International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Dialogs with GenAI NPCs: Exploring Player Interactions with Speech Agents in a VR Game

Nima Zargham, Leandro Tonini, Dmitry Alexandrovsky, Emma Grace Ruthven, Maximilian A. Friehs, Leon Tristan Dratzidis, Bastian Dänekas, Ioannis Bikas, Lennart E. Nacke, Sven Zebel & Rainer Malaka

To cite this article: Nima Zargham, Leandro Tonini, Dmitry Alexandrovsky, Emma Grace Ruthven, Maximilian A. Friehs, Leon Tristan Dratzidis, Bastian Dänekas, Ioannis Bikas, Lennart E. Nacke, Sven Zebel & Rainer Malaka (02 Feb 2026): Dialogs with GenAI NPCs: Exploring Player Interactions with Speech Agents in a VR Game, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2026.2620647

To link to this article: https://doi.org/10.1080/10447318.2026.2620647

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC

View supplementary material

Published online: 02 Feb 2026.

Submit your article to this journal

Article views: 261

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hihc20

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2026.2620647

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

Dialogs with GenAI NPCs: Exploring Player Interactions with Speech Agents in a VR Game

Nima Zargham[a*], Leandro Tonini[b*], Dmitry Alexandrovsky[c], Emma Grace Ruthven[d], Maximilian A. Friehs[b#], Leon Tristan Dratzidis[a], Bastian Dänekas[a], Ioannis Bikas[a], Lennart E. Nacke[e], Sven Zebel[b] and Rainer Malaka[a]

[a]Digital Media Lab, University of Bremen, Bremen, Germany; [b]Psychology of Conflict Risk and Safety, University of Twente, Enschede, Netherlands; [c]HCI and Accessibility, Karlsruhe Institute of Technology, Karlsruhe, Germany; [d]Psychology, University of Strathclyde, Glasgow, Scotland; [e]HCI Games Group, Stratford School of Interaction Design and Business, University of Waterloo, Waterloo, Canada

ABSTRACT

Generative artificial intelligence (GenAI) can potentially enhance player experiences by enabling more dynamic and adaptive interactions. This exploratory research investigates player perceptions and interactions with GenAI-based non-player characters (NPCs) in a speech-based Virtual Reality (VR) game. "Office Whispers," is a VR adventure-puzzle game featuring four GenAI-based NPCs with diverse human characteristics. Our findings from the user-study show that players had an overall positive experience and found the game novel. Players appreciated the freedom of expression and felt deeply immersed when NPCs responded in a believable manner. However, issues such as unnatural conversational flow, incorrect or inconsistent responses, and uninformative dialogue disrupted immersion and gameplay. Our study contributes new insights into the potential and the current limitations of GenAI-based NPCs in speech-based games. We outline design implications and offer guidance for game designers and developers on adapting AI-based NPCs to support more immersive and engaging player experiences.

CCS CONCEPTS

• Human-centered computing → Natural language interfaces;
• Applied computing → Computer games.

KEYWORDS

Generative AI; voice interaction; virtual reality; NPCs; video games

1. Introduction

The rise of large language models (LLMs) and generative AI (GenAI) technologies has unlocked new opportunities for innovation in gaming (Bubeck et al., 2023; Garcia-Pi et al., 2023; Shoa et al., 2023). LLMs are increasingly used to facilitate game development, including procedural content generation (PCG) to create game assets (Maleki & Zhao, 2024), maps and levels (Todd et al., 2023), narrative support (Kumaran et al., 2023), quest descriptions (Värtinen et al., 2024), and dialogue generation (Chen Gao & Emami, 2023). Researchers have also explored the use of GenAI in designing virtual characters (See Figure 1 for an illustration of a player in virtual reality conversing with an agent). For instance, Wang et al. (2023) introduced an LLM-driven learning agent capable of autonomous exploration, skill acquisition, and discovery. Similarly, Park et al. (2023) presented virtual agents that simulate realistic human behaviors such as planning, memory, reflection, and social interaction. These agents, demonstrated in a Sims-like sandbox environment, autonomously exhibit believable individual and social

CONTACT Maximilian A. Friehs m.a.friehs@utwente.nl Psychology of Conflict Risk and Safety, University of Twente, Enschede, Netherlands
*Both authors have contributed equally to this research.
†Additional affiliation: School of Psychology, University College Dublin, Ireland.
Supplemental data for this article can be accessed online at https://doi.org/10.1080/10447318.2026.2620647.

© 2026 The Author(s). Published with license by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

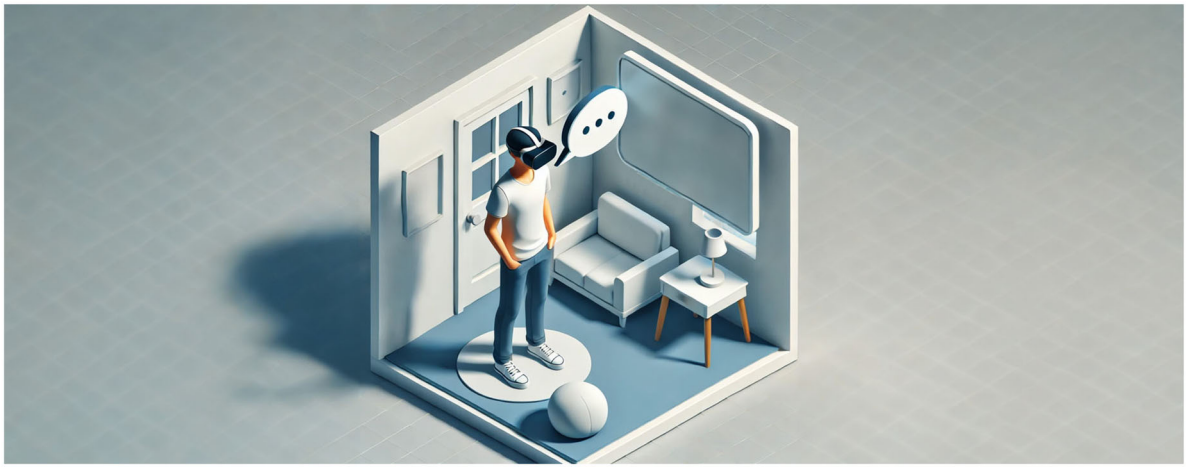

Figure 1. AI generated image depicting a person playing a speech-based VR game. The image was generated using DALL-E 3.

behaviors. With the rapid improvement in the interactive capabilities of GenAI models, agents powered by these technologies will soon become ubiquitous in gaming (Rao et al., 2024). Hence, it is essential to evaluate their adaptation in games with the current performance of this technology and, more importantly, to understand how players perceive their interactions with such games and AI-driven NPCs.

As the gaming industry grows and adapts to increasing consumer demand, video games strive to deliver more immersive experiences for players (Cairns et al., 2014; Leroy, 2021). Immersion is a state of deep absorption in gameplay that involves a loss of awareness of time, a detachment from the real world, and a deep involvement and sense of being in the task environment (Jennett et al., 2008), resulting from a positive gaming experience. One notable direction is the integration of cutting-edge technologies, such as virtual reality (VR), which moves users experientially closer to a natural interaction with the game environment. VR games enhance players' sense of spatial presence, enabling them to feel more connected to the game world (Ishiguro et al., 2019). Another crucial element that impacts immersion for many players in modern games is social interaction (Lee et al., 2006; Zhao et al., 2018). Multiplayer games allow for natural language communication among players and foster a sense of community. Previous research has shown that social interactions in online gaming form a considerable element in the enjoyment of playing (Cole & Griffiths, 2007) and the level of engagement of the players (Chen et al., 2006). In contrast, social interaction in single-player games is either completely absent or limited to scripted interactions with NPCs, mostly relying on dialogue trees (Zargham et al., 2020). Typically, these interactions lack the dynamic and responsive nature of real-time human interactions (Zargham, Friehs, et al., 2024), which can make single-player games feel less engaging and more rigid (Chen et al., 2006). The predictability and perception of having no choice diminishes replay value and can lead to boredom and decreased player engagement over time (Bowey et al., 2019).

Research has shown that speech interaction in games leads to elevated immersion (Zargham et al., 2022) and stimulates social engagement among players (Hicks et al., 2018). Using speech in single-player video games can foster more natural interactions, enhance player engagement, and improve accessibility (Zargham, Friehs, et al., 2024). This opens up new avenues for exploration, where speech interaction can serve as a bridge for fostering collaborative experiences within gaming environments traditionally seen as solitary.

Despite the recent advances over the past few years and increased research on speech interaction in video games (Allison, 2020; Anzai et al., 2021; Hedeshy et al., 2022; Hong et al., 2021; Zargham et al., 2022; Zargham, Fetni, et al., 2024), only a few studies have examined player interactions with NPCs that use GenAI to support dynamic, real-time conversations (Christiansen et al., 2024; Dratzidis et al., 2025; Zargham et al., 2025). The Player-NPC dialogues in current video games are mainly prescripted, limiting the range of interactions players can have with NPCs. They often lack flexibility and dynamism, making conversations predictable and repetitive. Moreover, NPCs usually cannot adapt their responses based on the player's unique actions or dialogue choices, potentially reducing the sense of

immersion and limiting player agency. As a result, there is still limited understanding of the benefits and challenges of integrating AI-driven, adaptive NPCs into speech-based games, particularly regarding their impact on player immersion, engagement, and overall experience. Recent work has begun exploring LLM-based agents in game-like contexts, such as social VR environments (Garcia-Pi et al., 2023; Wan et al., 2024). For example, Celeste AI (2022) in VRChat acts as a virtual companion capable of answering players' questions. However, in video games, such agents remain largely underexplored.

In this work, we present an exploratory study investigating players' experiences in a single-player VR game where they interact verbally with LLM-based NPCs. Our previous research investigated how players converse and interact with NPCs that varied in humanlike qualities within a speech-based game. We found that players adapted their speech patterns based on NPC attributes such as gender and hierarchical role. However, in that work, the interaction system relied on a rule-based system using a dictionary of predefined voice commands, which restricted the range of possible player utterances and limited conversational flexibility. Building on these findings, the present work uses LLMs to support open-ended, dynamic speech interaction and expand the vocabulary of possible player commands. To explore these interactions, we developed a speech-based adventure and puzzle game in VR and conducted a user study to evaluate the resulting player experience. The game features multiple NPCs with distinct characteristics, including gender, age, appearance, and hierarchical role, allowing us to examine how players engage with diverse LLM-driven characters in an immersive environment.

Research shows that speech-based games can raise concerns related to environmental awareness, as players may feel self-conscious when speaking aloud in non-immersive settings such as PC games (Zargham, Fetni, et al., 2024). Prior work suggests that pairing speech interaction with immersive technologies, such as VR, can reduce external distractions and support deeper player engagement (Zargham, Friehs, et al., 2024). Recent studies further indicate that combining speech input, LLMs for dynamic dialogue, and VR environments can substantially enhance immersion and overall game engagement (Christiansen et al., 2024). Based on these insights, we designed our game for VR to provide a more immersive, distraction-reduced setting where players can engage naturally in spoken interaction with GenAI-based NPCs.

In this work, we pose the following research questions:

RQ1: How do players experience a speech-based VR game that features GenAI-based conversational NPCs?

RQ2: How do players perceive GenAI-based NPCs with distinct characteristics in a speech-based VR game?

Our findings indicate that the game provided players with a deeply immersive, enjoyable and subjectively novel experience. Players reported a heightened sense of freedom to express themselves due to the integration of GenAI in NPC interactions. However, common challenges with GenAI, such as hallucinations or irrelevant responses, disrupted player immersion and hindered gameplay, leading to frustration. This study provides new insights into the use of LLMs and GenAI technologies to support communication with NPCs in video games. We specifically examine the social dynamics of these interactions, focusing on how players engage with NPCs that embody distinct human characteristics and how these differences influence engagement, immersion, and overall experience. By identifying key challenges and potential benefits, we shed light on how LLM-based NPCs shape social presence, communication patterns, and relationship-building in virtual environments. The implications of our work offer guidance for game designers seeking to integrate speech interaction and adaptive dialogue into future games.

2. Related work

As the gaming industry continues to push for enhanced player experience, researchers and developers explore innovative approaches to elevate gameplay (Cairns et al., 2014). This section provides an overview of prior research on immersion in games, speech interaction in games, LLMs in game

development, and social interactions with virtual agents. We highlight theoretical and empirical gaps that motivate our research.

