My Co-worker ChatGPT: Development of an XR Application for Embodied Artificial Intelligence in Work Environments

Philipp Krop*
Human-Computer Interaction Group
University of Würzburg

David Obremski
Psychology of Intelligent Interactive Systems
University of Würzburg

Astrid Carolus
Media Psychology
University of Würzburg

Marc Erich Latoschik
Human-Computer Interaction Group
University of Würzburg

Carolin Wienrich
Psychology of Intelligent Interactive Systems
University of Würzburg

Abstract

With recent developments in spatial computing, work contexts might shift to augmented reality. Embodied AI - virtual conversational agents backed by AI systems - have the potential to enhance these contexts and open up more communication channels than just text. To support knowledge transfer from virtual agent research to the general populace, we developed *My CoWorker ChatGPT* - an interactive demo where employees can try out various embodied AIs in a virtual office or their own using augmented reality. We use state-of-the-art speech synthesis and body-scanning technology to create believable and trustworthy AI assistants. The demo was shown at multiple events throughout Germany, where it was well received and sparked fruitful conversations about the possibilities of embodied AI in work contexts.

Index Terms: Virtual Humans, Embodied Artificial Intelligence, Augmented Reality, Work Environments.

1 Introduction

Artificial Intelligence (AI) assistants have become an increasingly important factor in personal and work contexts [9]. Recent developments in spatial computing led to the rise of new use cases for AI assistants, like the introduction of embodied AI. Embodied AI relates to AI that can interact through embodied interfaces, such as embodied conversational agents [10], often using natural communication channels. Although embodied conversational agents have been an important topic in research for a long time [8], they could not reach the working population because of their limited capabilities. Recent advancements in AI now allow embodied AI to have more natural conversations than traditional text-based interfaces because they can interact spatially and integrate natural interaction modalities, such as speech, facial expressions, and gestures. They thus are more scalable than their predecessors and have the potential to be the next step in designing believable, trustworthy AI systems for personal and work contexts.

However, consumers are still unfamiliar with embodied AI in the workplace, as knowledge transfer from research often takes too long or fails to reach businesses. Especially German companies are behind their competitors in other European countries with a similar GDP in adapting technological advancements [2]. The reasons for this are multifaceted. Employees' mistrust in new technologies such as AI, e.g., due to perceived risks to their privacy [7], or unfamiliarity with the technology [4], might be the root causes here. To successfully transfer the potential of embodied AI in the workplace from research to industry, it is imperative for researchers to develop and frequently show interactive demonstrators to both stakeholders and employees.

Figure 1: A screenshot of the embodied AI shown either in the user's office using augmented reality (left, center) or in a virtual office (right). The AI's embodiment is sourced from either the Unity Asset store (left), the *Rocketbox* library (center), or a digital replica of a person constructed using photogrammetry (right).

However, to the best of our knowledge, no suitable demonstrator that lets a person experience how work contexts can be enhanced with embodied AI exists yet. To close this gap, we developed *My CoWorker ChatGPT*, an easily accessible and transportable demonstrator where stakeholders at public events can get a glimpse of how work contexts can be enhanced with embodied AI. The demonstrator provides an XR-based scenario in which the user can naturally interact with an embodied AI using speech in a work environment. The embodied AI's behavior is controlled by GPT-4o, and its responses are played back to the user using state-of-the-art speech synthesis, virtual human replicas, predefined facial expressions, and gestures. A screenshot of the system is shown in Figure 1.

2 System Architecture

The demo runs on a laptop with an Intel Core i7-10875H CPU, an NVIDIA GeForce RTX 2070 SUPER GPU, and 16 GB of RAM connected to a *Quest 3* headset via link cable. It was developed using *Unity Engine 2022.3.19f1* and builds upon the *Reality Stack* framework by Kern & Latoschik [5] and the *Meta SDK* plugin to implement believable virtual humans in augmented reality. The embodiments of the virtual agent were sourced from the *Unity Asset Store*, the *Rocketbox* library [3], or scanned using the virtual human rendering pipeline by Achenbach et al. [1]. The embodied AIs receive two prompts before a user can interact with them: One prompt with generic information about the event (such as venue location, the event's topic, and some talking points that can spark discussions about the AI's role in human-AI interaction), and a second prompt specifying who the virtual agent embodies. This can either be a fictional person (when the AI is embodied with a *Rocketbox* character

