 13, 2 (1989), 259–294. doi:10.1207/s15516709cog1302_7

[10] Microsoft Corporation. 2025. Azure AI Speech Service. https://azure.microsoft.com/en-us/products/ai-services/ai-speech. Accessed: Oct. 1, 2025.

[11] Matthew A. d'Alessio, Loraine L. Lundquist, Joshua J. Schwartz, Vicki A. Pedone, James Pavia, and J. Fleck. 2019. Social presence enhances student performance in an online geology course but depends on instructor facilitation. *Journal of Geoscience Education* 67, 3 (2019), 222–236. doi:10.1080/10899995.2019.1580179

[12] ElevenLabs. 2025. ElevenLabs Text-to-Speech AI. https://elevenlabs.io/. Accessed: Oct. 1, 2025.

[13] Morad Elfleet and Mathieu Chollet. 2024. Investigating the Impact of Multimodal Feedback on User-Perceived Latency and Immersion with LLM-Powered Embodied Conversational Agents in Virtual Reality. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*. 1–9. doi:10.1145/3652988.3673965

[14] Epic Games. 2024. MetaHuman Creator. https://www.unrealengine.com/en-US/metahuman. Accessed: Oct. 1, 2025.

[15] Silvia García-Méndez, Francisco de Arriba-Pérez, and María del Carmen Somoza-López. 2025. A Review on the Use of Large Language Models as Virtual Tutors. *Science & Education* 34 (2025), 877–892. doi:10.1007/s11191-024-00530-2

[16] Katarzyna Gogol, Martin Brunner, Thomas Goetz, Romain Martin, Sonja Ugen, Ulrich Keller, Antoine Fischbach, and Franzis Preckel. 2014. "My Questionnaire is Too Long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology* 39, 3 (2014), 188–205. doi:10.1016/j.cedpsych.2014.04.002

[17] Liuying Gong, Jingyuan Chen, and Fei Wu. 2025. Is ChatGPT a Competent Teacher? Systematic Evaluation of Large Language Models on the Competency Model. *IEEE Transactions on Learning Technologies* (2025). doi:10.1109/TLT.2025.3564177

[18] Denmar Mojan Gonzales, Snehanjali Kalamkar, Sophie Jörg, and Jens Grubert. 2025. Behavioral and Symbolic Fillers as Delay Mitigation for Embodied Conversational Agents in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics* (2025). doi:10.1109/TVCG.2025.3616865

[19] Matt Gottsacker, Hiroshi Furuya, Zubin Datta Choudhary, Austin Erickson, Ryan Schubert, Gerd Bruder, Michael P. Browne, and Gregory F. Welch. 2024. Investigating the relationships between user behaviors and tracking factors on task performance and trust in augmented reality. *Computers & Graphics* 123 (2024), 104035. doi:10.1016/j.cag.2024.104035

[20] Chad Harms and Frank Biocca. 2004. Internal Consistency and Reliability of the Networked Minds Measure of Social Presence. In *Annual International Presence Workshop*. 246–251.

[21] Derek Jacoby, Tianyi Zhang, Aanchan Mohan, and Yvonne Coady. 2024. Human Latency Conversational Turns for Spoken Avatar Systems. (2024). doi:10.48550/arXiv.2404.16053

[22] 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. doi:10.48550/arXiv.2406.12216

[23] Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and Inducing Personality in Pre-trained Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 10622–10643.

[24] John A. Johnson. 2014. 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 (2014), 78–89. doi:10.1016/j.jrp.2014.05.003

[25] Simon Christophe Jolibois, Akinori Ito, and Takashi Nose. 2025. The Development of an Emotional Embodied Conversational Agent and the Evaluation of the Effect of Response Delay on User Impression. *Applied Sciences* 15, 8 (2025). doi:10.3390/app15084256

[26] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* 103 (2023), 102274. doi:10.1016/j.lindif.2023.102274

[27] Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. 2025. AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports* 15, 1 (2025), 17458. doi:10.1038/s41598-025-97652-6

[28] Daehwan Kim and Dongsik Jo. 2022. Effects on Co-Presence of a Virtual Human: A Comparison of Display and Interaction Types. *Electronics* 11, 3 (2022). doi:10.3390/electronics11030367

[29] Junyeong Kum and Myungho Lee. 2022. Can Gestural Filler Reduce User-Perceived Latency in Conversation with Digital Humans? *Applied Sciences* 12, 21 (2022). doi:10.3390/app122110972

[30] 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

[31] Evdoxia Eirini Lithoxoidou, George Eleftherakis, Konstantinos Votis, and Tony Prescott. 2025. Advancing affective intelligence in virtual agents using affect control theory. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 127–136. doi:10.1145/3708359.3712079

[32] François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality Generation for Dialogue. In *Proceedings of the 45th Annual Meeting*. ACL, Prague, Czech Republic, 496–503. https://aclanthology.org/P07-1063/

[33] 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

[34] Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, Ryan P. McMahan, and Joseph J. LaViola Jr. 2025. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces (CUI '25)*. Association for Computing Machinery, New York, NY, USA, Article 49, 15 pages. doi:10.1145/3719160.3736636

[35] Antje S Meyer. 2023. Timing in conversation. *Journal of Cognition* 6, 1 (2023), 20. doi:10.5334/joc.268

[36] 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).

