Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options

Frederik Roland Christiansen
Aalborg University
Aalborg, Denmark
fredch21@student.aau.dk

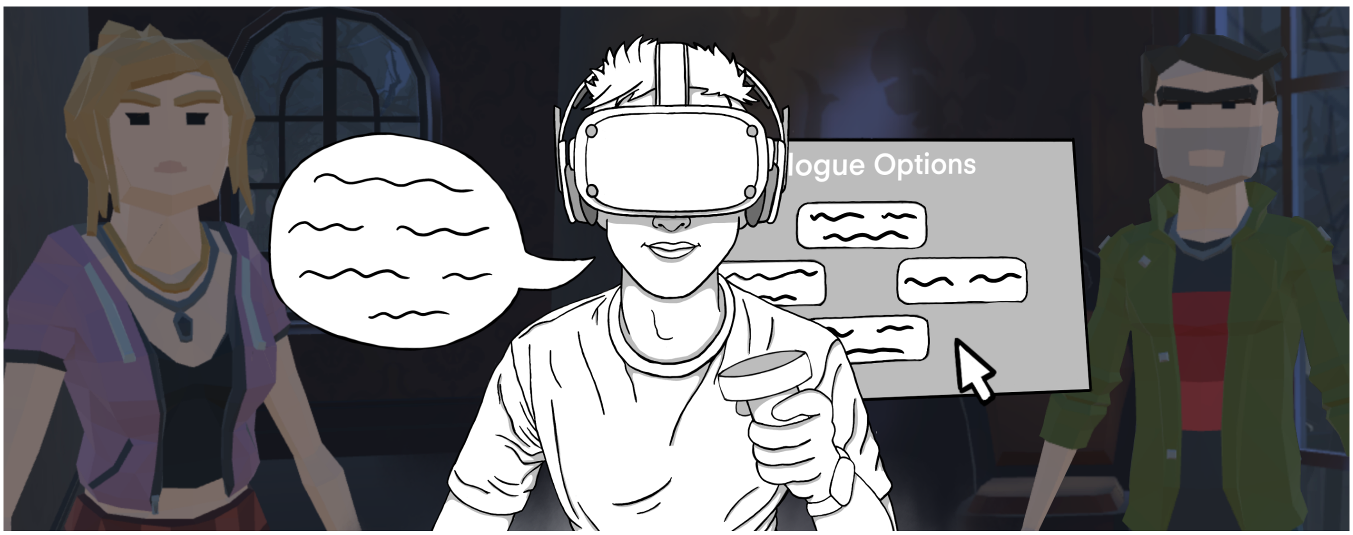
Linus Nørgaard Hollensberg
Aalborg University
Aalborg, Denmark
lnho19@student.aau.dk

Niko Bach Jensen
Aalborg University
Aalborg, Denmark
nbfj21@student.aau.dk

Kristian Julsgaard
Aalborg University
Aalborg, Denmark
kjulsg21@student.aau.dk

Kristian Nyborg Jespersen
Aalborg University
Aalborg, Denmark
knje21@student.aau.dk

Ivan Nikolov
Aalborg University
Aalborg, Denmark
iani@create.aau.dk



ABSTRACT

Combining modern technologies like large-language models (LLMs), speech-to-text, and text-to-speech can enhance immersion in virtual reality (VR) environments. However, challenges exist in effectively implementing LLMs and educating users. This paper explores implementing LLM-powered virtual social actors and facilitating user communication. We developed a murder mystery game where users interact with LLM-based non-playable characters (NPCs) through interrogation, clue-gathering, and exploration. Two versions were tested: one using speech recognition and another with traditional dialog boxes. While both provided similar social presence, users felt more immersed with speech recognition but found it overwhelming, while the dialog version was more challenging. Slow NPC response times were a source of frustration, highlighting the need for faster generation or better masking for a seamless experience.

CCS CONCEPTS

• **Human-centered computing** → **Auditory feedback**; **Virtual reality**.

