frontiers | Frontiers in Virtual Reality

TYPE Technology and Code
PUBLISHED 05 June 2025
DOI 10.3389/frvir.2025.1555173

OPEN ACCESS

EDITED BY
David Swapp,
University College London, United Kingdom

REVIEWED BY
Peter Fromberger,
University Medical Center Göttingen, Germany
Mario Lorenz,
Chemnitz University of Technology, Germany
Daniel Hernández García,
Heriot-Watt University, United Kingdom

*CORRESPONDENCE
Alon Shoa,
Shoa.Alon@Gmail.com
Doron Friedman,
doronf@runi.ac.il

RECEIVED 03 January 2025
ACCEPTED 25 April 2025
PUBLISHED 05 June 2025

CITATION
Shoa A and Friedman D (2025) Milo: an LLM-based virtual human open-source platform for extended reality.
Front. Virtual Real. 6:1555173.
doi: 10.3389/frvir.2025.1555173

Milo: an LLM-based virtual human open-source platform for extended reality

Alon Shoa* and Doron Friedman*

Advanced Reality Lab (ARL), School of Communications, Reichman University, Herzliya, Israel

Large language models (LLMs) have made dramatic advancements in recent years, allowing for a new generation of dialogue agents. This allows for new types of social experiences with virtual humans, in both virtual and augmented reality. In this paper, we introduce an open-source system specifically designed for implementing LLM-based virtual humans within extended reality (XR) environments. Our system integrates into XR platforms, providing a robust framework for the creation and management of interactive virtual agents. We detail the design and architecture of the system and showcase the system's versatility through various scenarios. In addition to a straightforward single-agent setup, we demonstrate how an LLM-based virtual human can attend a multi-user virtual reality (VR) meeting, enhance a VR self-talk session, and take part in an augmented reality (AR) live event. We provide lessons learned, with focus on the possibilities for human intervention during live events. We provide the system as open-source, inviting collaboration and innovation within the community, paving the way for new types of social experiences.

KEYWORDS

virtual human, virtual agent, virtual reality, large language model, open-source, XR

1 Introduction

In recent years, the fields of deep neural networks (DNNs) and large language models (LLMs) have witnessed groundbreaking advancements. We suggest that these innovations pave the way for the development of a new generation of virtual agents, and in this paper we focus on the enhanced ability to engage in natural dialogue. Specifically, our research focuses on leveraging LLMs to facilitate rich, immersive interactions within virtual reality (VR) and mixed reality environments. This integration promises not only to enhance the user experience by providing more lifelike and responsive interactions but also to expand the practical applications of virtual agents in everyday scenarios.

We present our system – Milo – an LLM-based conversational agent, that integrates into XR scenes in Unity. We present several use-cases that were explored using Milo as the virtual agent. Milo can be configured to handle different types of behaviors, and can be easily extended to support new use-cases. We provide the system as open-source[1], including a tutorial guide for Unity developers, including non-programmers.

Our contribution includes describing the design and implementation of the system as well as making it available to the research community. Additionally, we describe several

1 https://github.com/Advanced-Reality-Lab/Milo

scenarios using the system, including a dyadic interactions in both VR and mixed reality, multi-user VR, and enhancing self-talk in VR. We present lessons learned from two live events and how a human operator can contribute to such experiences.

2 Background

Considerable research has been conducted on dialogue systems, encompassing text-based dialogue, audio-based dialogue, and virtual agents (Patlan et al., 2021). Developing animated virtual agents capable of engaging in dialogue presents a multitude of challenges, including speech understanding, natural language dialogue management, speech generation, turn-taking, and the synchronization of verbal and non-verbal communication channels. These challenges have been extensively addressed in the literature for several decades; some early attempts include (Cassell, 2001; McNeill, 1992). The integration of immersive extended reality (XR), encompassing both virtual reality (VR) and augmented reality (AR), introduces additional layers of complexity, although many of the core challenges remain fundamentally similar. In VR, the focus often shifts toward achieving behavioral realism (Slater, 2009), including the nuanced use of proxemics and eye gaze, as opposed to the emphasis on photo-realism often pursued as a goal in non-immersive agents. In addition to fully automated virtual humans, alternative configurations that blend human operator input with software automation are also viable, allowing for hybrid control of avatars (e.g., see (Kishore et al., 2016; Friedman and Hasler, 2016)).

A variety of approaches have been employed to manage dialogue interactions in VR. Experimental setups frequently adopt the wizard-of-Oz methodology (Bradley et al., 2009), where the virtual human operates with a predefined set of responses, and the decisions regarding what to say and when are entirely controlled by a human operator (e.g., (Maulsby et al, 1993)). This approach allows for the study of human interaction dynamics without the complexity of fully autonomous dialogue systems, providing a controlled environment for research (e.g., (Nakash et al., 2022)).