*e-mail: philipp.krop@uni-wuerzburg.de

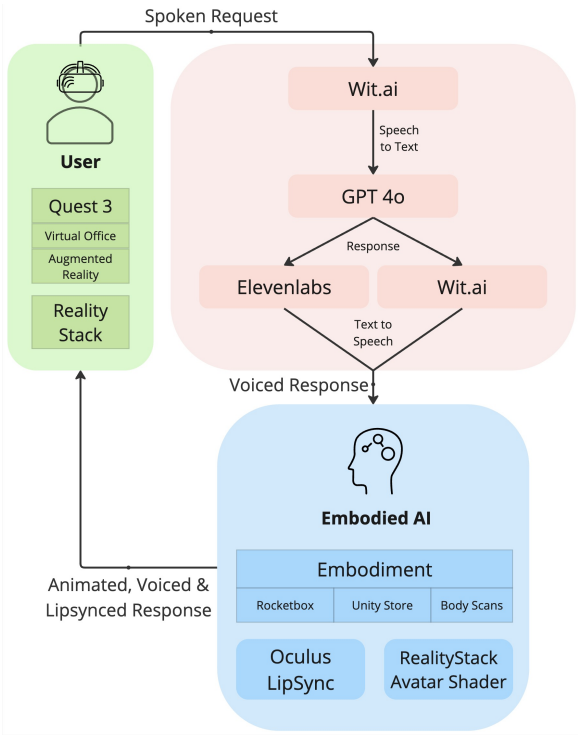

Figure 2: The architecture of *My CoWorker ChatGPT*. Users see either a virtual office or their own office using augmented reality. When a user speaks to the embodied AI, the request is first translated to text using *Wit.ai* and sent to *GPT-4o*. It is then sent to *Elevenlabs* or *wit.ai* for speech synthesis. This voiced response is then played back through the embodied AI using a fitting animation and lipsync.

or one from Unity's asset store) or a real person when the AI is embodied using a body scan. The AI is animated using animations from the *Mixamo* library.

Users can interact with the embodied AI by pressing the *A*-button on their controller to speak with the AI. The user's voice inputs are transcribed using *Wit.ai*, sent to *ChatGPT* to generate a response, and then sent to *Elevenlabs* or *Wit.ai* for speech synthesis using respective API calls. If the embodied AI is embodied with a body-scanned character, we use a speech model trained with *ElevenLabs* to imitate the original person's look and voice. The voiced response is then streamed back to the user and lip-synced using *Oculus Lipsync*. A schematic of the architecture is shown in Figure 2.

3 Initial Reception & Future Work

The demo was shown at various events throughout Germany - from public scientific events to more industry-oriented events like the *Landshut Leadership Forum* and a presentation in the German parliament. It was well received and served as a starting point for fruitful conversations with stakeholders and employees about the role of embodied AI in future work contexts and how human-AI interaction should be designed. Informal feedback indicated that users perceived the interaction with the embodied AI as more natural and personal than chat-based systems and liked that they could select a virtual representation fitting for a specific task.

We plan to expand the system in future work. Currently, only one exemplary work context - an office - is supported. Future work will expand this to provide users with a wider variety of virtual environments and embodiments. We hope this will make the AI more trustworthy, as previous research has shown that users prefer virtual agents they perceive as congruent to the task and environment [6]. In addition, we plan the embodied AI's believability by inferring the AI's emotional state and generating fitting animations based on its response, as currently, only predefined animations are supported. Finally, we plan to formally evaluate the demonstrator as a test bed for consecutive studies on human-AI interaction in work contexts.

4 Conclusion

We presented *My Co-worker ChatGPT*, an interactive demo showcasing what human-AI interaction could be like in future work contexts. Users can interact with a GPT-4o-controlled embodied AI in both a virtual office (using VR) or in their own office (using AR) via speech. They can select from various virtual representations, including virtual replicas of real people, that can speak with their natural voice using state-of-the-art voice synthesis. It is designed to be accessible and transferrable, running on a *Meta Quest 3* and a laptop with consumer hardware. Although a formal evaluation is outstanding, feedback from various events shows that it provides an easy way for stakeholders to understand what work with an embodied AI could be like and spark interesting conversations about the future of human-AI interactions.

Acknowledgments

The authors thank Samantha Straka for her work on the first version of the demo. This work was funded by the German Federal Ministry of Labour and Social Affairs [DKI.00.00030.21].

References

[1] J. Achenbach, T. Waltemate, M. E. Latoschik, and M. Botsch. Fast generation of realistic virtual humans. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*, pp. 1–10, 2017. 1

[2] European Commission. Digital economy and society index (DESI) 2021, 2021. Accessed: 2024-12-21. 1

[3] M. Gonzalez-Franco, E. Ofek, Y. Pan, A. Antley, A. Steed, B. Spanlang, A. Maselli, D. Banakou, N. Pelechano, S. Orts-Escolano, et al. The rocketbox library and the utility of freely available rigged avatars. *Frontiers in virtual reality*, 1:20, 2020. 1

[4] M. C. Horowitz, L. Kahn, J. Macdonald, and J. Schneider. Adopting ai: how familiarity breeds both trust and contempt. *AI & society*, 39(4):1721–1735, 2024. 1

[5] F. Kern and M. E. Latoschik. Reality stack i/o: A versatile and modular framework for simplifying and unifying xr applications and research. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 74–76, 2023. 1

[6] P. Krop, M. J. Koch, A. Carolus, M. E. Latoschik, and C. Wienrich. The effects of expertise, humanness, and congruence on perceived trust, warmth, competence and intention to use embodied ai. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24. Association for Computing Machinery, New York, NY, USA, 2024. doi: 10.1145/3613905.3650749 2

[7] M. K. Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, 2018. 1

[8] B. Lugrin. Introduction to socially interactive agents. In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition*, pp. 1–20. Association for Computing Machinery, New York, NY, USA, 2021. 1

[9] C. Rzepka and B. Berger. User interaction with ai-enabled systems: A systematic review of is research. In *ICIS 2018 Proceedings*, vol. 39. AIS Electronic Library, San Francisco, 2018. 1

[10] C. Wienrich and M. E. Latoschik. extended artificial intelligence: New prospects of human-ai interaction research. *Frontiers in Virtual Reality*, 2:94, 2021. doi: 10.3389/frvir.2021.686783 1