KEYWORDS

Large Language Models (LLM), VR, Presence, NPC, Speech Recognition, Social Actors, Immersive systems

VRST '24, October 9–11, 2024, Trier, Germany

ACM ISBN 979-8-4007-0535-9/24/10
https://doi.org/10.1145/3641825.3687716

ACM Reference Format:
Frederik Roland Christiansen, Linus Nørgaard Hollensberg, Niko Bach Jensen, Kristian Julsgaard, Kristian Nyborg Jespersen, and Ivan Nikolov. 2024. Exploring Presence in Interactions with LLM-Driven NPCs: A Comparative Study of Speech Recognition and Dialogue Options. In *30th ACM Symposium on Virtual Reality Software and Technology (VRST '24), October 9–11, 2024, Trier, Germany*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641825.3687716

1 INTRODUCTION

1.1 Presence in mediated experiences

Research into mediated experiences has a long history and includes technologies such as home theatre, videoconferencing, IMAX films, which are technologies designed to give the user a sense of unmediated reality despite the mediated nature of the interaction [32]. Virtual reality (VR) research in more recent years has built upon this [51, 54–56]. *Presence* particularly is used as a term to describe the ability of such mediated experiences to deliver a sense of 'being there' [25], or when the participant is 'concerned' by the things occurring in the virtual environment [36]. *Immersion* on the other hand affects the amount of presence experienced by users [10, 18, 45], which in turn strengthens users' feelings of relatedness, self-expansion, and enjoyment [5]. The importance of the sense of presence has been pointed out concerning the usefulness and validity of surgical skill training [44], and military training [24], the effectiveness of fear-based exposure therapy [57], increasing cognitive abilities in VR [46], as well as to increasing the users' experience, as they become more engaged and immersed due to an increase in arousal and emotional response in video games and cinematic virtual experiences [27, 32].

Slater, in his 2009 article [55], lays out three concepts that are used together to explain what makes up presence. These concepts are 1) immersion, 2) place illusion (PI), and 3) plausibility illusion (Psi). PI is the extent to which the user experiences the illusion of "being there" in the virtual environment, despite knowing they are not. Immersion can then be thought of as the capabilities of the system to *deliver* this illusion, whereas PI is the extent to which the user *experiences* the feeling of "being there" within the virtual environment. Psi is the degree to which the user finds the environment plausible or believable, i.e., how realistic or coherent they find it to be. Increases in Psi are achieved when 1) the world reacts to users' actions, such as virtual non-player characters (NPCs) making "eye contact" if looked at or responding if being spoken to, 2) when the environment reacts to simply the virtual existence of the user, such as NPCs engaging with the user without first being addressed, and 3) when events in the environment stay true to the users' expectations of what should occur within the setting [51]. PI and Psi affect users' experience, and the experiences of these illusions are what is understood as the sense of presence.An even better view of what presence is given by the six factors presented by Lombard and Ditton [32] - "Presence as social richness", "presence as realism", "presence as transportation", "presence as immersion", "presence as medium as a social actor", and "presence as a social actor within a medium". We will particularly focus on the last one. This factor can be seen as either real actors (TV personalities) or virtual actors (video game NPCs) within a medium evoke a sense of presence in users, by taking the role of social actors in para-social interactions. This leads to psychological processes in users that cause them to overlook the mediated nature of the exchange, which enhances the felt sense of presence and makes them respond to social cues as in non-mediated communication. For 3D environments involving NPCs, this definition of presence is very applicable. And since realistic NPC behavior is important for the perceived realism and believability of the world [35], we will include this definition of presence in our study. Among immersive technologies, VR headsets offer a high degree of system immersion [40], as it blurs the boundaries between the physical and virtual world [37, 60]. This is found to induce a greater sense of presence compared to other mediated experiences [45]. High levels of system immersion can lead to greater spatial understanding [6], and can evoke even stronger emotional responses in users [19, 45, 65].

1.2 Challenges with existing dialogue system in games

The current challenges in traditional video game dialogue systems stem from their reliance on time-consuming and non-scalable human-written scripts [23], which can easily break the illusion of reality, particularly when NPCs fail to recognize players with whom they have previously interacted [23, 62]. A literature review by Silva et al. [53] emphasizes the need for solutions capable of controlling NPC behavior dynamically, and they propose that artificial intelligence (AI) can provide a framework for achieving this goal. Large language models (LLMs) emerge as a promising solution, offering the ability to generate dynamic and context-aware dialogue [28]. Managing contextual awareness is also crucial, and one approach involves using "Game Tokens" to inform the AI about the player's knowledge [23].

Despite challenges, AI-driven dialogue systems offer several advantages, including improved user experience, enhanced replayability, and cost savings compared to human-authored content [2]. But while increasing user freedom of expression, AI must be capable of handling various user inputs to avoid breaking immersion [23].