2.1. Immersion in games

Immersion is a central concept in game research and has been strongly linked to player engagement, enjoyment, and presence (Cairns et al., 2014; Leroy, 2021). VR, in particular, has been shown to increase the sense of presence in the virtual world and perceptual realism compared to traditional screen-based gaming (Winkler et al., 2020; Yao & Kim, 2019). The rapid expansion of the VR gaming sector highlights a major transformation in how games are designed and experienced (Zhang, 2023). As the technology becomes more accessible and affordable, big game franchises seize the opportunity to offer players immersive experiences, producing VR versions of popular franchises such as *Half-Life: Alyx* (Valve Corporation, 2020). Previous studies have further demonstrated the effectiveness of VR in domains such as education (Faiqotuzzulfa & Putra, 2023) and exposure therapy (Alexandrovsky et al., 2020; Mulvaney et al., 2024). In gaming and training contexts, VR environments with embodied interaction have been effective in simulating performance scenarios, such as public speaking (El-Yamri et al., 2019; Wolf In Motion Ltd, 2016). However, most VR research has focused on visual and embodied interaction, with less emphasis on spoken, social interaction with AI-driven characters. Although social VR settings and their use in psychology are still in their infancy, the dynamic social interactions in VR show promise as a tool to facilitate interventions (Ganschow et al., 2024; Mulvaney et al., 2024). A study by Guimarães et al. (2020) showed that social interactions with virtual agents in VR create a stronger sense of social presence than traditional platforms like desktop computers. Speech interaction is another technology that has been shown to enhance player immersion and the sense of presence in video games by allowing players to use an intuitive and natural form of interaction (Allison et al., 2018; Osking & Doucette, 2019; Zargham et al., 2022; Zargham, Fetni, et al., 2024; Zhao et al., 2018). While studies indicate that speech interaction has the potential to increase immersion (Zargham, Fetni, et al., 2024), very few studies have looked into the combination of speech, VR, and GenAI-based communication in a single system and analyzed their combined effects on player experience. Our study addresses this gap by situating speech-based interaction with GenAI-based NPCs in a VR adventure/puzzle game to examine how these technologies jointly affect immersion, player experience, engagement, and perceived freedom of expression.

2.2. Speech-based games

Even though speech technology has been around for decades, it has rarely been employed in video games (Allison et al., 2020), largely due to technological and hardware constraints (Kinoshita et al., 2020; Nima, 2024). However, with technological advances and the widespread availability of microphones in gaming devices, the video game industry has shown increasing interest in voice interaction (Allison et al., 2020). Game developers are now experimenting with this modality as a means of interacting with games (Allison et al., 2019; Carter et al., 2015). Although in most games using this feature, speech remains an optional or side feature, mainly used as a novelty rather than a core gameplay component (Allison et al., 2020). Empirical work has shown that speech-based interaction can enhance immersion and engagement, particularly when it aligns with the player's role or character identity (Allison et al., 2018, 2019; Suh et al., 2021; Zargham, Fetni, et al., 2024). A meaningful way to incorporate speech into games is through conversations with NPCs. Literature suggests that interacting with NPCs can evoke emotional responses and influence decision-making (Henrik, 2016). Well-designed NPC interactions can create a sense of social presence and companionship (Guimarães et al., 2022). When players hear NPCs respond directly to their spoken input, it can create a deeper emotional connection and foster empathy with these virtual characters (Bonfert et al., 2021; Chen et al., 2022; McLean et al., 2021; Nass & Brave, 2005), possibly leading to more meaningful and memorable gaming experiences. In a previous study, we found that integrating politeness-based speech mechanics into NPC interactions increased players' cognitive engagement and perceived realism (Zargham et al., 2025). The study also found that players adjusted their speech patterns based on NPC attributes such as

gender and hierarchy, suggesting that players adapt their behavior in response to real-world social norms reflected in the game. In a recent study, we explored the use of speech interaction in single-player video games, focusing on NPC interactions (Zargham, Friehs, et al., 2024). Topic experts recognized the significant potential of speech interaction to enhance immersion, engagement, and entertainment. However, concerns were raised about privacy and environmental limitations. However, a major limitation of most speech-based games is their reliance on pre-scripted or command-based dialogue systems. These systems restrict player expression, reduce conversational flexibility, and often lead to frustration when speech input falls outside predefined patterns (Zargham et al., 2020, 2025). While such approaches offer technical reliability, they undermine the potential of speech as a natural interaction modality. Recent studies further highlight players' preference for open-ended communication to enable greater engagement with in-game characters (Christiansen et al., 2024; Zargham, Friehs, et al., 2024; Zargham et al., 2024). Our work directly addresses this limitation by replacing rigid, command-based dialogue with open-ended, LLM-supported conversational interaction, enabling players to express themselves freely while maintaining narrative and gameplay coherence.

2.3. LLMs in game development

The recent advent of LLMs has significantly expanded the design space of conversational user interfaces (CUIs), enabling them to support users in more complex tasks and contexts with greater sophistication (Ruan et al., 2023; Yang et al., 2023; Zhao et al., 2024). Yet, LLMs are still relatively limited in speech-related applications (Min & Wang, 2024). In social VR platforms, studies such as Wan et al. (2024) and Pan et al. (2025) integrated GenAI-based agents into VRChat. These systems demonstrated improved contextual coherence and potential benefits for language learning and reduced speaking anxiety, while also highlighting challenges such as disrupted conversation flow and limited emotional support. However, these works primarily focused on social VR or educational settings, rather than structured video game contexts. In gaming contexts, Volum et al. (2022) and Rao et al. (2024) explored LLM-driven NPCs in Minecraft. While players appreciated the novelty and helpfulness of such agents, recurring issues emerged, such as NPCs generating out-of-character responses, hallucinations, or irrelevant dialogue, which disrupted immersion and task flow. Christiansen et al. (2024) compared open speech interaction with LLM-based NPCs to traditional dialogue trees in a mystery game. While unrestricted speech enhanced immersion and engagement, it also led to cognitive overload and confusion due to a lack of structure and occasional conversational breakdowns. More recently, conversational AI tools such as InWorld AI Inc. (n.d.), and Convai Inc. (n.d.) enable the development of GenAI-based NPCs for games, virtual worlds, and immersive environments. Yet, the evaluation of their impact on player experience, social presence, and immersion in controlled VR gameplay scenarios remains limited. Our work adds to the body of literature by empirically investigating how players experience LLM-based NPCs in a single-player VR game with structured objectives, moving beyond sandbox or social VR environments. We further analyze how different NPC characteristics influence interaction dynamics.

2.4. Social interactions with virtual agents

Social presence, the feeling of being with another social entity, has been identified as a key factor in immersive and engaging virtual experiences (Guimarães et al., 2020). Prior research shows that interacting with virtual agents using speech can elicit emotional responses, empathy, and a sense of perceived companionship (Guimarães et al., 2022; Henrik, 2016; Nass & Brave, 2005). In VR contexts, interactions with embodied agents have been found to produce stronger social presence than interactions on desktop platforms (Guimarães et al., 2020). However, in the context of video games, most NPCs rely on static scripts and lack the adaptive conversational capabilities needed for sustained social interaction. Recent studies suggest that LLM-based agents can support richer dialogue and potentially deeper social bonds (Garcia-Pi et al., 2023; Wan et al., 2024). Yet, empirical research specifically examining how players socially engage with GenAI-driven NPCs that have distinct human-like characteristics (e.g., gender, age, hierarchy) remains scarce, especially in speech-based VR games. Our work contributes to the understanding of these social dynamics by analyzing how players adapt their speech, behavior, and

Figure 2. The two game versions along with the corresponding NPCs for each level.

attitudes toward different NPCs, and how these interactions influence their sense of social presence, immersion, and relationship-building inside the game.

3. Game design

We developed a speech-based VR game titled "Office Whispers," an adventure/puzzle game which puts players in a virtual office environment with multiple rooms where they must use speech interactions to communicate with various characters. Players assume the role of an office employee who needs to locate a lost laptop for an important pitch presentation. Using speech interactions, players communicate with NPCs to gather clues and progress through the game. The game consists of four NPCs: the manager (Kai), the maintenance worker (Ali), the designer (Parker), and the intern (Aria) (Figure 2). By assigning each NPC a clear position within the office hierarchy, we aimed to observe how social dynamics influence players' communication patterns and interactions.

In our prior work, we observed that NPC attributes such as gender and hierarchical role can shape player perceptions and interaction styles (Zargham et al., 2025). Hence, in this work, we introduced a controlled gender-role manipulation. Specifically, we created two versions of the game where the same four hierarchical positions (manager, designer, maintenance worker, intern) were preserved, but the gender associated with each role was swapped. This manipulation was designed to explore whether such surface-level changes would lead to measurable differences in overall player experience, character perception, or preferences within a VR gameplay context. In the first version, the roles were filled by a male maintenance worker, a female intern, a male designer, and a female manager. In the second version, these roles were mirrored (Figure 2).

3.1. Mechanics

While playing "Office Whispers," players interact with the NPCs through speech commands. Players' movement within the game environment is facilitated through teleportation and directional movement using VR joysticks. Additionally, players can move physically through the room. However, to prevent

Figure 3. The chat box above NPCs highlights both the players' recognized input and the NPCs' statements.

collisions in the physical world, on top of drawing a virtual barrier, experimenters advised players to restrict their physical movement to no more than two steps in any direction. The game includes multiple interactable objects, some necessary to solve the game's puzzles. Examples of these items include mugs, keyboards, documents, donuts, or a computer mouse. Items could be grabbed through a grabbing motion by using palm triggers. Players are required to solve ten puzzles to complete the game. Each puzzle's solution is associated with specific NPCs: eight have solutions known only to individual characters, necessitating at least two interactions with each NPC. The remaining two puzzles can be solved by any NPCs, providing flexibility in how players approach these challenges. The puzzles can be solved in any order, allowing players to choose their preferred sequence of tasks.

The following is the list of puzzles and the corresponding NPC who could support the player to solve them:

- Determine why everyone in the office is annoyed (All NPCs).
- Discover the access code to the server room and restart the server (Ali).
- Identify who broke into the server room and inform the maintenance personnel (Parker).
- Find out what the gift for Parker's birthday is (Kai).
- Locate the birthday gift and deliver it to Parker (Ali).
- Find out what the second gift for Parker's birthday is (All NPCs).
- Locate the second birthday gift and deliver it to Parker (Aria).
- Locate Aria's USB drive and return it (Parker).
- Find out where your missing laptop is (Aria).
- Discover the keycode to the storage room (Kai).

We implemented a to-do list within the game to give players an overview of the puzzles and their progress. The list displays unsolved puzzles in white. When a puzzle is completed, its entry on the list changes to green. To converse with the NPCs, players had to stand close to the NPC and speak their inquiry out loud. When in close range, the character starts reacting to the players speech. The conversations between the player and the NPCs were live transcribed and displayed in a chat-like text box above the character, detailing the player's recognized input and the NPCs' statements.

This was meant as a visual aid to promote better accessibility and understanding of the conversations (Figure 3). Players were free to move around and interact with the environment during conversations. They also had the option to leave conversations at any time. When NPCs were unable to assist with a puzzle directly, they could provide hints by suggesting other characters who might be able to help the player solve the puzzle. However, these hints were not always accurate, as NPCs could

speculate about who might have the necessary information. Players could ask NPCs questions like, "Do you know who could help me with this?" to receive guidance on whom to approach for specific puzzles, which could encourage players to converse more with NPCs. The NPCs remained stationary at their designated locations within the office. Character walking was not implemented, as it was deemed unnecessary for the game's design and focus on conversational interactions. Moreover, the NPCs were limited to interacting with the player could not communicate with each other.

3.2. Implementation

The game environment and logic were created with Unity 3D[1]. We used Meta Oculus Quest 2[2] for the VR setup.

We used the InWorld AI platform (Inworld AI, 2024) to design and deploy the conversational NPCs in our game. InWorld is a commercial middleware system for interactive characters that integrates speech recognition, dialogue management, memory, and LLM–based response generation into a single pipeline. While the platform abstracts many underlying technical components (e.g., the specific LLM architecture), it enables developers to control NPC behavior through structured character profiles, world knowledge settings, and interaction logic. Each NPC in InWorld is defined by a "Core Description" detailing their motivations, flaws, and actions. Additionally, the NPCs can also be assigned individual attributes, including names, pronouns, roles, and interests. These configuration elements collectively function as persistent character prompts that guide the LLM's generative behavior during player interactions. To ensure comparability across characters, we kept core traits such as motivations, interests, and background consistent. However, variations in gender, voice, and office roles were introduced as part of our research goal to investigate the social dynamics of interactions with NPCs. All NPC configurations were created and managed using InWorld's character editor. Speech input from players was captured via the Meta Oculus Quest 2 microphone and processed using InWorld's built-in speech recognition pipeline. At the time of development (May–July 2023), the speech recognition engine and model configurations were not externally configurable or publicly documented by InWorld.

4. Study design

We conducted an exploratory user study to assess players' experiences when interacting with GenAI-based NPCs through speech in a VR game. To broaden the ecological validity of our findings and explore whether interaction patterns differed across geographically distinct contexts, we collected data at two sites, one in Canada and one in the Netherlands. Prior work suggests that communication styles and comfort with verbal interaction may vary depending on participants' social and linguistic backgrounds (Nass & Brave, 2005; Zargham, Friehs, et al., 2024). Conducting the study at two international universities allowed us to explore whether consistent patterns emerged across different participant pools. Further, the two sites provided access to participants with diverse linguistic and cultural backgrounds, which is particularly relevant given that our system relies on spoken interaction with AI-driven characters. Rather than aiming to conduct a formal cross-cultural comparison, we aimed to examine whether players interacting with the same system in two distinct settings exhibited broadly similar or divergent patterns of experience, communication, and perception of GenAI-based NPCs.