Another viable approach is structured dialogue. This method typically separates natural language understanding (NLU) from response generation. The NLU component leverages state-of-the-art machine learning techniques to parse user utterances into distinct intents. Subsequently, a dedicated system component is tasked with deciding the appropriate responses, often utilizing *ad hoc* mappings from intent templates to intent responses. For instance, Traum et al. (Traum et al., 2015) employed this method with pre-recorded responses for reconstructing holocaust survivor testimonies, with the primary challenge being the development of an NLU system capable of mapping arbitrary questions to the most pertinent responses.

LLMs open new opportunities for dialogue to support non-trivial social experiences. LLMs such as GPT-3 showed a remarkable ability to perform multiple tasks close to state-of-the-art (SOTA) results in a zero-shot fashion, which led to the emergence of "prompt engineering": constructing the input to the model using natural language, such that the output is more likely to be in the correct format, given the specific task. Reynolds et al. (Reynolds and McDonell, 2021) show an example of prompt engineering for a machine translation task; by providing the model a better sequence of tokens (designed prompt sentences), the model provides better answers than the zero-shot or few-shot examples using the "Simple Colon" as they describe.

LLMs are still not without limitations, such as their poor performance in multi-hop reasoning, potential bias, and inability to process long-term dependencies (Borji, 2023). Fully automated dialogue still involves challenges in steering the conversation, or occasionally problematic responses. Borji et al. explore many of these issues and present a categorization of failures. Our approach to minimize errors is to allow easily configuring specific LLMs, setting up prompts, and importantly–human in the loop interventions in live events (Section 4.3).

3 System

Milo is a virtual agent server that can communicate with Unity applications with a simple application programming interface (API). The system interacts with multiple clients and has multiple working configurations (see Figure 1 for system architecture). Overall, it can operate in either one of two modes:

Chat mode: the agent continuously listens to the participant and automatically reacts when detecting a long period of silence (simple detection of pause with a duration parameter that can be configured). This mode is used for dialogue scenarios.

Assist mode: the agent listens into a multi-party conversation, and intervenes in the conversation only when invoked by one of the participants or by an operator. Future research is required for allowing the system to automatically decide when to intervene in a conversation, and this is most likely application dependent.

Configurations: Milo can be configured in several ways: i) mode (Assist Chat), ii) audio input (RTP microphone), iii) prompt template, iv) underlying LLM (OpenAI models/local model from Huggingface), and iv) text-to-speech (TTS) output voice.

Milo logs all audio received, in raw format, and logs the transcription and responses, in text format. The synthesised audio is not saved.

The system has a simple API:

Agent: The agent controls the flow of events, from receiving the audio, transcribing it, sending the text to an LLM, and synthesizing the response. The agent can be configured to work with models from OpenAI or custom self-contained models from huggingface. Server: The server has an HTTP interface that includes: i) StartConversation, ii) EndConversation, and iii) GetFile (for audio speech files). Unity SDK: The SDK has few simple scripts for connecting to Milo and for sending and receiving the audio. It is available as an easy-to-install Unity package. Operator control: The server has an operation GUI for monitoring and controlling the agent behavior with a human in the loop (see Sections 4.2 and 4.3) for description of the operator control flow in the various use cases).

The process for a simple conversation use case is as follows: Upon a trigger fron the VR application (e.g., button press, gaze activation), the app calls a StartConversation from the SDK, which sends a request to the server. Optionally, the VR can receive a conversation starter text for the agent to speak when the scenario begins. Next, the system starts to listen to the head mounted display (HMD) microphone, and sends detected audio to the server via RTP. The server, when receiving a StartConversation, registers the client,