In this paper, we explore the problem of ways to communicate with LLM-based NPCs within a VR video game setting. To do this we first conduct a large literature review of existing research that uses generative AI and LLMs in VR and video games. Most studies identified, used AI as a tool for idea generation or content creation before any user interaction. Only one study, in our sample, focused on real-time user interaction to generate NPC quest dialogue using GPT2. However, this was not evaluated with users directly affecting the AI and was not conducted within VR. Therefore, to address this gap and assess how players currently interact with NPCs in VR, a state-of-the-art (SOTA) analysis was performed, compiled of 20 of the most popular industry VR game titles. This analysis indicated that current user-NPC interactions, where the user controls the flow of the conversation, predominantly rely on a dialogue options menu through non-diegetic user interfaces (UIs). Only one game allowed for the use of voice commands but was restricted to 55 predefined voice lines. To address these gaps in knowledge we set out to build a VR experience where users can freely interact with ChatGPT-4 LLM-based NPCs, to solve a murder mystery. To test how users' presence and experience are affected by the ways they communicate with the NPCs we built two versions of the experience utilizing two distinct communication modalities - unrestricted verbal communication through a speech recognition layer and more traditional predefined dialogue options. We test the experience of the users through a Game Experience Questionnaire (GEQ) [20], for both social and in-game interactions. We show that even though there was no statistically significant difference between the results of the two methods of interaction, users displayed stronger behavioral involvement with the speech recognition version while finding the dialogue options more challenging. We also show that more users managed to correctly solve the murder mystery, using the more traditional dialogue option, as they felt overwhelmed by all the possibilities of being able to ask the NPCs anything. Finally, we also show that there is more work needed in lowering the generation times of NPCs to make the dialogue flow better, as well as preventing the LLM from hallucinating when not given enough information about the context and environment.

The main contributions of this paper are as follows:

- Conducting a literature review of current research into the development of AI-driven NPCs and different user interactions with NPCs in VR;
- Developing a VR murder mystery game, where all NPCs are LLM-based;
- Creating two ways of interacting with the NPCs using different modalities - through speech recognition and through traditional dialogue options, to test how each would change the social presence of users;
- Analysis of current shortcomings of LLM-based NPCs and possible ways to mitigate them.

2 BACKGROUND RESEARCH

2.1 Research on AI used in narrative content

To evaluate the existing literature on the use of generative AI and LLMs in video games, we conducted a structured literature review (SLR) on the ACM Digital Library. We choose to focus on this library, because of the large number of journals and conferences connected to VR and games present.

2.1.1 The screening process. The SLR followed a five-step selection and screen process. Step 1 was the initial search of the articles in the ACM Digital Library. Our search terms are as shown here - *("generative artificial intelligence" OR "generative AI" OR "language model" OR "text generation model" OR "textbased AI" OR "textbased artificial intelligence" OR "large language model")* for the AI part and *("virtual reality" OR "VR" OR "simulated reality" OR "SR" OR "video game" OR "video games")* for the VR part. We required that articles should mention one of a few variations of the terms generative AI or LLMs in combination with either VR, video games, or simulated reality. The inclusion of video games in the search aimed to capture potential similarities in the implementation of generative AI between the two fields of VR and video games. The search only included articles between 2013 and 2023, which yielded **472** results. Step 2 involved removing 20 articles due to inaccessibility, which left **452** results. In Step 3, articles lacking source code transparency were excluded, aligning with the FAIR (Findable, Accessible, Interoperable, and Reusable) principle in research [4]. Given our focus on exploring the current implementation of generative AI in virtual experiences, the articles needed to show their implementation. Articles without a link to or direct inclusion of source code were excluded. 328 articles were excluded, which left **124** articles. Step 4 involved reading throughout the title, abstract, introduction, conclusion, and images for the topics of virtual reality, video games, and narrative settings, as articles that did not research at least one of these topics were irrelevant to the search. Of the 124 remaining articles only **8** articles, featured one of the required topics. Step 5 was performed by fully analyzing the remaining articles to assess how they used AI in VR or video games.