[37] 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 Embodied AR Instructors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, Barcelona, Spain. doi:10.1145/3772318.3791902

[38] 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

[39] Mahdi Nabiyouni, Siroberto Scerbo, Doug A. Bowman, and Tobias Höllerer. 2017. Relative Effects of Real-world and Virtual-World Latency on an Augmented Reality Training Task: An AR Simulation Experiment. *Frontiers in ICT* Volume 3 - 2016 (2017). doi:10.3389/fict.2016.00034

[40] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. doi:10.1111/0022-4537.00153

[41] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, Austin Erickson, Zubin Choudhary, Yifan Li, and Gregory Welch. 2020. A Systematic Literature Review of Embodied Augmented Reality Agents in Head-Mounted Display Environments. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE 2020)*. The Eurographics Association, Virtual Conference. doi:10.2312/EGVE.20201234

[42] OpenAI. 2024. GPT-4o – Fast, intelligent, flexible GPT model. https://platform.openai.com/docs/models/gpt-4o. Accessed: Oct. 1, 2025.
[43] OpenCV Contributors. 2025. OpenCV: Open Source Computer Vision Library. https://opencv.org/. Accessed: Oct. 1, 2025.
[44] David A. Robb, José Lopes, Muneeb I. Ahmad, Peter E. McKenna, Xingkun Liu, Katrin Lohan, and Helen Hastie. 2023. Seeing eye to eye: trustworthy embodiment for task-based conversational agents. *Frontiers in Robotics and AI* Volume 10 - 2023 (2023). doi:10.3389/frobt.2023.1234767
[45] N. Sautchuk-Patricio and P. Henning. 2024. Addressing Trust Concerns in Educational Environments: Developing an Explainable Embodied Conversational Agent. In *EDULEARN24 Proceedings* (Palma, Spain) *(16th International Conference on Education and New Learning Technologies)*. IATED, 3033–3039. doi:10.21125/edulearn.2024.0805
[46] Henk Schmidt. 2018. The single-item questionnaire. *Health Professions Education* 4, 1 (2018), 1–2. doi:10.1016/j.hpe.2018.02.001
[47] 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, 383–392. doi:10.1007/978-3-319-07668-3_37
[48] 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
[49] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178. doi:10.1016/j.csl.2020.101178
[50] Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences* 119, 4 (2022), e2116915119. doi:10.1073/pnas.2116915119
[51] Ozlem Topsakal and Eray Topsakal. 2022. Framework for a foreign language teaching software for children utilizing AR, Voicebots and ChatGPT (large language models). *The Journal of Cognitive Systems* 7 (2022), 33–38. doi:10.52876/jcs.1227392
[52] 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
[53] Kanta Yamaoka, Ko Watanabe, Koichi Kise, Andreas Dengel, and Shoya Ishimaru. 2022. Experience is the Best Teacher: Personalized Vocabulary Building within the Context of Instagram Posts and Sentences from GPT-3. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 313–316. doi:10.1145/3544793.3560382
[54] Fu-Chia Yang, Pedro Acevedo, Siqi Guo, Minsoo Choi, and Christos Mousas. 2025. Embodied Conversational Agents in Extended Reality: A Systematic Review. *IEEE Access* 13 (2025), 79805–79824. doi:10.1109/ACCESS.2025.3566698
[55] Fu-Chia Yang, Kevin Duque, and Christos Mousas. 2024. The Effects of Depth of Knowledge of a Virtual Agent. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (2024), 7140–7151. doi:10.1109/TVCG.2024.3456148
[56] Fu-Chia Yang, Siqi Guo, and Christos Mousas. 2025. Exploring familiarity and knowledgeability in conversational virtual agents. *ACM Transactions on Applied Perception* 23, 1 (2025), 1–28. doi:10.1145/3757062
[57] Habeeb Yusuf, Arthur Money, and Damon Daylamani-Zad. 2025. Pedagogical AI conversational agents in higher education: a conceptual framework and survey of the state of the art. *Educational technology research and development* 73 (2025), 815–874. doi:10.1007/s11423-025-10447-4
[58] Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. 2025. A Comprehensive Review of AI-based Intelligent Tutoring Systems: Applications and Challenges. arXiv:2507.18882 [cs.IR] doi:10.48550/arXiv.2507.18882