4.1. Procedure

The study was conducted in controlled lab environments at both sites. Each location was spacious enough to support a VR play area with a minimum size of 3 meters by 3 meters. This ensured that participants had sufficient space to move safely and interact with the virtual environment without too many physical constraints. Players could also play the game seated. An experimenter was present throughout each session to ensure participant safety, provide technical assistance when needed, and take observational notes on participant behavior and verbal remarks. Before beginning the experiment, all participants were given a written informed consent form. The form described the purpose of the research, the study procedure, potential risks (e.g., motion sickness, cognitive fatigue), data collection

methods, including audio recording, and compensation details. They were also informed they could withdraw from the study or stop playing the game anytime. Participants were also informed that all collected data would be anonymized and handled confidentially. After providing consent, participants received a written briefing describing the VR game narrative, central goals, and interaction possibilities, along with a list and images of the NPCs they would encounter. Following the briefing, participants were instructed about the game controls and were given the opportunity to familiarize themselves with the VR setting and controls before beginning the actual game session. Upon the introduction phase, the experimenters informed them of a 20-minute gameplay time limit. This time limit was set to ensure consistency across the study and mitigate the risk of fatigue. Shorter sessions could also reduce the risk of motion sickness. After the play session, participants were asked to fill out the post-exposure questionnaires. A semistructured interview, which was audio recorded, was held at the end of the session. The interviews took an average of 7.54 min ($SD = 1.98$). Each session lasted approximately $40 - 55$ min. Our study protocol was reviewed and approved by the University of Waterloo Research Ethics Board (REB #45294) and the University of Twente Ethics Committee (#230076).

4.2. Pretest

We conducted preliminary tests with two participants before running the main study. This aimed to identify technical problems related to the game and challenges with speech recognition, puzzles, and the overall study procedure. Several gameplay issues were identified during these sessions, and appropriate measures were taken to address them. One key piece of feedback was that the NPCs' responses were too long, which participants found frustrating. In response, we shortened the NPC dialogues, limiting each response to a maximum of 50 words. Additionally, based on feedback about the evaluation length, we reduced the number of questions to streamline the process.

4.3. Participants

Participants were recruited in two different locations as part of an internationally joined research effort. The two locations were Canada and Netherlands. We used a convenience sampling approach for participant recruitment through word of mouth, university mailing lists, and intranet structures of university facility-wide shared research subject pools. Participation in the study was voluntary, and participants each received 20$ in cash for taking part in the study. We recruited $N = 48$ individuals between 18 to 52 years ($M = 23.58$, $SD = 6.15$). 21 participants self-identified as male (43.7%), 26 as female (54.2%), and one as diverse (2.1%). We recruited 20 participants in Canada and 28 participants in the Netherlands. Regarding residency, 20 participants (41.7%) were from Canada, 17 (35.4%) from the Netherlands, nine (18.8%) from Germany, and two (4.2%) from other countries. All participants had prior experience with video games. 19 people (39.6%) played video games frequently (seven daily and 12 several times a week), while 29 participants (60.4%) reported not playing video games often (nine people once per week and 20 once monthly). While 40 participants (83.3%) had prior experience with voice-controlled applications, only 7 participants (14.5%) had previously played a voice-controlled video game. Requirements for inclusion in the study were an age above 18 years and English language proficiency.

4.4. Measures

We used a mixture of questionnaires to assess player experience.

The post-exposure questionnaires included demographic questions, the Player Experience Inventory (PXI) (Abeele et al., 2020), and a custom-designed questionnaire aimed at evaluating players' experience with the game and their interactions with the NPCs. The custom-designed questionnaire consisted of seven questions, covering participants' favorite and least favorite characters, the attractiveness of these characters, and their enjoyment of interacting with individual characters. The questionnaire also included questions about players' perceived performance, willingness to play similar games, and overall game experience, all of which were indicated on seven-point Likert scales. To facilitate recall, questions

regarding the featured NPCs were supplemented with pictures of the corresponding NPCs. Additionally, participants were asked to partake in a semi-structured interview to further evaluate the qualitative aspects of the player experience and individual preferences (Wilson, 2013). The interview questions were specifically crafted by the research team for this study, aiming to capture participants' experiences with the central element of voice interaction with NPCs. The interview consisted of ten questions, covering likes and dislikes about the game, the most and least interesting aspects, the most attractive aspects, the players' experiences interacting with NPCs, their opinions about individual NPCs, and their overall experience with speech interaction. After the interview, participants were debriefed, compensated, and given the opportunity to ask further questions related to the research topic. The custom-designed questionnaire and the interview questions can be found in the Supplementary Material.

4.5. Data analysis

4.5.1. Quantitative analysis

The analyses of quantitative data were performed with Python (3.11) using the Python data stack, including pandas (The pandas development team, 2020), numpy (Harris et al., 2020), and pingouin-stats (Vallat, 2018) packages. To determine if the data met the normality assumptions of parametric tests, we visually inspected the distributions of the data on Q–Q plots. All collected data fell approximately on a line ($R^2 = [0.89 - 0.98]$), indicating that the data was normally distributed. Hence, for our analysis, we mainly employed parametric t-tests to examine if differences between groups were significant. For the significance tests, we used an alpha level of $\alpha = 0.05$.

4.5.2. Qualitative analysis

Post-study interviews were transcribed using Amberscript (2024)[3]. Recordings were labeled with participant IDs, and no personally identifiable information was included in the file metadata or spoken identifiers. After transcription was completed, the audio files were deleted from the platform, and only the anonymized text transcripts were retained for analysis. The interview data were analyzed using a domain summary approach (Braun et al., 2019; Connelly & Peltzer, 2016), where themes were organized around a common topic rather than a shared meaning. This approach aimed to capture the diverse interpretations and perspectives related to a specific subject or area of focus (Morgan, 2022). The analysis began with data familiarization and categorization (Braun & Clarke, 2019). Two researchers independently reviewed the responses to achieve an understanding of the content. This process helped identify emerging patterns, ideas, and concepts in the participants' answers. To develop the coding scheme, a randomly selected subset of ten interviews was initially coded by one researcher informed by relevant literature and insights from earlier studies. The resulting preliminary codes were reviewed and iteratively refined in discussion with the co-researchers and two external HCI experts. After consensus on the codebook was reached, one author applied the finalized coding scheme to the full dataset using ATLAS.ti (version 24.1.0) [4]. While formal interrater reliability metrics were not computed due to this workflow, reliability and consistency were supported through collaborative codebook development, expert review, and iterative discussions during the analysis process. Both latent (interpreting the commenters' intended meaning) and semantic (verbatim participant words) codes were generated. Responses could be assigned multiple codes where applicable. Each code was recorded as occurring once per answer, with additional occurrences noted if the sentiment was repeated in responses to later questions. This inductive coding process aimed to produce a codebook that reflected the frequency of participants' responses and the salience of their sentiments, especially when participants emphasized certain aspects across different questions. This process led to extracting key insights and findings from the analyzed responses. These insights are presented in Subsection 5.4. The codebook can be found in the Supplementary Material.

5. Results

This section presents the key results of the user study, beginning with the quantitative survey data analysis, followed by the qualitative findings from interviews and observations during the study sessions.

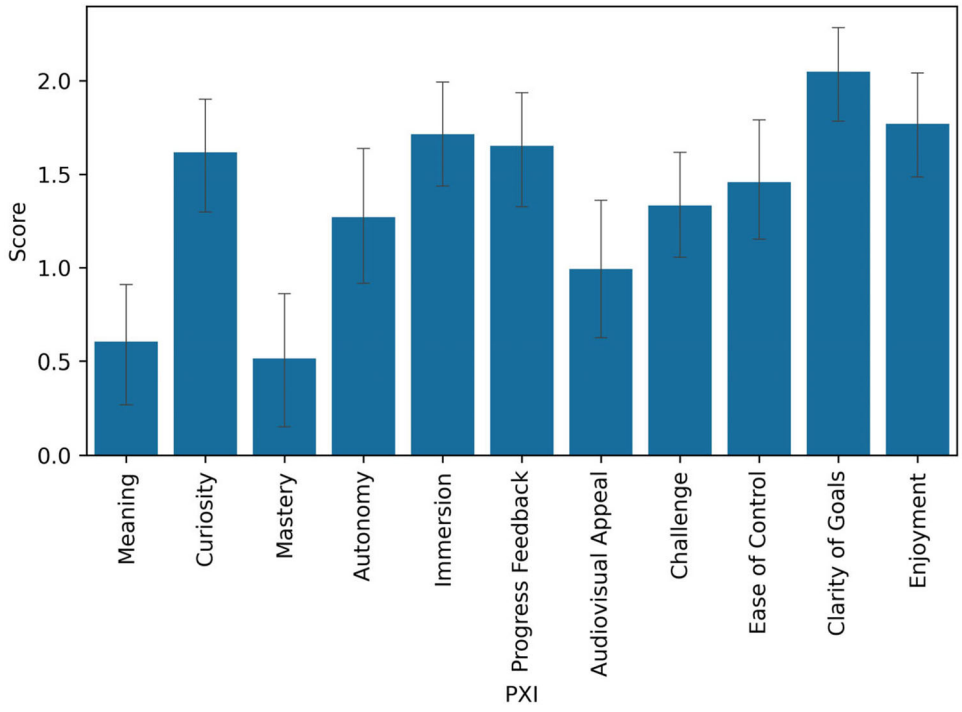

Figure 4. Mean scores of the PXI questionnaire subscales.

5.1. Player experience

The player experience was assessed using the PXI on a 7-point Likert scale ranging from −3 to +3, with 0 as the neural point (Abeele et al., 2020). The overall mean scores for all subscales of the PXI were rated above neutral (Figure 4), with the highest score found in the "Clarity of Goals" construct ($M = 2.049$, $SD = 0.88$) and the lowest score on the "Mastery" ($M = 0.51$, $SD = 1.33$). To compare the player experience of both game versions, we conducted independent ttests on each subscale of the PXI between the game versions, which showed no significant differences ($p > 0.05$). The Bayes Factors were in the range $BF10 = [0.29 - 1.41]$ indicating low evidence for the alternative hypothesis and suggesting equivalent experiences in both versions (Figure 5).

Further, we checked if the PXI scores differed between the two experiment sites and the participants' gender. Between the sites, only "Clarity of Goals" was rated significantly higher in Canada than in the Netherlands ($t = 3.394$, $p_{bonf} = 0.01$, Cohen's $d = 0.92$, $BF10 = 23.19$). We did not observe significant differences between the participants' genders on the PXI ($BF10 = [0.30 - 1.11]$).

5.2. Custom survey questions

Regarding the custom-designed questionnaire, the participants rated their overall game experience with a mean score of 5.75 ($SD = 0.93$), suggesting a generally positive experience. Likewise, participants' willingness to play similar games had a mean score of 5.73 ($SD = 1.33$), suggesting a positive attitude toward engaging with games of a similar type in the future. The players' self-rated performance showed a mean score of 4.27 ($SD = 1.27$), indicating that participants rated their performance rather positively. This measure reflects subjective self-assessment rather than objective task success or efficiency. As objective performance metrics such as puzzle completion time or success rate were not recorded, these findings should be interpreted as reflecting differences in perceived competence rather than measurable gameplay performance. We performed independent t-tests to examine differences in ratings between the game versions, the experiment locations, and the participants' genders. We only observed a significant difference in self-rated performance between the two experiment sites ($t = 2.39$, $p = 0.02$, Cohen's $d = 0.68$, $BF10 = 2.77$). All other tests showed no significant differences with small Bayes Factors ($BF_{10} < 0.51$), suggesting low evidence for differences between the groups.

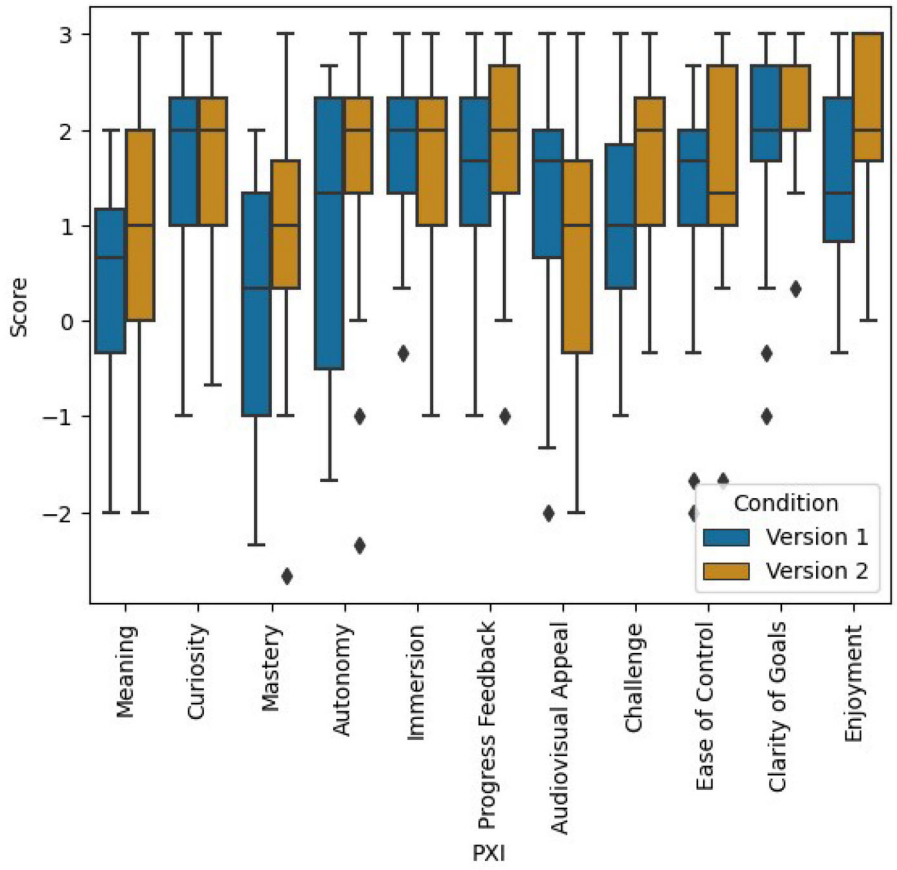

Figure 5. Mean scores of the PXI questionnaire subscales for both game versions.