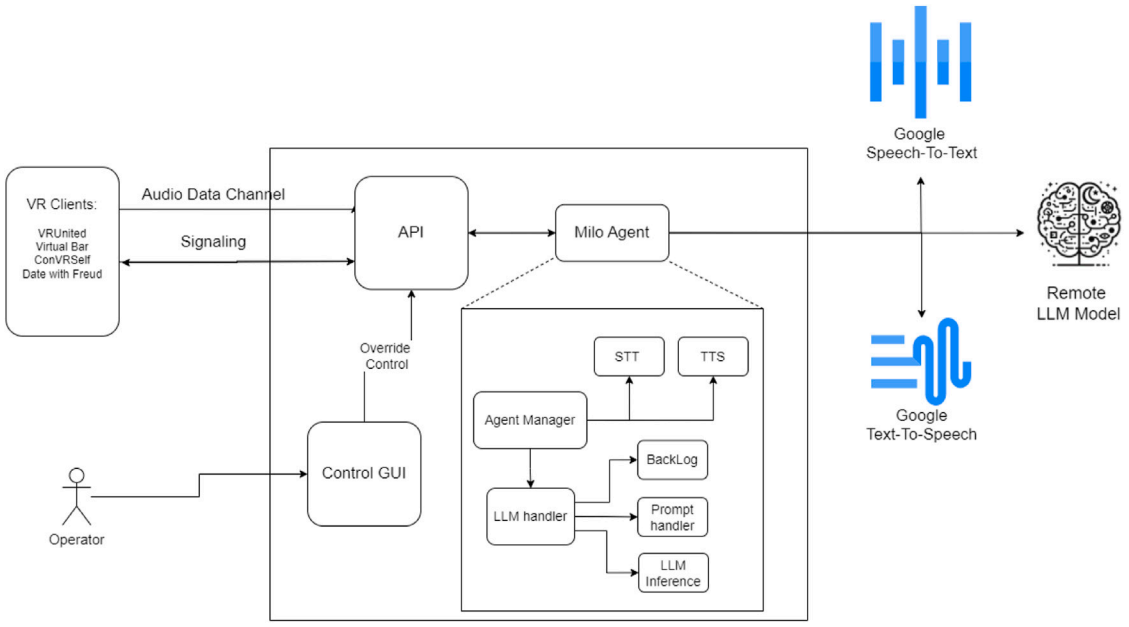

FIGURE 1
Milo Architecture Diagram. The diagram illustrates how the VR application functions as a client that connects to the system's API. Operator control is integrated to allowing overriding generated content. The Agent Manager handles connections to speech-to-text (STT), text-to-speech (TTS), and large language model (LLM) services, with the STT and TTS modules interfacing with Google's services.

creates a Milo Agent, and starts its main process. The main process for the agent is to listen to the received audio and transcribe it. When a duration of silence (configurable parameter) is detected (using voice-activity detection via the Google API), the agent assumes end of sentence and creates a response (generated audio). When the audio is ready, the server sends a signal to the VR client, the client downloads the audio from the server and plays it.

The Unity SDK controls all the process of communication with the server, and can initiate appropriate animations if an animator controller exists for the character. The animation controller needs to include at least the following states: talking, listening, and an idle state. Additionally, the developer manually sets gaze targets in the virtual space, and percentage per each target. In real time, the virtual human jumps among these targets according to the assigned probabilities. For example, in the multi-user VR scenario (Section 4.3), a virtual Einstein was set to randomly look equally at all other three participants and switch gaze every few seconds. Integrating a more realistic gaze algorithm is left for future work.

From the participant perspective, a simple scenario such as meeting an LLM-based virtual human in VR (Section 4.1) proceeds as follows. The participant enters the VR scene and can look around. An operator using the controllers triggers a startConversation request. The agent says a prerecorded conversation starter, and meanwhile the audio from the microphone is being sent to the agent by the AudioStreamer for transcription and response generation. Once the response is ready, Milo sends a trigger with the text response, and the VR app starts fetching the audio content, switching to a Talking state in the animation for the duration of the audio. The system then continues to manage a spoken conversation between the participant and the virtual human.

FIGURE 2
A participant interacting with the virtual barman. Subtitles are automatically added for improved comprehension.

4 Scenarios

4.1 The Barman

The Barman is a virtual scene that we developed to showcase our LLM-based agent platform. In this scenario, participants enter a virtual bar populated with several virtual humans and engage in conversation with the bartender, who can be either male or female (Figure 2). Utilizing this bar scene, we explored the application of a prompt-designed persona based on the Big Five personality traits —- a well-established model in personality psychology comprising five distinct dimensions.

We crafted personas for two virtual bartenders with contrasting personalities, setting their traits at various levels such as "high,"

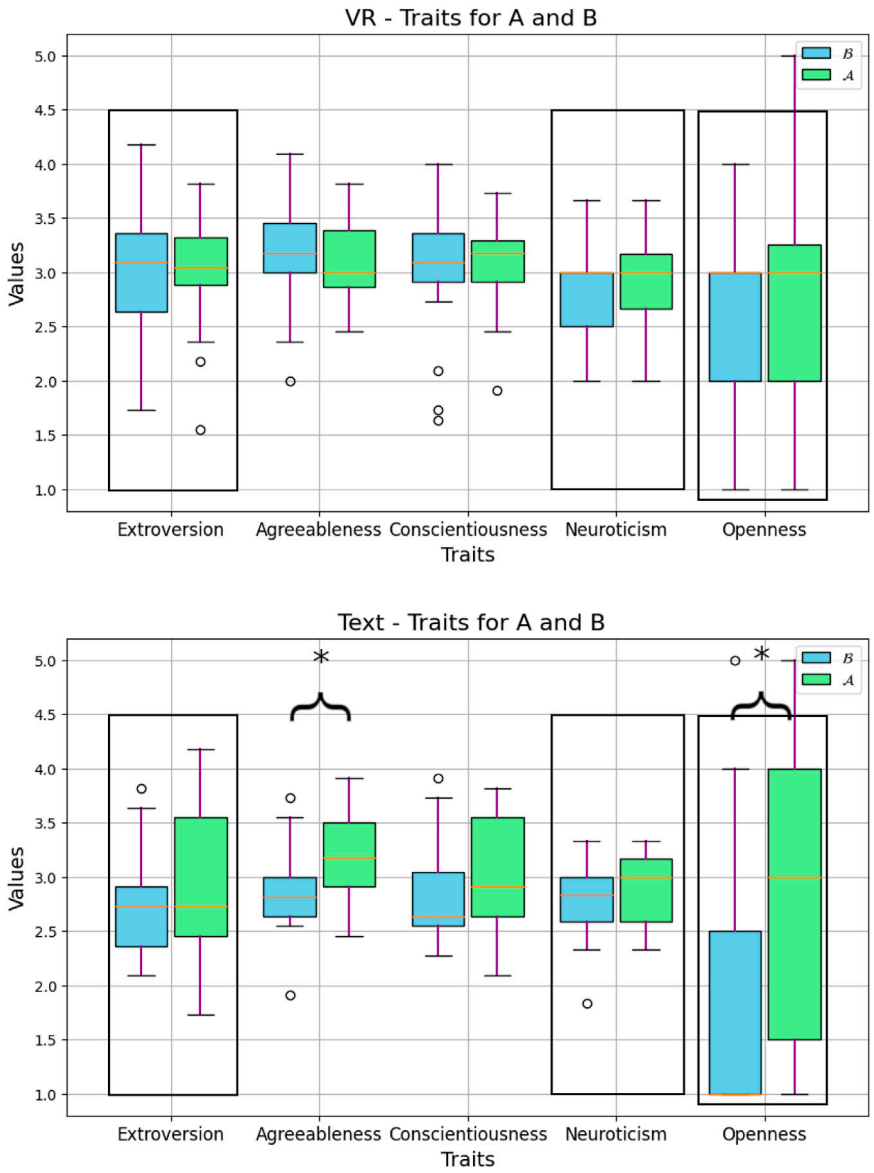

FIGURE 3
Traits for $\mathcal{A}$ and $\mathcal{B}$ (Table 1) Bar plots of personality assessments. Significant differences are marked with *. The traits we have manipulated are boxed.

"somewhat," and "low" (Table 1). We used OpenAI GPT-3, considered the best model at the time of development. In addition to the VR scene, we implemented a web-based text version with the same prompt. This allowed us to compare the effectiveness of our virtual bartenders' interactions against those conducted via a traditional text interface, serving as our baseline. We hypothesized that GPT-3 could infer personality traits solely from their names, assigning categorical values (high, somewhat, low) without any additional background information or detailed persona descriptions. The experiment employed a 2 × two factorial design with a between-subjects factor (interaction modality: text vs VR) and a within-subjects factor (personality type: two distinct designs). Each participant interacted with both personality types, with the order of presentation counterbalanced to control for sequence effects. Participants were randomly assigned to either the text-based condition (*via* web chat) or the immersive VR condition, ensuring that no individual experienced both modalities. After each conversation, participants completed a personality assessment of the bartender using items adapted from the IPIP (Markey and Markey, 2009).

Despite the informal nature of the study, and a range of limitations, we report the results; this can be considered a preliminary exploratory study, suggesting that more systematic studies are required. Significant differences were observed in this informal evaluation (Figure 3); the data is based on 17 participants in the text condition and 13 in the VR condition. Participants in the text-based sessions were partially proficient at deducing the intended personality traits of the bartender, whereas guesses regarding personality traits in VR sessions seemed almost random. In the text condition, the participants correctly judged one of the three traits significantly higher ($p < 0.05$) for personality A as compared to B – openness. The difference for agreeableness was

TABLE 1 The opposing personalities explored in the evaluation, for *A* and *B* (same *A* and *B* as in the bar plot).

Trait	Personality $\mathcal{A}$	Personality $\mathcal{B}$
Extroversion	Low	High
Openness	Low	High
Neuroticism	Somewhat	Low
Agreeableness	–	–
Conscientiousness	–	–

also judged to be different, although the prompt did not refer to this trait. For the VR condition there was no significant difference in estimated personality on any of the traits.

Several factors may explain these disparities. One notable issue was the accuracy of speech recognition, which, despite being fairly reliable (using a Google-based API in English), was not flawless and likely impacted the quality of interactions negatively. More profound differences, however, stemmed from the experiential nature of the VR environment—participants' physical sensation of being in a bar, the sense of co-presence with the bartender, and the visual and non-verbal communication dynamics of the bartender's avatar. These elements suggest a complex interplay between the dialogue content and the immersive environment, underscoring the need for further investigation into how being embodied as avatars in VR may modify the way LLMs are perceived.

Additionally, based on this scenario we created a tutorial for developing a simple VR scene, interacting with an LLM-based virtual human using Milo. The tutorial is intended for individuals with some experience with Unity but no advanced programming skills are required. It is part of the open-source system, and explains how to set up Milo and add it to a Unity project.

The bartender was used to test a hypothesis that we can generate a generate a prompt using only a description of the trait in a.