5.3. Ratings of the NPCs

The Designer (Parker) was rated as the most favorite character, closely followed by the Manager (Kai) and Maintenance (Ali). Meanwhile, the Intern (Aria) was rated least favorable. Similarly, the Intern (Aria) was rated the most disliked character, closely followed by the Manager (Kai). The ratings are depicted in Figure 6. The distribution of the favorite and disliked characters between the two versions can be seen in Figure 7.

We analyzed the relationship between participants' gender and their choice of favorite characters. Female participants showed a balanced preference, selecting male and female characters equally (13 males, 13 females), whereas male participants tended to slightly favor female characters more frequently, with 13 choosing female characters compared to eight selecting male characters. A similar pattern emerged for disliked characters: female participants again showed a balanced distribution (13 males, 13 females), while male participants were more likely to dislike male characters (12 males, 9 females).

5.3.1. Reasons for favouring characters

Participants selected their favorite characters for various reasons. Twelve favored the character's helpfulness and ability to provide accurate information as the primary factor, with one noting, "She usually had the right information or knew what to do" (P24). Seven valued politeness and friendliness, while four appreciated characters who greeted them warmly. Three selected characters based on relatability, favoring characters that reminded them of themselves or someone they knew in real life. One player described, "He talked in the same cadence as some people I know personally" (P09). Three participants preferred NPCs who provided short, direct responses, while three others favored those who seemed more approachable, though the specific character varied among them. Additionally, three players chose characters who lied or hid facts, finding this behavior fun. Conversely, two participants selected their

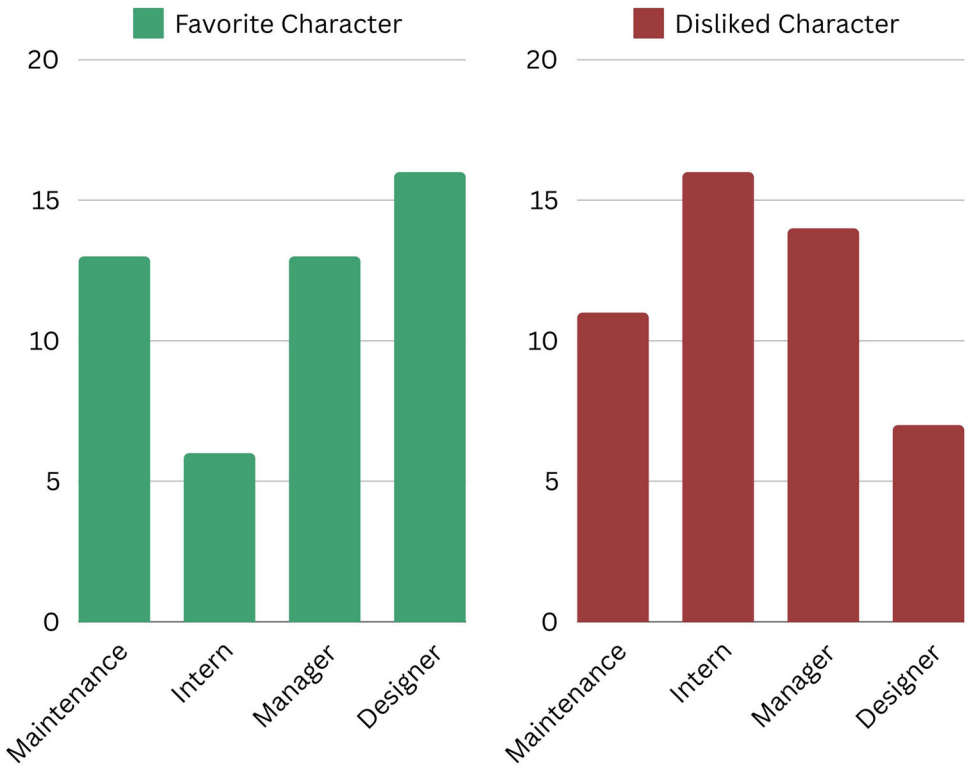

Figure 6. Frequency of characters selected as favorite and most disliked in the game (combined game versions).

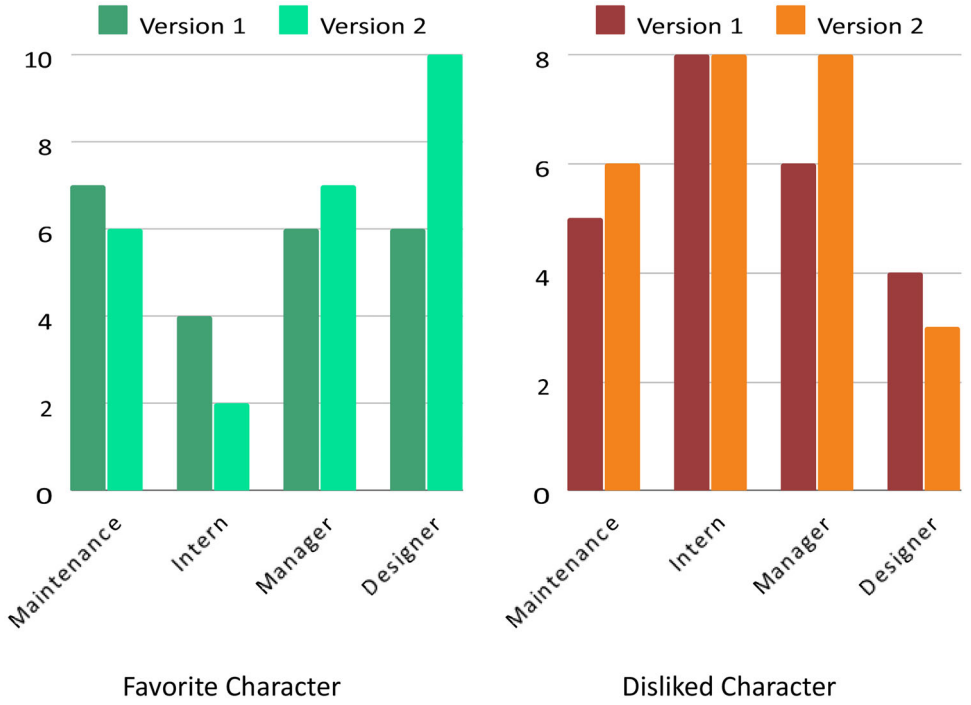

Figure 7. Frequency of characters selected as favorite and most disliked in the game divided by game versions.

favorite character for their honesty. Attractiveness and physical appearance influenced the choices of the three participants: "Parker seemed cute and kind" (P44). Additionally, three participants liked Parker specifically due to the game story, as it was their birthday. They perceived this character to be happier: "Parker was in a good mood because of his birthday" (P34). Other reasons for selecting a favorite character included humorous responses, being the least annoying, seeming realistic, appearing intelligent, or having more interactions with the player.

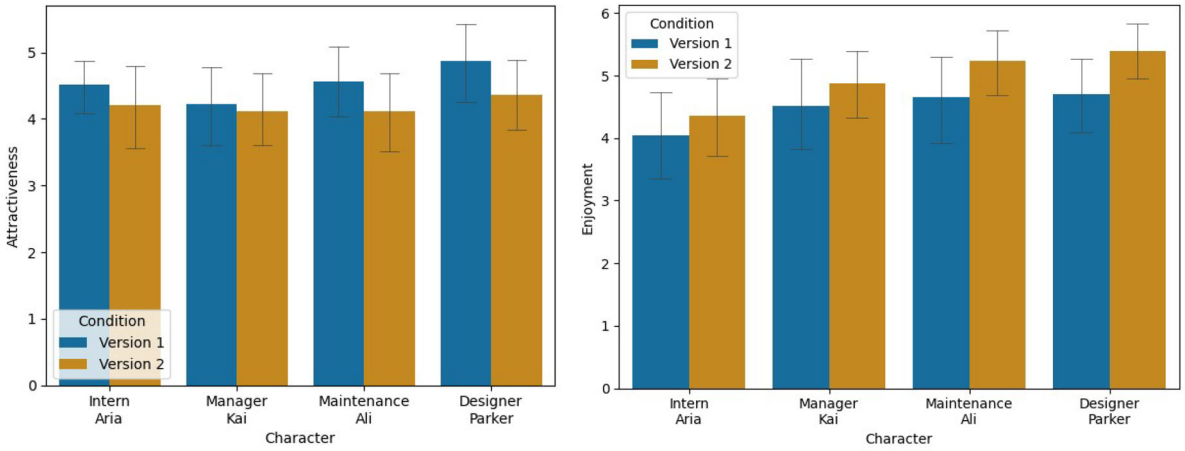

Figure 8. Mean scores of character attractiveness and interaction enjoyment.

Table 1. Pearson correlations between rated character attractiveness and enjoyment of interaction with the character.

Character	Pearson's *r*	*p bonf*	*BF*10	$1-\beta$
Intern – Aria	0.385	0.028	6.162	0.783
Manager – Kai	0.457	0.004	32.133	0.917
Maintenance – Ali	0.590	<0.001	2268.387	0.996
Designer – Parker	0.565	<0.001	874.909	0.991
Total	0.566	<0.001	896.614	0.991

5.3.2. Reasons for disliking characters

Sixteen cited a lack of helpful information, with one saying, "He didn't contribute much to the tasks of the game" (P24). Six felt certain characters were withholding information, as one player remarked, "They did not answer questions they were meant to know" (P20). Five participants disliked characters they perceived as dishonest and three participants mentioned receiving contradictory information from the characters as the main reason for disliking them. Impoliteness was another factor, with four players feeling their interactions with specific characters were rude. One user commented, "It felt like he antagonized me for no reason but did not tell me if I did anything wrong. Maybe he had a bad day" (P16). Other reasons included minimal interactions, long responses, unappealing tone of voice or speech patterns, misrecognition, visual design and appearance.

5.3.3. Attractiveness and enjoyment of interactions

When assessing individual characters' attractiveness, we conducted independent *t*-tests to examine differences in ratings between the game versions, the experiment locations, and the participants' gender. No significant differences were found in attractiveness ratings or interaction enjoyment ($p > 0.05$, $BF10 = 0.06$) across characters (Figure 8).

We further explored participants' ratings of perceived attractiveness and enjoyment of interactions by conducting a Pearson's correlation. The analysis shows strong correlations between attractiveness and enjoyment for each character. Additionally, we combined each participant's attractiveness and enjoyment ratings to calculate total appeal across the NPC characters. The results of the correlation analysis are displayed in Table 1.

5.4. Interview findings

The responses from two participants were excluded from the analysis due to technical issues with the audio recordings. The remaining interview responses were analyzed and coded, resulting in a coding scheme with five main themes: *Player Experience and Impressions*, *Game Design and Mechanics*, *Virtual Reality*, *Speech Interaction*, and *NPC Design and Interaction*.

5.4.1. Player experience and Impressions

All but one participant expressed positive feedback regarding the speech interaction as a modality to interact with the game characters. Specifically, 28 participants (58.3%) liked speech interaction with NPCs the most about the game, and 22 participants (45.8%) found it to be the most interesting part of gameplay. Additionally, eight participants (16.6%) noted that this feature was the attractive part of the game. One player remarked, "I thought speech interaction with the NPCs was really, really good. It makes it feel more personal in a way. I honestly think a much deeper connection can be formed with the characters" (P13). 56.2% of the participants highlighted the novelty of using speech interaction to converse with NPCs in the game. One participant noted, "It was unlike anything I've ever experienced before. This was very novel and fluid" (P2). The game's immersive quality was emphasized by 28 participants (58.3%), with 15 (31.2%) explicitly mentioning that speaking to the game characters enhanced their sense of immersion. Thirteen participants (27%) noted that the combination of VR and speech interaction significantly contributed to this immersive experience. One player commented, "With the VR glasses, it makes you feel like you are really talking to that person" (P7). Another added: "You are already wearing the VR glasses, so you are very immersed in the game, but talking to them just makes it feel more real" (P20). However, two participants noted that occasional lapses in conversation realism detracted from the immersion. One player said, "Maybe in an actual workplace, if I ask a question and they are like, 'I'm sorry, I don't understand that,' I'd be like, 'screw you'" (P43).

Nine participants (18.7%) praised the game's realism, with one participant saying, "It seemed so real. It just seemed like you were there. You just felt part of the game" (P29). Many participants found the speech interaction with NPCs to be enjoyable and engaging. Fifteen participants (31.2%) described these interactions as fun and entertaining, with four (8.3%) specifically finding them humorous. Players used various positive descriptors to convey their experience with the speech interaction, including "enjoyable" (16 mentions), "interesting" (11 mentions), "nice" (8 mentions), "cool" (4 mentions), "good" (5 mentions), "natural" (3 mentions), "intuitive" (2 mentions), "relaxing" (2 mentions), "impressive" (2 mentions), among others.

On the negative side, three participants found the speech interaction awkward or weird, with one player noting that this type of game was not their preferred style.

Freedom of expression during interactions with NPCs emerged as a key theme, with 16 participants (33.3%) appreciating this feature for providing them greater control and agency in the game. One participant shared, "It gives you a lot of agency—what you want to do and your plan of action. It feels more like you control the game instead of the game controlling how you are supposed to play it" (P20). Another added, "I liked that you could really talk to them. Sometimes, you have games where you can only say certain things, but you are quite limited. In this game, I feel like they could answer everything you say. It's what people are like" (P17).

Few participants also saw potential in games utilizing speech interaction for educational or therapeutic purposes. One player suggested, "You can practice different languages with people or practice saying things out loud. Those things that you find difficult to speak about out loud in your real life" (P28). Another player added that such games can help with stress relief: "This kind of game can help relieve stress when you are having a very bad day. You can talk to someone" (P39).

5.4.2. Game design and mechanics

The esthetics and design of the game received favorable feedback from 18 participants (37.5%) who appreciated the visual appeal. On the other hand, three participants disliked the overall esthetics (6.25%) Two participants noted that the poor visual quality negatively impacted their immersion. One player commented, "The graphics were very game-like. You do not dive into it and feel like it is another reality. You know, it is a video game, and it puts you down to earth" (P21).

Feedback on the game's puzzles was generally positive, with 17 participants (35.4%) enjoying solving the tasks presented. Eight participants (16.6%) specifically highlighted that the puzzles were the game's most interesting aspect, and two appreciated the variety of the tasks. However, not everyone shared this sentiment—six participants (12.5%) did not enjoy the puzzles overall, and four mentioned that they

disliked a specific puzzle (8.3%). Four participants (8.3%) found the puzzles to be the least interesting part of the game.

Seven players (14.5%) saw navigating back and forth between characters as cumbersome, with three (6.25%) specifically noting that this was the least interesting part of the game. Moreover, four participants (8.3%) found the gameplay time too short. Two noted that it was difficult to fully grasp the NPCs' characters within the 20-minute gameplay period.

5.4.3. Virtual reality

The feedback on VR controls was mixed. Thirteen participants (27%) found the controls cumbersome and challenging to use, while eight (16.6%) appreciated the VR controls, describing them as intuitive and simple. Similarly, opinions on the virtual environment were also divided. Twelve participants (25%) found the virtual environment appealing, noting that it contributed positively to their overall experience. However, eleven participants (22.9%) felt that the environment was not particularly interesting. Six participants (12.5%) highlighted navigation within the VR space as complex and found it challenging to move through the game.

Conversely, three participants (3.2%) found the movement within VR to be exciting, adding to their sense of immersion. A concern raised was regarding motion sickness, with four participants (8.3%) reporting mild symptoms. Fortunately, none of these cases were severe enough to detract significantly from the overall experience.

5.4.4. Speech interaction

Thirty-seven participants (77%) reported experiencing speech recognition issues during gameplay. Interestingly, 24 of these participants (50%) attributed the problems to themselves, citing unclear speech or accents as the cause rather than the recognition system itself. Sixteen participants (33.3%) noted that their pronunciation difficulties led to misrecognition of their intents. On the positive side, 13 players praised the speech recognition system, highlighting its ability to accurately recognize their intents. However, two participants noticed a slight delay in NPC responses, affecting their experience. Five participants (10.4%) pointed out issues with turn-taking in conversations, where NPCs interrupted the player, making it difficult for the player to stop them. Additionally, four players mentioned putting in extra effort, such as enunciating, speaking louder, or slowing down, to ensure that NPCs understood them. One player compared this to "talking to a baby" due to the need for exaggerated enunciation (P15).

Nine players (18.7%) appreciated the ability to directly speak to characters without having to select dialogue options, finding it more immersive and realistic. Three participants (6.2%) liked not having to press extra buttons when talking to characters, though two others recommended a "hold to talk" mechanism to avoid unintended recognition.

One of the most prominent issues reported was the NPCs providing unhelpful responses, such as generic statements like "Sorry, I don't know about that!" Fourteen participants (29.1%) expressed frustration with these replies, leading some to perceive the NPCs as "not intelligent enough." Eight participants (16.6%) highlighted instances where NPCs gave out-of-context responses that did not align with their inquiry or game's tasks.

Moreover, 13 participants (27%) mentioned that NPCs sometimes provided incorrect information or "hallucinated" facts, hindering their game progression. Four participants (8.3%) found these madeup responses the least interesting aspect of the game. One player shared, "It was challenging because I was always questioning, 'Did he lie? Did he not lie?'" (P8). Despite the issues, six participants (12.5%) appreciated the uncertainty in NPC responses, with four (8.3%) finding it the game's most interesting aspect. Two found these events playful and humorous, with one player noting: "It didn't feel really artificial when the NPC was lying to me. It just felt more human" (P25).

Fifteen participants (31.2%) felt that conversations with NPCs were somewhat artificial, with seven (14.5%) describing them as robotic. One player stated, "It feels less like I'm having a conversation with a human being, but rather with just a response machine" (P2). Five participants (10.4%) found the NPC responses dry or flat, and two others wished for more organic, humanlike conversations.

Participants expressed several desires for more natural NPC interactions. Eight players (16.6%) suggested adding more conversational formalities to enhance realism. One player noted, "They didn't really keep the conversation going to get more information … In real life, people would at least try to help or ask more questions to show they care" (P11). Two mentioned that the NPCs always got the last word, which damaged the realism of the interaction. One player said, "They always ended conversations with sentences like 'If I could help you with something, just ask me again.' That made it seem programmed" (P16). Three players hoped for more context-aware NPCs, wishing them to know about past and present events by the player and other NPCs, as well as a better spatial understanding of the game's world and location of objects and characters. One player added, "The characters should have a more natural response to my behavior. Where I'm actually standing in the room, how I'm talking to them. It will convince me to respect the character even more" (P33).

Three participants recommended requiring more set-up or chitchat before directly asking NPCs questions. One player mentioned, "I thought I needed to have some opening sentences before getting to the point. But after a while, I realized I could directly ask the question. In the real world, that doesn't work. You have to introduce yourself or talk a bit and then ask your questions" (P37). Two other players indicated that they took advantage of the fact that they did not have to keep conversational formalities in the game, with one player saying, "I didn't have to close the conversation. Once I got the information I needed, I could just turn around and leave because I was aware that I was playing a game" (P42).

5.4.5. NPC design and interaction

Thirteen participants (27%) perceived that each NPC had a distinct personality, with five specifically appreciating this uniqueness. However, four participants (8.3%) felt that the NPCs' personalities were too similar, and two expressed a desire for more variety and enhancement in this area. Seventeen players (35.4%) mentioned that they found certain characters more interesting than others, highlighting a disparity in character appeal. Four players (8.3%) appreciated the diversity among the characters, recognizing it as a positive aspect of the game. Twelve participants (25%) noted that conversing with the NPCs initially felt weird, but they became more comfortable as the game progressed.

Three participants (6.2%) noted that they interacted differently with the characters based on their roles, suggesting a hierarchical interaction format. One player shared, "I sort of interacted with them based on how I would in a real office. Like with the manager, I would probably ask just direct questions, whereas the others, I might be more relaxed towards" (P31).

When asked about the most attractive aspects of the game, only two participants mentioned specific NPCs. Both were male participants who selected a female NPC. The lack of movement among NPCs was an important point of criticism. Seven participants (14.5%) wished the characters could move, with one finding this the least interesting aspect of the game. Participants also expressed dissatisfaction with the characters' bodily gestures and facial expressions. Six players (12.5%) wished for improvements in these areas, while three (6.2%) found the NPC gestures creepy. One player elaborated, "[The NPCs] didn't have a very natural backchanneling. When you talk to real people, they blink, move their heads, and say 'mhm' once in a while" (P33).

Three participants (6.25%) hoped for more emotional expression from the NPCs, noting that the lack of emotion made the NPCs feel less natural. Regarding the NPCs' voices, five participants (10.4%) found them monotonous and lacking emotion, expressing a desire for more modulation. Conversely, three participants (6.2%) found a specific character's voice engaging. Six participants (12.5%) mentioned a lack of proactivity in NPCs, with one player suggesting, "The conversations would be a little bit more natural if they also approach you and you do not only approach them" (P24). Additionally, two participants wished for cross-conversations between NPCs to enhance the game's realism, and two others wanted more support and prompts from NPCs to guide them toward progress.

5.5. Experimenters' observations

Throughout the study sessions, experimenters observed various patterns and events related to the game, NPCs, and participant behavior. A common behavior observed was that when the game failed to

recognize their voice input, many participants would instinctively move closer to the in-game characters, even if they were already within the communication range set in the game. Conversely, some participants backed away when they felt their voices were perceived as too loud.

A recurring issue involved NPCs failing to provide the necessary information, even though they were the correct character to ask. Instead, NPCs sometimes gave vague or incorrect answers or made-up facts (hallucinations), disrupting player progress. This inconsistency led to players receiving contradictory answers from different NPCs, which caused confusion and made players speculate about which character could be trusted. In a few instances, the NPCs did not stay in character and deviated from their designated roles. This was mainly caused by the speech recognition system's misinterpreting user intents. In one instance, the words "clips" and "mailing" were recognized as "ships" and "sailing." From this instance forward, the NPC began addressing the player as a ship captain, using pirate-themed jokes and metaphors such as "I hope you find that treasure!" or "Good luck on the office seas!"

Furthermore, we noticed that, for a few participants new to virtual reality, the novelty of the technology overshadowed other aspects of the game, including speech interaction. In post-session interviews, they focused more on the immersive VR environment, controls, and movement, paying less attention to speech mechanics and game design, as their first-time VR experience dominated their attention.

6. Discussion

This explorative study investigated how players perceive and experience speech interactions with GenAI-based NPCs in a VR game and how they communicate with characters displaying various humanlike traits. Overall, participants had a positive experience playing "Office Whispers," as reflected in the PXI ratings, custom-designed questions, and interview insights. Most participants enjoyed the social aspects of the game and had a favorable experience interacting with the NPCs through speech. While the game's esthetics, puzzles, and design were generally well-received, improvements in graphics and narrative were requested to enhance overall immersion and player engagement. Some felt that the lo-fi graphics hindered their immersion, and the office-based premise lacked excitement for others. This could be due to the mundane or repetitive nature of an office-based setting in contrast to more imaginative or fantastical environments. To engage a broader audience, games might explore varied narrative environments and settings to increase appeal.

Eventually, we interpreted our findings to provide answers to the following research questions:

RQ1: How do players experience a speech-based VR game that features GenAI-based conversational NPCs?
RQ2: How do players perceive GenAI-based NPCs with distinct characteristics in a speech-based VR game?

6.1. Player experience

The PXI results suggest that players had an overall positive experience. This is further supported by the participant's rating in the custom questionnaire concerning the overall experience and willingness to play similar games. These results were mostly equivalent between the two game versions, participants' gender, and the experiment sites, indicating that the yielded experience was stable and unaffected by these factors. The perception of characters, including their attractiveness and the enjoyment of interacting with them, was also consistent across both game versions, participants' gender, and experiment sites.

The study was conducted at two sites to explore whether player interactions and experiences remained consistent across distinct institutional and geographical contexts. Overall, we observed largely comparable patterns of interaction and experience across the two sites. The only statistically notable differences emerged in the PXI dimension "Clarity of Goals" and in players' self-reported performance in the custom questionnaire, where participants from the Canadian site reported higher scores. The Canadian sample consisted primarily of native English speakers, whereas the Netherlands sample included a higher proportion of non-native speakers. Given that the game relies heavily on spoken interaction and verbal comprehension of objectives, differences in English proficiency may have

influenced participants' understanding of the game goals and their perceived performance. However, these differences should be interpreted with caution, as they may have been influenced by other potential location-based confounding factors.

Immersive Gameplay and Realism: We observed that participants found the game highly immersive, as evidenced by their ratings on the PXI *Immersion* subscale and interview feedback. One example of this immersion was participants instinctively moving closer to NPCs to improve voice recognition, showing an intuitive adjustment to the interaction. While much of this immersive experience can be attributed to the virtual reality technology (Ishiguro et al., 2019), interviews revealed that interacting with NPCs through speech significantly contributed to players' sense of immersion, with onethird of participants noting that it made them feel more integrated into the game's world. This supports previous research, suggesting that speech interaction in games can enhance the sense of immersion (Osking & Doucette, 2019; Zargham, Fetni, et al., 2024, Zargham, Friehs, et al., 2024, Zargham et al., 2025; Zhao et al., 2018). Additionally, the freedom to choose the words to interact with NPCs—enabled by the use of GenAI-based characters—further enhanced players' immersive experience, supporting the idea that incorporating GenAI-based NPCs can amplify the immersive qualities of VR experiences (Christiansen et al., 2024). Overall, our results suggest that combining VR with speech interaction not only maintains but enhances immersion. Integrating VR and speech technology allows for more natural and intuitive interactions, validating the previous work (Christiansen et al., 2024; Zargham et al., 2020). The increased immersion and entertainment, along with positive feedback on the player experience, underscore the potential of speech-based VR games to deliver novel and engaging experiences.