4.2 Enhancing VR self-conversation

Osimo et al., (2015) presented an intriguing protocol whereby a participant can engage in self-talk in VR. The protocol includes two avatars in a conversation, one is a look-alike of the participant and the other is a counselor (in the original study Sigmond Freud was selected). The participants keep "switching bodies" between the avatars, each time listening to themselves speak from the other avatar, then respond, essentially externalizing the process of self talk. This allows participants to view themselves from outside and offer themselves support, which was found to be beneficial for counseling purposes (Slater et al., 2019).

While VR self-talk is generally successful, authors note that sometimes the participant can "get stuck", or perhaps "run out" of things to say to themselves. Thus, we have integrated Milo with the self-talk VR application, referred to as ConVRSelf, as an AI agent that listens to the (self-) conversation. The sessions were intended for addressing mild psychological challenges with the general population. When the participants were in the body of the therapist they could opt to receive a new perspective from an LLM, ideally helping them remove their block and revitalize the self-conversation.

The virtual agent listens to the conversation and transcribes it in real-time. As to when it responds, we have explored two options, and both are technically supported by Milo: i) when the participant requests assistance, or ii) when the operator requests intervention. The operator is the person controlling the experiment using the GUI. In the latter case, the operator can see the generated text in the operator GUI, and decide if and when it is appropriate to send it to the application.

In a preliminary study, participants went through a psycho-education training and were then asked to experience the ConVRself application and discuss a personal challenge of medium importance and difficulty. The experimenter was a trained clinical psychologist. The instructions included explanations about resilience skills, about the ConVRself application, and the AI button was mentioned briefly as an option "to use if you are stuck". Surprisingly, we observed limited engagement with the AI features of the ConVRself application. Specifically, only one out of twelve participants pressed the AI assistance button during their sessions. We hypothesized that the immersive nature of the ConVRself experience might have been overwhelming, thereby diverting the participants' attention away from the AI assistance option. Indeed, the ConVRself sessions were reported to be powerful for most participants, some responded with strong emotions or even cried.

To address this issue, we first modified the protocol to allow the operator to activate the AI. This change aimed to enhance the frequency and relevance of AI interactions during the experiments. However, feedback from pilot studies indicated that the AI's presence needed to be more pronounced within the experience. To address this, in a subsequent iteration, the AI was visually represented as an avatar resembling Albert Einstein in the virtual environment. This avatar was activated by the participants' gaze; specifically, when embodied as the therapist, participants could seek support from the AI by directing their gaze towards "him." In this enhanced setup, the AI's presence was more prominently felt, as it was actively engaged by the participants who utilized it to respond during the sessions. Figure 4 illustrates the various VR scenarios implemented.

Using these and a few other improvements we have reached a version of AI-enhanced ConVRself that seems to be usable and useful, and the results of an evaluation study will be reported elsewhere. A video of the full experience, including self-talk VR and conversation with the LLM, is available[2].

4.3 Live XR events with virtual public figures

Milo was used to control virtual humans in two live events, showcasing the integration of both VR and AR into physical events. Both events were part of academic conferences.

The first was a live VR panel session broadcast at the conference "XR for the People" held at our university in June 2022. Instead of a remote video conference panel, the conference included a multi-user VR panel with three (real) participants from three different

2 https://www.youtube.com/watch?v=JBPzG02ofsA

FIGURE 4
Snapshots from AI-enhanced VR self talk using Milo. Left: a button-activated voice agent. Right: a side view of the scene: the LLM-based agent is embodied as a virtual Einstein, the participant faces a counselor avatar, in this case embodied as Barack Obama.

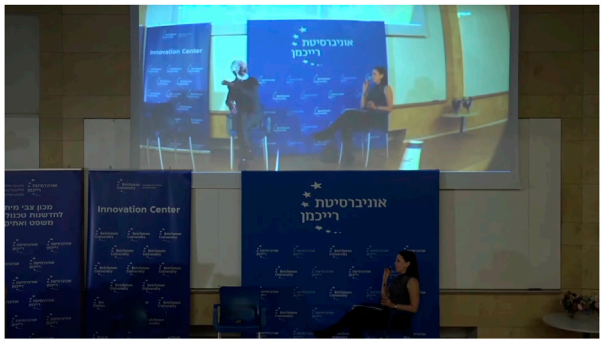

FIGURE 5
Virtual Freud with real speaker (Prof. Gurit Birnbaum), on stage during the live event. The augmented view from the Oculus Pro device was projected above the real speaker.