However, the player's immersion proved rather fragile. Speech recognition errors, turn-taking errors, and GenAI hallucinations quickly broke the immersive experience. Speech interaction was shown to heighten engagement while amplifying the perceptibility of system flaws. Although several participants praised the game's realism, particularly in terms of interaction and the overall game world, the limitations in visual design, character movement, gestures, and facial expressions, as well as the NPCs' lack of full contextawareness, negatively impacted immersion and realism. Several players reported that NPC responses failed to align with the events in the environment or with their prior actions, which they felt reduced immersion and narrative coherence. Furthermore, the emotional expression of NPCs was another area for improvement, with several participants describing their voices as monotonous. Such factors disrupt the interaction flow, reminding players of the artificial nature of NPCs. Addressing these limitations is crucial for maintaining realism and immersion.

6.1.1. Speech interaction

Even though the majority of participants (77%) reported issues with speech recognition, most players responded positively to the speech interaction feature in the game, with 58.3% identifying it as the most liked aspect and 45.8% finding it the most interesting part of the gameplay. The novelty of this feature was another key factor, with over half of the participants (58.3%) emphasizing that they had never experienced such a mechanic in a game before. This points to the scarcity of speech interaction in video games, a point also highlighted in previous literature (Allison et al., 2019; Zargham, Friehs, et al., 2024). Beyond novelty, this suggests that speech interaction currently represents a gap in mainstream game design, making even imperfect systems feel innovative and agency-enhancing to players.

The ability to speak directly to NPCs provided a new and immersive way to interact with game characters, as highlighted by our participants' remarks that speech interaction "makes it feel more personal" and allows for a "deeper connection" with the characters. These reactions align with theories of social presence and parasocial interaction, indicating that people may perceive agents more as social actors when communication resembles human-human dialogue (Seaborn et al., 2022). While many players found speech innovative, technical limitations detracted from the experience for several. Interestingly, half of the participants attributed the recognition problems to their own unclear speech or accents, indicating that players were somewhat forgiving of these flaws. Nevertheless, speech recognition issues, turn-taking problems, and lack of proactive interactions were raised as limitations of speech interactions with NPCs. These challenges forced players to over-enunciate, wait their turn to speak, and exercise more patience, disrupting the natural flow of conversation. Many of these technical concerns have

been researched, and efforts are currently being made to address these to offer more seamless user experiences (Reicherts et al., 2021; Skantze, 2021; Zargham, Fetni, et al., 2024). Interestingly, some players did not attribute the "mistakes" of the NPCs to technical errors but rather to the character traits of that NPC. Some participants stated that NPCs were "lying" even though the NPCs were not programmed to lie, and they should have behaved equally for all participants. This fits past research that humans tend to attribute intention and meaning to actions performed by artificial actors (Nass & Brave, 2005).

6.1.2. GenAI for communication

Participants found interactions with the GenAI-based NPCs more organic and less scripted than typical NPC dialogues in most games. This stemmed from two key factors: the direct natural interaction provided by speech, as opposed to the traditional method of selecting dialogue options, and the freedom to formulate commands without preset limits. Previous studies have identified communication constraints in games as limiting, leading to player frustration and reduced agency (Zargham, Fetni, et al., 2024; Zargham et al., 2025). In contrast, the freedom of expression in our game enhanced players' sense of control and engagement within the game world. Interview responses indicate that flexibility in communication could enhance players' sense of agency and control and deepen their immersion in the game. This flexibility allows for a more personalized and dynamic gaming experience, particularly beneficial for individuals who may not be comfortable with ordinary control schemes. Nevertheless, many participants expressed dissatisfaction with the artificiality of conversations, calling for more human-like interactions, such as courtesy, context-awareness, and more natural responses. Some participants perceived NPCs as cold or merely "response machines." Participants called for more natural conversation flow and behavior, with suggestions for integrating speech into gameplay activities or requiring more conversational formalities before diving into questions. These findings highlight the need for more nuanced, lifelike NPC interactions. Enhancing speech systems to incorporate conversational formalities such as politeness, context awareness, and smoother conversation dynamics, such as embedding speech interactions into gameplay activities, could create more natural, engaging experiences.

Overall, regarding RQ1, our study reveals that players experience GenAI-based speech interaction in VR as highly immersive, agency-enhancing, and novel, yet also somewhat fragile. Players' immersion and their sense of realism are easily disrupted by technical errors, shallow conversational dynamics, and limited contextawareness. While speech shifts player engagement from menudriven interactions toward more interpersonal-style communication, this shift also raises expectations that current GenAI NPCs may not yet fully meet.

6.2. Perceptions on GenAI-based NPCs

When designing the NPCs, we maintained consistency in core traits, such as interests and hobbies, to isolate the effects of role and gender. We introduced variations in gender, voice, and office roles to diversify the characters. While many players perceived some differences and found certain NPCs more engaging, others felt the characters were too similar, pointing to "monotone" voices and "AI-like" responses. While we deliberately manipulated the mapping between gender and hierarchical office roles, our findings suggest that gender alone did not significantly influence players' perceptions of NPCs. Instead, players' preferences and judgments appeared to be more strongly driven by the narrative framing of the role (e.g., manager vs. intern), the quality of interaction, and perceived personality traits. However, it is important to emphasize that our manipulation was intentionally minimal. We altered gender presentation and voice while keeping dialogue content and personality consistent to ensure comparability across versions. Our design does not allow strong claims about nuanced constructs such as perceived empathy, authority, or player identification. Future work should systematically examine how social attributes shape player–NPC relationships.

Moreover, it is important to note that participants' perception of, and attitudes toward, individual NPCs could be shaped by a combination of interacting factors beyond the characters' narrative traits or designed personalities. Specifically, visual appearance, voice characteristics (e.g., pitch, tonality, and expressiveness), response timing, and the frequency of speech recognition errors or LLM-generated

inconsistencies may all have influenced character evaluations. Because these elements varied perceptually across characters and play sessions, and were not all experimentally controlled or measured independently, our character-specific ratings should be interpreted as holistic user perceptions rather than isolated judgments of character design or personality alone. Participants also expressed interest in learning more about each NPC's unique traits, such as their hobbies. These findings suggest that, while NPCs should be designed with distinct personalities, their differences should be reinforced through varied vocal expressions, emotional depth, and conversational complexity to make interactions more engaging and authentic. Research also suggests that when NPCs have strong personalities and are well-staged, it can create an illusion of intelligence, and players may overlook technical flaws such as recognition failures or hallucinations (Denisova & Cairns, 2015). Participants frequently noted the helpfulness of the characters' responses as the main reason for favoring them over others. When the gen-AI system produced problematic or insufficient responses—often occurring unpredictably with any character—it led to player frustration, causing those NPCs to be viewed unfavorably. This underscores that, regardless of the form of interaction, many players prioritize game progression and become frustrated by factors that create barriers to their advancement. The Designer was the most frequently chosen favorite character, while the Intern was the most frequently disliked. Interviews revealed that some participants adjusted their speech patterns based on the perceived hierarchy within the virtual office, suggesting that players' perceptions of NPCs' roles influenced their interaction dynamics. The favored Designer had a role similar to the player's, while the disliked Intern held a lower position in the game's story. Similar hierarchy-related speech patterns were observed in previous work (Zargham et al., 2025). These dynamics suggest that roles assigned to NPCs can potentially reinforce real-world biases and stereotypes. Game designers should carefully consider these behavioral tendencies when designing GenAI-based NPCs to ensure interactions promote inclusivity and avoid perpetuating hierarchical or discriminatory patterns. GenAI NPCs could function as socio-cultural artifacts. They may unintentionally reproduce or amplify players' implicit biases, even when not explicitly designed to do so.

For a minority of players, a character's attractiveness played a role in determining their preference for specific NPCs, particularly among male participants. Moreover, although we did not observe any significant differences concerning participants' gender and their NPC attractiveness ratings, we saw that male players showed a slight tendency to favor female characters, while female participants exhibited a more balanced preference. This suggests that gendered biases may influence how players engage with and perceive NPCs, as seen in prior research (Zargham et al., 2025). Our data further revealed a correlation between a character's perceived attractiveness and the enjoyment players reported during interactions, with more attractive characters often linked to more enjoyable experiences. This mirrors well-established findings in social psychology, where attractiveness is linked to perceptions of competence and likability (Lorenzo et al., 2010), suggesting that GenAI-driven NPCs might be subject to similar heuristic judgments. These results highlight the importance of visual and narrative design when creating engaging NPCs.

Ultimately, regarding RQ2, we observed that players perceive GenAI-based NPCs primarily through their functional reliability, role-based expectations, and interaction quality, rather than gender or predefined personalities. When GenAI failures occur, players could also interpret them as social behaviors. This reinforces the need for more consistent, context-aware, and emotionally expressive NPC designs. Character distinctiveness must be supported across dialogue, voice, behavior, and responsiveness to create coherent identities.

All in all, our work highlights the immense potential of combining speech interaction, VR, and LLMs for highly immersive gameplay. While free-form speech increases player agency, subtle guidance mechanisms are required to prevent confusion and help maintaining the narrative flow. Designers should prioritize contextawareness, ensuring NPCs respond dynamically to game world changes, player actions, and narrative progression to maintain realism and immersion. Typical speech interaction concerns such as recognition errors, turn-taking issues, and lack of proactivity can disrupt player experience. Improvements in these aspects can create smoother interactions and enhance engagement.

Moreover, NPCs should exhibit conversational depth and coherence, avoiding irrelevant or out-of-context responses that break immersion.

Incorporating formalities such as courtesy and varied conversational styles can lead to more human-like interactions. Other than conversational capabilities, NPCs should exhibit natural behaviors through better voice modulation, facial expressions, gestures, and movement. This could make interactions feel more authentic. Addressing these factors can lead to more believable, engaging, and immersive GenAI-based NPCs, ultimately enhancing the overall gameplay experience.

7. Limitations and future work

Several limitations should be considered when interpreting our findings. First, some participants perceived the 20-minute gameplay as too short and desired a longer session to engage more deeply in conversations with the NPCs rather than focusing primarily on the game's tasks. While this duration was chosen to limit fatigue and VR motion sickness and to ensure comparability across sessions, it may have constrained the depth of social interaction and reduced the emergence of more complex conversational dynamics. Future studies could adopt longer or multiple-session designs and control for fatigue to explore more in-depth character interactions. Further, we did not measure any objective performance metrics, such as puzzle completion time, success rate as our study was primarily concerned with players' subjective experiences and social perceptions during interactions with GenAI-based NPCs. We focused on perceived performance as a measure of players' sense of competence during interactions, which are known to influence immersion and engagement. Future work should combine subjective and objective measures to better understand how interaction quality relates to actual in-game performance. Moreover, we witnessed that several participants struggled with VR controls and navigation, and some were first-time VR users. This novelty effect likely increased cognitive load and may have shifted attention away from speech interaction toward motor control. Extended VR familiarization and improvements in control design and movement mechanics could have improved the player experience. Some players initially found conversations with NPCs awkward, but became more comfortable as the game progressed. This suggests that speech-based games may require an adjustment period for players to become familiar with this interaction style, especially those with less experience in such forms of interaction. Technically, the NPCs lacked certain features such as context-aware dialogue grounding, proactive conversation initiation, and dynamic character movement. These constraints likely reduced the perceived naturalness of the characters, thereby potentially impacting immersion and character evaluation. Furthermore, we did not systematically disentangle the effects of visual design, voice characteristics, response latency, and LLM reliability on character perception. As a result, character ratings in this study reflect a mixture of aesthetic design, narrative framing, and technical performance. Future work should independently manipulate these factors to isolate their respective contributions. Our participants raised concerns about the context awareness of the NPCs. This was not directly manipulated or measured within the experimental design, but rather emerged from participant comments and observed breakdowns in interaction during gameplay. Future work should explicitly and systematically manipulate contextual variables such as NPC access to real-time game state information, environmental changes, or player action history to empirically test how different degrees of context integration affect player experience. Furthermore, participants occasionally referenced gendered or hierarchical interpretations of NPC roles (e.g., perceiving certain characters as more authoritative or submissive based on voice, dialogue style, or appearance). Such perceptions could reflect or reinforce broader social stereotypes (Zargham et al., 2025), which could bias character liking, trust, or perceived competence. We acknowledge that our design may have unintentionally activated certain such biases. We also note a potential confound related to character names, as they could carry cultural, linguistic, or gendered associations that influence perceived personality. Finally, our participants were recruited from two sites (Canada and the Netherlands). We acknowledge that participant samples at both sites were heterogeneous, including individuals from multiple nationalities and native languages. This diversity makes it difficult to attribute observed differences directly to national or cultural factors. Furthermore, other potential confounding variables, such as language proficiency, prior VR experience, and familiarity with

AI technologies, may have influenced player behavior and interaction styles. Hence, we treated any site-related differences as exploratory observations rather than cultural effects.

8. Conclusion

This work examined player interactions with GenAI-based NPCs in a speech-based VR game. We developed an adventure and puzzle game in which players used speech to communicate with the NPCs to progress through the game. An user study with 48 participants examined how players experience such a game, their perceptions of interactions with GenAI-based NPCs, and how they adjust their communication strategies when engaging with characters exhibiting distinct traits and social roles. Our findings indicate that players generally enjoyed the game, appreciated the freedom of natural language expression, and experienced a strong sense of immersion when interacting with GenAI-based NPCs. At the same time, limitations in speech recognition, restricted behavioral depth of the characters, and challenges with generative AI, such as hallucinations and inconsistent responses, disrupted immersion and occasionally undermined the perceived coherence and reliability of the characters' responses. Beyond these empirical observations, this study makes a conceptual contribution to ongoing discussions in HCI and game user research regarding the role of generative AI in interactive systems. In our study, GenAI-based NPCs were perceived as dynamic and partially unpredictable social actors within game worlds. Players' experiences and character perceptions emerged not only from the visual and narrative design of the NPCs, but also from the interactional behavior of the underlying AI system, including its responsiveness, variability, breakdowns, and occasional failures. This positions conversational AI in games as a fundamentally interactiondriven element rather than a static design component. This work further provides an empirical approach for studying player-NPC relationships in systems where dialogue and behavior are not predetermined but dynamically generated, contributing to emerging evaluation practices for generative AI in interactive environments. Overall, our findings highlight both the immense experiential potential and current limitations of integrating speech interaction, VR, and LLM-based NPCs into game design. While such systems can significantly enhance immersion and player agency, their success depends on carefully designed interaction frameworks that manage uncertainty, conversational reliability, and system limitations. For HCI, this raises important questions about how users attribute agency, trust, and intention to AI-driven characters. For game research, it opens new perspectives on character design, narrative structure, and player-NPC relationships in generative systems. By situating our findings within these broader concerns, this work contributes a step toward understanding how GenAI-based NPCs can be meaningfully and responsibly integrated into future interactive and immersive experiences.

Notes

1. https://unity3d.com/unity.
2. https://www.meta.com/quest/products/quest-2/.
3. https://www.amberscript.com/en/.
4. https://atlasti.com.

Acknowledgment

No additional funding was received for this work.

Author contributions

CRediT: **Nima Zargham**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Leandro Tonini**: Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing; **Dmitry Alexandrovsky**: Conceptualization, Methodology, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing; **Emma Grace Ruthven**: Formal analysis, Investigation, Validation, Writing – original draft, Writing – review & editing; **Maximilian A. Friehs**: Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Validation, Writing –

original draft, Writing – review & editing; **Leon Tristan Dratzidis**: Investigation, Software, Supervision, Validation, Writing – original draft, Writing – review & editing; **Bastian Dänekas**: Investigation, Software, Validation, Writing – original draft, Writing – review & editing; **Ioannis Bikas**: Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; **Lennart E. Nacke**: Supervision, Writing – original draft, Writing – review & editing; **Sven Zebel**: Supervision, Writing – original draft, Writing – review & editing; **Rainer Malaka**: Methodology, Resources, Supervision, Writing – review & editing.

Disclosure statement

No potential conflict of interests was reported by the author(s).

Declaration on generative AI

During the preparation of this work, the authors used ChatGPT version 4o to revise sentences and Grammarly for grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed. The authors assume full responsibility for the content of the publication.

ORCID

Nima Zargham http://orcid.org/0000-0003-4116-0601
Leandro Tonini http://orcid.org/0009-0004-4354-139X
Dmitry Alexandrovsky http://orcid.org/0000-0001-9551-719X
Emma Grace Ruthven http://orcid.org/0009-0003-9267-161X
Maximilian A. Friehs http://orcid.org/0000-0002-9362-4140
Leon Tristan Dratzidis http://orcid.org/0009-0005-3504-0621
Bastian Dänekas http://orcid.org/0000-0002-4240-0761
Ioannis Bikas http://orcid.org/0000-0002-1807-3764
Lennart E. Nacke http://orcid.org/0000-0003-4290-8829
Sven Zebel http://orcid.org/0000-0002-0905-5687

References

Abeele, V. V., Spiel, K., Nacke, L., Johnson, D., & Gerling, K. (2020). Development and validation of the player experience inventory: A scale to measure player experiences at the level of functional and psychosocial consequences. *International Journal of Human-Computer Studies*, *135*(2020), 102370. https://doi.org/10.1016/j.ijhcs.2019.102370

Alexandrovsky, D., Volkmar, G., Spliethöver, M., Finke, S., Herrlich, M., Döring, T., Smeddinck, J. D., & Malaka, R. (2020). *Playful user-generated treatment: a novel game design approach for VR exposure therapy* [Paper presentation]. Proceedings of the Annual Symposium on Computer-Human Interaction in Play (Virtual Event, Canada) (CHI PLAY '20), New York, NY, USA (pp. 32–45). Association for Computing Machinery. https://doi.org/10.1145/3410404.3414222

Allison, F. J. (2020). *Voice interaction game design and gameplay* [PhD. Dissertation]. University of Melbourne.

Allison, F., Carter, M., & Gibbs, M. (2020). Word Play: A history of voice interaction in digital games. *Games and Culture*, *15*(2), 91–113. https://doi.org/10.1177/1555412017746305

Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018). *Design patterns for voice interaction in games* [Paper presentation]. Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play (Melbourne, VIC, Australia) (CHI PLAY '18), New York, NY, USA (pp. 5–17). Association for Computing Machinery. https://doi.org/10.1145/3242671.3242712

Allison, F., Newn, J., Smith, W., Carter, M., & Gibbs, M. (2019). *Frame analysis of voice interaction gameplay* [Paper presentation]. Proceedings of the 2019 CHI Conference on Human [Paper presentation]. Factors in Computing Systems (Glasgow, Scotland UK) (CHI '19), New York, NY, USA (pp. 1–14). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300623

Anzai, S., Ogawa, T., & Hoshino, J. (2021). Speech recognition game interface to increase intimacy with characters. In Jannicke Baalsrud Hauge, Jorge C. S. Cardoso, Licínio Roque, & Pedro A. Gonzalez-Calero (Eds.), *Entertainment Computing – ICEC 2021*. Springer International Publishing. 167–180.

Bonfert, M., Zargham, N., Saade, F., Porzel, R., & Malaka, R. (2021). *An evaluation of visual embodiment for voice assistants on smart displays* [Paper presentation]. CUI 2021-3rd Conference on Conversational User Interfaces, New York, NY, USA (pp. 1–11). ACM.

Bowey, J. T., Friehs, M. A., & Mandryk, R. L. (2019). *Red or blue pill: Fostering identification and transportation through dialogue choices in RPGs* [Paper presentation]. Proceedings of the 14th International Conference on the

Foundations of Digital Games (San Luis Obispo, California, USA) (FDG '19), New York, NY, USA (p. 28). Association for Computing Machinery. https://doi.org/10.1145/3337722.3337734

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2019). *Thematic analysis* (pp. 843–860). Springer Singapore. https://doi.org/10.1007/978981-10-5251-4_103

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712.

Cairns, P., Cox, A., & Nordin, A. I. (2014). *Immersion in digital games: Review of gaming experience research* (Vol. 12, pp. 337–361). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118796443.ch12https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118796443.ch12

Carter, M., Allison, F., Downs, J., & Gibbs, M. (2015). *Player identity dissonance and voice interaction in games* [Paper presentation]. Proceedings of the 2015 [Annual Symposium on Computer-Human Interaction in Play (London, United Kingdom) (CHI PLAY '15), New York, NY, USA (pp. 265–269). Association for Computing Machinery. https://doi.org/10.1145/2793107.2793144

Celeste-AI. (2022). *Celeste-AI*. https://github.com/Celeste-AI/Celeste-AI (accessed 2025 January 20).

Chen Gao, Q., & Emami, A. (2023). The Turing quest: Can transformers make good NPCs? In Vishakh Padmakumar, Gisela Vallejo, & Yao Fu (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)* (pp. 93–103). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-srw.17

Chen, V. H.-H., Duh, H. B.-L., Phuah, P. S. K., & Yan Lam, D. Z. (2006). Enjoyment or engagement? Role of social interaction in playing massively mulitplayer online role-playing games (MMORPGS). In Richard Harper, Matthias Rauterberg, & Marco Combetto (Eds.), *Entertainment Computing - ICEC 2006* (pp. 262–267). Springer Berlin Heidelberg.

Chen, Y. H., Keng, C.-J., & Chen, Y.-L. (2022). How interaction experience enhances customer engagement in smart speaker devices? The moderation of gendered voice and product smartness. *Journal of Research in Interactive Marketing*, *16*(3), 403–419. https://doi.org/10.1108/jrim-03-2021-0064

Christiansen, F. R., Hollensberg, L. N., Bach Jensen, N., Julsgaard, K., Jespersen, K. N., & Nikolov, I. (2024). *Exploring presence in interactions with LLM-driven NPCs: A comparative study of speech recognition and dialogue options* [Paper presentation]. Proceedings of the 30th ACM Symposium on Virtual Reality Software and Technology (Trier, Germany) (VRST '24), New York, NY, USA (pp. 6–11). Association for Computing Machinery. https://doi.org/10.1145/3641825.3687716

Cole, H., & Griffiths, M. D. (2007). Social interactions in massively multiplayer online role-playing gamers. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society*, *10*(4), 575–583. https://doi.org/10.1089/cpb.2007.9988

Connelly, L. M., & Peltzer, J. N. (2016). Underdeveloped themes in qualitative research: Relationship with interviews and analysis. *Clinical Nurse Specialist*, *30*(1), 52–57. https://doi.org/10.1097/NUR.0000000000000173

Convai Inc. (n.d.). *Convai*. https://convai.com/. (accessed 2025 January 20).

Denisova, A., & Cairns, P. (2015). *The placebo effect in digital games: Phantom perception of adaptive artificial intelligence*. In Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play (London, United Kingdom) (CHI PLAY '15) (pp. 23–33). Association for Computing Machinery. https://doi.org/10.1145/2793107.2793109

Dratzidis, L. T., Zargham, N., & Malaka, R. (2025, August). "Follow my lead": Role of AI-Based NPC autonomy in player-NPC collaboration. In *International Conference on Entertainment Computing* (pp. 60–77). Springer Nature Switzerland.

El-Yamri, M., Romero-Hernandez, A., Gonzalez-Riojo, M., & Manero, B. (2019). Designing a VR game for public speaking based on speakers features: A case study. *Smart Learning Environments*, *6*(1), 12. https://doi.org/10.1186/s40561-019-0094-1

Faiqotuzzulfa, F., & Putra, S. A. (2023). Virtual reality's impacts on learning results in 5.0 education: A Meta-Analysis. *International Transactions on Education Technology*, *1*(1), 10–18. https://doi.org/10.33050/itee.v1i1.172

Ganschow, B., Zebel, S., van Gelder, J.-L., & Cornet, L. J. M. (2024). Feeling connected but dissimilar to one's future self reduces the intentionbehavior gap. *PLOS One*, *19*(7), e0305815. https://doi.org/10.1371/journal.pone.0305815

Garcia-Pi, B., Chaudhury, R., Versaw, M., Back, J., Kwon, D., Kicklighter, C., Taele, P., & Seo, J. H. (2023). AllyChat: Developing a VR conversational AI agent using few-shot learning to support individuals with intellectual disabilities. In José Abdelnour Nocera, Marta Kristín Lárusdóttir, Helen Petrie, Antonio Piccinno, & Marco Winckler (Eds.), *Human-Computer Interaction – IN-TERACT 2023* (pp. 402–407). Springer Nature Switzerland.

Guimarães, M., Prada, R., Santos, P. A., Dias, J., Jhala, A., & Mascarenhas, S. (2020). *The impact of virtual reality in the social presence of a virtual agent* [Paper presentation]. Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (Virtual Event, Scotland, UK) (IVA '20), New York, NY, USA (p. 238). Association for Computing Machinery. https://doi.org/10.1145/3383652.3423879

Guimarães, M., Santos, P. A., & Jhala, A. (2022). Emergent social NPC interactions in the Social NPCs Skyrim mod and beyond. arXiv preprint arXiv:2207.13398.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hedeshy, R., Kumar, C., Lauer, M., & Staab, S. (2022). *All birds must fly: The experience of multimodal hands-free gaming with gaze and nonverbal voice synchronization* [Paper presentation]. Proceedings of the 2022 International Conference on Multimodal Interaction (Bengaluru, India) (ICMI '22), New York, NY, USA (pp. 278–287). Association for Computing Machinery. https://doi.org/10.1145/3536221.3556593

Henrik, W. (2016). *The non-player character: Exploring the believability of NPC presentation and behavior* [PhD. Dissertation]. Department of Computer and Systems Sciences, Stockholm University. https://api.semanticscholar.org/ CorpusID:64022074

Hicks, K., Gerling, K., Dickinson, P., Linehan, C., & Gowen, C. (2018). Leveraging icebreaking tasks to facilitate uptake of voice communication in multiplayer games. In Adrian David Cheok, Masahiko Inami, & Teresa Romão (Eds.), *Advances in computer entertainment technology* (pp. 187–201). Springer International Publishing. https://doi.org/10.1007/978-3-319-76270-8_14

Hong, M., Choi, Y., & Cha, S. (2021). *"Anyway": Two-player defense game via voice conversation* [Paper presentation]. Extended Abstracts of the 2021 Annual Symposium on Computer-Human Interaction in Play (Virtual Event, Austria) (CHI PLAY '21)., New York, NY, USA (pp. 345–349). Association for Computing Machinery. https://doi.org/10.1145/3450337.3483509

Inworld AI. (2024). *FAQ – large language models (LLMs).* https://docs.inworld. ai/docs/tutorial-basics/faq/#llms (Accessed 2024 August 21).

InWorld AI Inc. (n.d.). *InWorld AI.* https://inworld.ai/. (Accessed 2025 January 20).