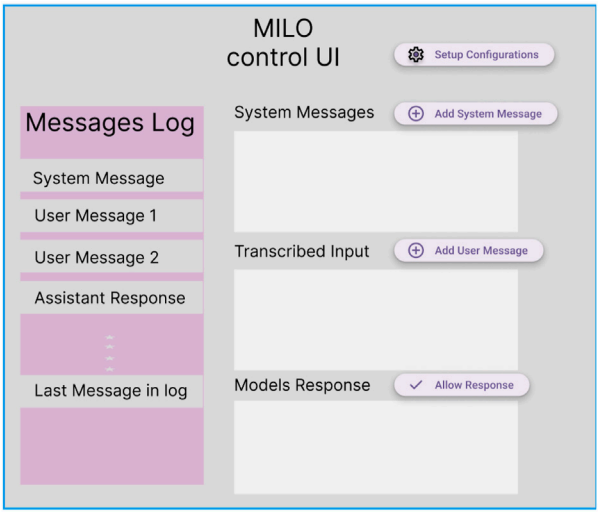

FIGURE 6
An illustration of the UI used in Date with Freud. The log can be switched (move up, down) and messages can be removed. An editable text box for the system messages and editable user messages is updated in real time with each new transcribed message. Finally, the UI shows the response output, generated by the LLM, which can be sent to the TTS to be transcribed.

continents, seating around a table in a (virtual) Sushi restaurant and discussing one of the panelist's new book. A physical moderator was on the physical stage and bridged between the VR panelists and the live audience. Additionally, a virtual Albert Einstein took part in the conversation, controlled by Milo. The event and lessons learned were described in a conference paper (Shoa et al., 2023)[3].

A Date With Freud: in March 2023, we conducted another live-stage event: a discussion between Prof. Gurit Birnbaum, a world expert on intimacy and relationship research, and a virtual Sigmond Freud. This time we used a video see-through AR setup (referred to in industry as mixed reality): the audience could see the presenter talk to an empty chair, while the scene augmented with a virtual Sigmond Freud was projected on a screen above her. The scene was captured and broadcast live from a 'mixed reality' Oculus Pro. device (Figure 5).

For live events, we incorporate a human operator, who functions similarly to a director in a real-time TV studio, overseeing not only the virtual actors but also other elements such as cameras, lights, and props. Additionally, the operator plays a crucial role as a safeguard for the responses generated by the LLM.

For the multi-user VR panel, we parallelized the content creation to have multiple (4) choices of generated text, allowing the operator to manually select the preferred response and send it to the avatar for enacting. We later removed this option because although it was carried out by multiple threads and simultaneous calls in practice it slowed down process. Nevertheless, we suggest such opportunities for utilizing the "human-in-the-loop" deserve further investigation.

The operator had the ability to view or edit the transcribed audio, edit prompts templates, and to initiate or stop a response. For the Freud live

3 https://www.youtube.com/watch?v=98QKzT1dkpo

event we have improved the control GUI (Figure 6), based on lessons learned in the first event, as well as the evolution of LLMs, such as the introduction of system messages in the OpenAI API. The UI allowed the operator to 'direct' the virtual Freud in real-time–e.g., trying to make it less cautious when discussing intimate relationships, or instructing it to use humor in specific cases (see video[4]).

Supporting extended reality (XR) introduces additional challenges and requirements. We have developed an intuitive interface that allows for easy placement of virtual humans, such as Freud, in specific locations, revisedlikee.g. on a chair. In VR, gaze targets for the virtual human are typically predefined manually by developers using Unity and specific assets. In contrast, our XR system empowers the director to set gaze targets directly within the physical scene, so that the event could be adapted to any physical location–in our case the director was able to easily set up the virtual Freud to switch gaze between real participant and the audience. Furthermore, the director has the capability to manually adjust eye gaze targets for Freud, such as the human participant on stage or different sections of the audience in the auditorium—center and both sides.

We implemented this scenario as a multi-user co-located XR scene utilizing the Photon Engine[5] to accommodate multiple individuals wearing HMDs. This capability was eventually not utilized in practice, as the presenter opted not to wear an HMD.

5 Discussion

We presented Milo–a system for the integration of LLM-based spoken dialogue within extended reality (XR) platforms. This framework simplifies the workflow for incorporating virtual humans into a diverse array of social scenarios, making it accessible for a broad spectrum of applications. Throughout this paper, we have explored various use cases of Milo and detailed the lessons learned from each, demonstrating the system's versatility and ease of deployment. Notably, we highlighted its efficacy in facilitating live events, where an operator can manage the virtual human in real-time, providing dynamic responses to unfolding situations.

The Milo system is available under an Academic Free License 3.0. Accompanying the system, we offer an accessible tutorial designed specifically for social science students who possess a background in Unity but lack extensive programming skills. Our objective is to democratize the use of LLM-based virtual characters, extending their potential applications across various fields and creative domains. We envision that Milo will serve as a catalyst for further exploration and innovation, benefiting not only students and researchers but also communicators and artists who wish to engage with this cutting-edge technology.

5.1 Ethical considerations

Our system combines two transformative yet ethically complex technologies: extended reality (XR) and generative artificial intelligence (AI). Each technology has its own set of ethical challenges, as documented in the literature. For instance, generative AI often raises concerns regarding misinformation, privacy, and autonomy, whereas XR has been scrutinized for issues related to user safety, psychological impacts, and the potential blurring of boundaries between reality and simulation (Slater et al., 2020). When these technologies are integrated, they create additional ethical dilemmas—most notably regarding the legal and moral implications of recreating identifiable individuals, whether celebrities or private persons. This raises important questions about consent, digital likeness rights, and the risk of misuse. Furthermore, it underscores the need for robust guidelines to ensure that such representations are both accurate and respectful.

Milo, as a system, introduces unique ethical and legal challenges because it deals with highly regulated areas, including user data privacy and AI usage. To address these challenges, we have implemented several measures:

User Consent: In all lab experiments, participants complete an informed consent form. Data is stored on local hard drives within encrypted folders to protect privacy.

Data Minimization and Retention: Data collection is optional, with files saved locally on the server hosting Milo. Audio recordings (in their raw format), transcriptions, and AI-generated responses are stored as text files. However, there is currently no automated mechanism for data deletion; data management must be handled manually to comply with cloud-based GDPR requirements.

Security Measures: All data transfers utilize secured channels (e.g., HTTPS for speech-to-text/text-to-speech services and OpenAI connections, or SSL for self-hosted models). Locally stored data must be manually secured using third-party encryption tools.

User Rights: Since the data is saved locally, the responsibility for GDPR compliance falls on the user or experimenter, who must manage data access or removal based on their own procedures.

In order to align with the EU Digital AI Act–during experiments, operational logs and monitoring data are maintained only while the experiment is active; these logs are not archived or audited after the experiment concludes. Likewise, audio data, transcriptions, and LLM responses are stored in a configurable local folder that should be manually encrypted if used.

Beyond these technical considerations, the integration of XR and generative AI offers the potential to create immersive social experiences that could help address social isolation and loneliness. However, there is also a risk of cultivating inauthentic interactions that may detract from genuine human contact and community engagement. As (Chalmers, 2005) posits, experiences in VR deserve the same seriousness as those in the real world, especially when involving multi-user environments. Therefore, it is crucial to continuously evaluate whether these virtual interactions ultimately enhance or undermine the quality of human relationships.

5.2 Limitations

Our system exhibits two primary limitations that impact its effectiveness in simulating naturalistic conversational dynamics. The first limitation concerns the mechanism of automatic turn-taking. Presently, the system manages dyadic interactions using a

4 https://www.youtube.com/watch?v=gh4LszoIQ94

5 www.photonengine.com

rudimentary approach that primarily detects periods of silence. To achieve a seamless conversational flow, it may be necessary to implement more sophisticated, multimodal cue-based techniques. This challenge becomes even more pronounced in multi-party scenarios, where the intricacies of conversational flow, such as managing interruptions and recognizing subtle cues for turn exchanges, are not sufficiently addressed. The design of agents capable of listening to a conversation and deciding opportune moments to intervene remains an open question and is heavily dependent on the specific application. Enhancing our model's ability to discern and respond to these nuances is a critical area for ongoing research and development.

Additionally, while the latency within our system is currently deemed acceptable based on user feedback, it falls short of replicating the rapid exchange typical of real-life conversations. Future improvements should focus on optimizing architecture and algorithms to further reduce latency, e.g., by introducing anticipatory processing.

5.3 Open source alternatives for commercial tools

Non-commercial tools can also be implemented and extended within Milo. The system includes a built-in module for using local or self-hosted large language models (LLMs), and it provides a simple service for loading models from Hugging Face and serving predictions over an SSL-secured API. Although there is no built-in implementation for speech-to-text (STT) and text-to-speech (TTS), many excellent open-source alternatives are available (e.g., Kaldi or SpeechBrain (Can et al., 2018; Ravanelli et al., 2021). Similar to the LLM module, a local or remote service must be developed for these speech functions; however, unlike the LLM where processing is handled on textual data, the speech modules must accommodate streaming audio to generate outputs in real time.

5.4 Future directions

The continuous evolution of LLMs opens up numerous possibilities for enhancements and innovations in virtual environments. One particularly promising extension of Milo is the deployment of multiple LLM-based virtual humans within a single scene. Such scenarios pose complex challenges in managing multi-party interactions, which include both verbal and non-verbal communication elements like turn-taking and gaze dynamics. Developing advanced models that can effectively interpret and respond to the intricate dynamics between multiple interactive agents will be crucial. Recent advancements in language generation have shifted the research focus from task-oriented dialogues, as seen in works like Budzianowski et al. (Budzianowski and Vulić, 2019), to more elaborate storytelling methods, exemplified by the narrative techniques in Fable (Maas, 2023). These storytelling approaches offer sophisticated frameworks for episodic generation through simulations, suggesting the potential to revisit and innovate upon pioneering methods such as those proposed by Bates et al. (Bates et al., 1992), who suggested, a few decades ago, combining multiple AI agents and real-time direction inspired by improvisational theater – as a framework for interactive storytelling.

Currently, LLMs are predominantly utilized in text-based applications, with some transition to audio interactions. A further transition into fully embodied virtual agents in VR presents a distinctly different set of challenges and experiences. Our preliminary studies have already indicated significant differences in the perception of personality traits, such as those defined by the Big Five inventory, between interactions in text and VR. This distinction highlights the vast potential for further exploratory research in this area.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding authors.