Ishiguro, T., Suzuki, C., Nakakoji, H., Funagira, Y., & Takao, M. (2019). Immersive experience influences eye blink rate during virtual reality gaming. *Polish Psychological Bulletin*, *50*(1), 49–53. https://doi.org/10.24425/ppb.2019.126018

Jennett, C., Cox, A. L., Cairns, P., Dhoparee, S., Epps, A., Tijs, T., & Walton, A. (2008). Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, *66*(9), 641–661. https://doi.org/10.1016/j.ijhcs.2008.04.004

Kinoshita, K., Ochiai, T., Delcroix, M., & Nakatani, T. (2020). *Improving noise robust automatic speech recognition with single-channel time-domain enhancement network* [Paper presentation]. In ICASSP 2020 – 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New York, US (pp. 7009–7013). IEEE. https://doi.org/10.1109/ICASSP40776.2020.9053266

Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023). Scenecraft: Automating interactive narrative scene generation in digital games with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, *19*(1), 86–96. https://doi.org/10.1609/aiide.v19i1.27504

Lee, K. M., Peng, W., Jin, S.-A., & Yan, C. (2006). Can robots manifest personality?: An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication*, *56*(4), 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x

Leroy, R. (2021). *Immersion, flow and usability in video games* [Paper presentation]. Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (Yokohama, Japan) (CHI EA '21), New York, NY, USA (pp. 37–48). Association for Computing Machinery. https://doi.org/10.1145/3411763.3451514

Lorenzo, G. L., Biesanz, J. C., & Human, L. J. (2010). What is beautiful is good and more accurately understood: Physical attractiveness and accuracy in first impressions of personality. *Psychological Science*, *21*(12), 1777–1782. https://doi.org/10.1177/0956797610388048

Maleki, M. F., & Zhao, R. (2024, November). Procedural content generation in games: A survey with insights on emerging LLM integration. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (vol. 20, pp. 167–178). https://doi.org/10.1609/aiide.v20i1.31877

McLean, G., Osei-Frimpong, K., & Barhorst, J. (2021). Alexa, do voice assistants influence consumer brand engagement?–Examining the role of AI powered voice assistants in influencing consumer brand engagement. *Journal of Business Research*, *124*, 312–328. https://doi.org/10.1016/j.jbusres.2020.11.045

Min, Z., & Wang, J. (2024). Exploring the integration of large language models into automatic speech recognition systems: An empirical Study. In Biao Luo, Long Cheng, Zheng-Guang Wu, Hongyi Li, & Chaojie Li (Eds.), *Neural information processing* (pp. 69–84). Springer Nature Singapore. https://doi.org/10.48550/arXiv.2307.06530

Morgan, H. (2022). Understanding thematic analysis and the debates involving its use. *The Qualitative Report*, *27*(10), 2079–2090. https://doi.org/10.46743/2160-3715/2022.5912

Mulvaney, P., Rooney, B., Friehs, M. A., & Leader, J. F. (2024). Social VR design features and experiential outcomes: Narrative review and relationship map for dyadic agent conversations. *Virtual Reality*, *28*(1), 45. https://doi.org/10.1007/s10055-024-00941-0

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press.

Nima, Z. (2024). *Expanding speech interaction for domestic activities* [PhD. Dissertation]. Universität Bremen.
Osking, H., & Doucette, J. A. (2019, June). Enhancing emotional effectiveness of virtual-reality experiences with voice control interfaces. In *International Conference on Immersive Learning* (pp. 199–209). Springer International Publishing.
Pan, M., Kitson, A., Wan, H., Prpa, M. (2025, July). ELLMA-T: An embodied LLM-agent for supporting english language learning in social VR. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference* (pp. 576–594).
Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative agents: Interactive simulacra of human behavior* [Paper presentation]. Proceedings of the 36th Annual ACM Symposium on, User Interface Software and Technology (San Francisco, CA, USA) (UIST '23)., New York, NY, USA (p. 22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763
Rao, S., Xu, W., Xu, M., Leandro, J., Lobb, K., DesGarennes, G., Brockett, C., Dolan, B. (2024). Collaborative quest completion with LLM-driven non-player characters in minecraft. arXiv preprint arXiv:2407.03460. https://arxiv.org/abs/2407.03460
Reicherts, L., Zargham, N., Bonfert, M., Rogers, Y., & Malaka, R. (2021). *May I interrupt? Diverging opinions on proactive smart speakers* [Paper presentation]. Proceedings of the 3rd Conference on Conversational User Interfaces (Bilbao (Online), Spain) (CUI '21), New York, NY, USA (p. 10). Association for Computing Machinery. https://doi.org/10.1145/3469595.3469629
Ruan, J., Chen, Y., Zhang, B., Xu, Z., Bao, T., Mao, H., Li, Z., Zeng, X., & Zhao, R. (2023). *Tptu: Task planning and tool usage of large language model-based AI agents.* https://openreview.net/forum? id=GrkgKtOjaH
Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2022). Voice in human–agent interaction: A survey. *ACM Computing Surveys*, *54*(4), 1–43. https://doi.org/10.1145/3386867
Shoa, A., Oliva, R., Slater, M., & Friedman, D. (2023). *Sushi with Einstein: Enhancing hybrid live events with LLM-based virtual humans* [Paper presentation]. Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents (Würzburg, Germany) (IVA '23), New York, NY, USA (p. 6). Association for Computing Machinery. https://doi.org/10.1145/3570945.3607317
Skantze, G. (2021). Turn-taking in conversational systems and humanrobot interaction: A review. *Computer Speech & Language*, *67*, 101178. https://doi.org/10.1016/j.csl.2020.101178
Suh, J., Bennett, C. C., Weiss, B., Yoon, E., Jeong, J., & Chae, Y. (2021). *Development of speech dialogue systems for social AI in cooperative game environments* [Paper presentation]. 2021 IEEE Region 10 Symposium (TENSYMP), 1, New York, NY, USA (pp. 1–4). IEEE. https://doi.org/10.1109/TENSYMP52854.2021.9550859
The Pandas Development Team. (2020). Pandas-Dev/Pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134
Todd, G., Earle, S., Nasir, M. U., Green, M. C., & Togelius, J. (2023). *Level generation through large language models* [Paper presentation]. Proceedings of the 18th International Conference on the Foundations of Digital Games (Lisbon, Portugal) (FDG '23), New York, NY, USA (p. 8). Association for Computing Machinery.https://doi.org/10.1145/3582437.3587211
Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026. https://doi.org/10.21105/joss.01026
Valve Corporation. (2020). *Half-Life: Alyx*. Windows. https://www.half-life.com/ en/alyx/
Värtinen, S., Hämäläinen, P., & Guckelsberger, C. (2024). Generating role-playing game quests with GPT language models. *IEEE Transactions on Games*, *16*(1), 127–139. https://doi.org/10.1109/TG.2022.3228480
Volum, R., Rao, S., Xu, M., DesGarennes, G., Brockett, C., Van Durme, B., Deng, O., Malhotra, A., & Dolan, B. (2022). Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. In Marc Alexandre Côté, Xingdi Yuan, and Prithviraj Ammanabrolu (Eds.), *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)* (pp. 25–43). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wordplay-1.3
Wan, H., Zhang, J., Suria, A. A., Yao, B., Wang, D., Coady, Y., & Prpa, M. (2024). *Building LLM-based AI agents in social virtual reality* [Paper presentation]. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI EA '24), New York, NY, USA (p. 7). Association for Computing Machinery. https://doi.org/10.1145/3613905.3651026
Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv preprint arXiv:2305.16291. https://arxiv.org/abs/2305.16291
Wilson, C. (2013). *Interview techniques for UX practitioners: A user centered design method*. Newnes.
Winkler, N., Röthke, K., Siegfried, N., & Benlian, A. (2020). Lose yourself in VR: Exploring the effects of virtual reality on individuals' immersion. *Proceedings of the 53rd Hawaii International Conference on System Sciences.* https://doi.org/10.24251/HICSS.2020.186
Wolf In Motion Ltd. (2016). *Speech Trainer*. Windows. https://store.steampowered. com/app/552770/Speech_Trainer/

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Yin, B., & Hu, X. (2023). Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–32. https://doi.org/10.48550/arXiv.2304.13712

Yao, S., & Kim, G. (2019). The effects of immersion in a virtual reality game: Presence and physical activity. In Xiaowen Fang (Ed.), *HCI in games* (pp. 234–242). Springer International Publishing. https://doi.org/10.1007/978-3-030-22602-2_18

Zargham, N., Bonfert, M., Volkmar, G., Porzel, R., & Malaka, R. (2020). *Smells like team spirit: Investigating the player experience with multiple interlocutors in a VR game* [Paper presentation]. Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (Virtual Event, Canada) (CHI PLAY '20), New York, NY, USA (pp. 408–412). Association for Computing Machinery. https://doi.org/10.1145/3383668.3419884

Zargham, N., Dratzidis, L. T., Alexandrovsky, D., Friehs, M. A., & Malaka, R. (2025). Gaming with etiquette: Exploring courtesy as a game mechanic in speech-based games. *International Journal of Human–Computer Interaction*, *41*(11), 6968–6986. https://doi.org/10.1080/10447318.2024.2387901

Zargham, N., Fetni, M. L., Spillner, L., Muender, T., & Malaka, R. (2024). *"I know what you mean": Context-aware recognition to enhance speech-based games* [Paper presentation]. Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24), New York, NY, USA (pp. 18). Association for Computing Machinery. https://doi.org/10.1145/3613904.3642426.

Zargham, N., Friehs, M. A., Tonini, L., Alexandrovsky, D., Ruthven, E. G., Nacke, L. E., & Malaka, R. (2024). Let's talk games: An expert exploration of speech interaction with NPCs. *International Journal of Human–Computer Interaction*, *0*(0), 1–21. https://doi.org/10.1080/10447318.2024.2338666

Zargham, N., Merkord, A., Safari, A., Dratzidis, L. T., & Malaka, R. (2025, August). "Can you feel me now?": Exploring player empathy in AI-based NPC conversations . In *International Conference on Entertainment Computing* (pp. 389–407). Springer Nature Switzerland.

Zargham, N., Pfau, J., Schnackenberg, T., & Malaka, R. (2022). *"I didn't catch that, but I'll try my best": Anticipatory error handling in a voice controlled game* [Paper presentation]. Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22), New York, NY, USA (pp. 13). Association for Computing Machinery. https://doi.org/10.1145/3491102.3502115

Zhang, R. (2023). Research on the progress of VR in game. *Highlights in Science, Engineering and Technology*, *39*, 103–110. https://doi.org/10.54097/hset.v39i.6507

Zhao, R., Wang, K., Divekar, R., Rouhani, R., Su, H., & Ji, Q. (2018). *An immersive system with multi-modal human-computer interaction* [Paper presentation]. 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), New York, NY, USA (pp. 517–524). IEEE.

Zhao, Z., Fan, W., Li, J., Liu, Y., Mei, X., Wang, Y., Zhao, X., Tang, J., & Li, Q. (2024). Recommender systems in the era of large language models (LLMs). *IEEE Transactions on Knowledge and Data Engineering*, *36*(11), 6889–6907. https://doi.org/10.1109/TKDE.2024.3392335

About the authors

Nima Zargham is a postdoctoral researcher in the Digital Media Lab at the University of Bremen. His research focuses on human-AI communication and collaboration. Some of his key research areas include AI persona design, proactive AI, game-based AI interactions, and the ethical and societal implications of AI.

Leandro Tonini holds a Bachelor's and Master's degree in psychology from University of Twente. He currently works at the Center for Forensic Psychiatry in Thuringia (Germany).

Dmitry Alexandrovsky holds a postdoctoral position at the HCI & Accessibility Lab at Karlsruhe Institute of Technology. He works in the area of VR and Accessibility, and develops new research perspectives on disengagement from digital technology.

Emma Grace Ruthven holds a BA (Hons) Psychology from the University of Strathclyde and MSc Forensic Psychology from Glasgow Caledonian University, she has spent time collaborating with researchers within the Psychology of Conflict, Risk and Safety department at the University of Twente.

Maximilian A. Friehs is an Assistant Professor in the Psychology of Conflict Risk and Safety section at University of Twente and a research fellow at University College Dublin. His work focusses on behavior change which includes human-mediated reality-interactions and performance enhancement in critical situations.

Leon Tristan Dratzidis is a research associate in the Digital Media Lab at the University of Bremen. His research focuses on human–agent collaboration and communication, specifically in the context of video games.

Bastian Dänekas is a research associate in the Digital Media Lab at the University of Bremen. His research focuses on the application of computer science in sports.

Ioannis Bikas is a research associate in the Digital Media Lab at the University of Bremen. His research focuses on skill acquisition and player performance in video games and esports.

Lennart E. Nacke is the director of the HCI Games Group and a full professor of user experience and design at the Stratford School of Interaction Design and Business, University of Waterloo. He focuses on player experience in video games, immersive VR environments, gamification, and the impact of generative AI.

Sven Zebel is an Associate Professor in the Psychology of Conflict Risk and Safety section at University of Twente and endowed professor for mediation at the Vrije Universiteit Amsterdam. His work focusses on conflict resolution and explores options for technology-based solutions.

Rainer Malaka is a Professor of Digital Media at the University of Bremen. He is IFIP Fellow and dean of the department for Mathematics and Computer Science of the University of Bremen. His research focuses on multimodal interaction, language understanding, entertainment computing, and artificial intelligence.