Ethics statement

Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

AS: Conceptualization, Investigation, Methodology, Project administration, Software, Visualization, Writing – original draft, Writing – review and editing. DA: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This research was supported by the Horizon 2020 Information and Communication Technologies (ICT) project - Leadership in Enabling and Industrial Technologies - SOCRATES (#951930), and by the Horizon3432020 FET Proactive project GuestXR (#101017884).

Acknowledgments

We thank our collaborators in the multiple projects mentioned in this paper: Yair Moss, Ramon Oliva, Maya Shekel, Mel Slater, and Moreah Zisquit.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Bates, J. (1992). Virtual reality, art, and entertainment. *Teleoperators and Virtual Environments,* 1 (1), 133–138. doi:10.1162/pres.1992.1.1.133

Borji, A. (2023). A categorical archive of chatgpt failures.

Bradley, J., Mival, O., and Benyon, D. (2009). Wizard of oz experiments for companions.

Budzianowski, P., and Vulić, I. (2019). Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv Prepr. arXiv:1907.05774*.

Can, D., Martinez, V. R., Papadopoulos, P., and Narayanan, S. S. (2018). "Pykaldi: a python wrapper for kaldi," in *2018 IEEE international conference on (IEEE)*.

Cassell, J. (1999). Embodied conversational agents: representation and intelligence in user interfaces. *AI magazine,* 22 (4), 67–67.

Chalmers, D. J. (2005). "The matrix as metaphysics," in *Philosophers explore the matrix*. Editor C. Grau (Oxford University Press), 132–176.

Maulsby, D., Greenberg, S., and Mander, R. (1993). Prototyping an intelligent agent through Wizard of Oz. In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. 277–284. doi:10.1145/169059.169215

Friedman, D., and Hasler, B. S. (2016). "The beaming proxy: towards virtual clones for communication," in *Human computer confluence transforming human experience through symbiotic technologies*, 156–174.

Kishore, S., Muncunill, X. N., Bourdin, P., Or-Berkers, K., Friedman, D., and Slater, M. (2016). Multi-destination beaming: apparently being in three places at once through robotic and virtual embodiment. *Front. Robotics AI* 3, 65. doi:10.3389/frobt.2016.00065

Markey, P. M., and Markey, C. N. (2009). A brief assessment of the interpersonal circumplex: the ipip-ipc. *Assessment* 16, 352–361. doi:10.1177/1073191109340382

McNeill, D. (1992). *Hand and mind: what gestures reveal about thought*. University of Chicago Press.

Nakash, T., Haller, T., Shekel, M., Pollak, D., Lewenchuse, M., Klomek, A. B., et al. (2022). "Increasing resilience and preventing suicide: training and interventions with a distressed virtual human in virtual reality," in *Proceedings of the 22nd ACM international conference on intelligent virtual agents* (New York, NY, USA: Association for Computing Machinery). doi:10.1145/3514197.3549613

Osimo, S. A., Pizarro, R., Spanlang, B., and Slater, M. (2015). Conversations between self and self as sigmund freud—a virtual body ownership paradigm for self counselling. *Sci. Rep.* 5, 13899. doi:10.1038/srep13899

Patlan, A. S., Tripathi, S., and Korde, S. (2021). A review of dialogue systems: from trained monkeys to stochastic parrots.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., et al. (2009). SpeechBrain: a general-purpose speech toolkit. ArXiv:2106.04624.

Reynolds, L., and McDonell, K. (2021). Prompt programming for large language models: beyond the few-shot paradigm, 1, 7. doi:10.1145/3411763.3451760

Shoa, A., Oliva, R., Slater, M., and Friedman, D. (2023). Sushi with einstein: enhancing hybrid live events with llm-based virtual humans, 1, 6. doi:10.1145/3570945.3607317

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Trans. R. Soc. B Biol. Sciences* 364, 3549–3557. doi:10.1098/rstb.2009.0138

Slater, M., Gonzalez-Liencres, C., Haggard, P., Vinkers, C., Gregory-Clarke, R., Jelley, S., et al. (2020). The ethics of realism in virtual and augmented reality. *Front. Virtual Real.* 1, 1. doi:10.3389/frvir.2020.00001

Slater, M., Neyret, S., Johnston, T., Iruretagoyena, G., de la Campa Crespo, M., Alabèrnia-Segura, M., et al. (2019). An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. *Sci. Rep.* 9, 10903–10913. doi:10.1038/s41598-019-46877-3

Traum, D., Jones, A., Hays, K., Maio, H., Alexander, O., Artstein, R., et al. (2015). "New dimensions in testimony: digitally preserving a holocaust survivor's interactive storytelling," in *Interactive storytelling: 8th international conference on interactive digital storytelling, ICIDS 2015, Copenhagen, Denmark, november 30-december 4, 2015, proceedings 8*. Springer, 269–281.