2.1.2 Analysis of SLR results. In three of the eight articles, AI was used as a collaboration tool, where AI and humans collaborated to create narrative dialogue for a screenplay [38, 68, 69]. Another article [30] also utilised AI for storytelling, but here AI was used to predict story-endings based on the script. A third way of collaborating was shown by Mascarenhas et al. [34]. Here, a framework was presented for the AI to build cognitive agents, while also helping with developing video games. Another way to help the developers make video games was by training LLMs on a large dataset of quest descriptions from MMORPGs, like Van Stegeren et al. [64]. The trained AI is then used to generate quest dialogue from a set of given prompts. The AI-generated quest was, however, inferior to the manually written ones when rated by users. Common among these six articles is that they all utilise the AI before any user interaction, as the AI is either used as an idea-generation tool or to create content. The end user is therefore not involved directly with the AI, but instead with the content which the AI helped produce.

Another result [29] uses AI only for speech recognition, which leaves only one paper [3] that uses AI which the user interacts with in real-time. Ashby et al. [3] trained the AI in the same way as Van Stegeren et al. [64] did, but instead used a knowledge graph (a database of relations for the NPC) to procedurally generate quests while the user interacts with the NPC. This lets the user take part in the quest-creating process, by conversing with the NPC. Ashby et al. found that users prefer the generated quest when the users themselves take part in the quest generation.

While these findings suggest that AI might not be desirable when creating game content [64], interacting with an AI-driven NPC can be preferable compared to a "normal" NPC [3]. However, none of the 8 studies were performed in VR. To assess how NPCs are currently implemented in VR games, we, therefore, performed a SOTA analysis on 20 VR games to subsequently shed light on how AI can be used to improve VR-NPC implementation. After looking at the left articles, it was felt that the SLR process was too strict, especially for more cutting-edge technology like LLM-driven NPC, so we expanded this part of the research. In more recent work [66], the number of required hyper-parameters for the most plausible LLM conversations consisted of at least 3 base and 5 context observations, for users to feel a meaningful conversation was done. It has also been shown that continuous learning and environmental feedback, as well as sticking to strict guidelines for developing NPC behavior can greatly enhance the conversational and planning capabilities of LLM-NPC [8, 17, 67]. In addition, introducing personality traits and environmental context can greatly benefit conversational variance [26, 49, 72].

2.2 NPC interactions in industry games

To understand how players interact with NPC's in VR video games, we performed a SOTA analysis of 20 of the most popular VR games (found from [70]), by using the search terms "VR", "Single-player" and "First Person" on the Steam Store. We analyzed how the games had implemented the player's ability to interact with the NPCs.

Game Title	Interaction Types				
	V	DO	PA	BL	OM
Alien Isolation (VR mod) [9]					✓
Phasmophobia [16]	✓				✓
Half-Life: Alyx [63]			✓	✓	✓
Asgard's Wrath [52]		✓	✓		✓
Boneworks [71]				✓	✓
TWD: Saints & Sinners [22]		✓	✓		✓
Skyrim VR [59]		✓	✓		✓
Resident Evil 7 [31]					✓
Moss [48]					✓
Stormland [15]			✓	✓	✓
Hitman 3 VR [1]			✓	✓	✓
Elite: Dangerous [11]					
No Mans Sky [14]		✓			
Fallout 4 VR [58]		✓	✓		✓
Karnage Chronicles [61]		✓			✓
The Room VR [13]					✓
The Persistence [33]			✓	✓	✓
Blood and Truth [12]			✓		
Vader Immortal: Ep1 [21]					✓
Red Matter [50]			✓		✓

Table 1: The types of interactions a player can perform to which an NPC reacts in the 20 VR games used in the SOTA analysis. The types of interaction modalities are: V (voice), DO (dialogue options), PA (proximity awareness), BL (body language), OM (object manipulation).

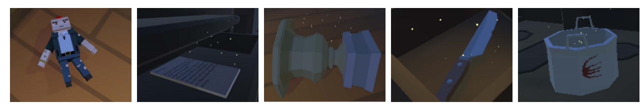
After looking through the trailer and videos of gameplay, games with no clear NPC interactions were excluded from the analysis. The remaining 20 games and their interaction types can be seen in Table 1. In 17 out of 20 cases, the games had the player interact with NPCs through interactions with objects in the scene, triggering a response from the NPCs. Another 9 out of the 20 games used acted voice where the player character is talking to the NPCs. This was often used in combination with proximity-aware NPCs, which would start speaking as the player entered their proximity to convey some story. Only one of the games allowed the player to talk to the NPCs in the form of predefined voice commands.

In cases where the player directly interacted through dialogue with the NPC, the most prominent interaction method was dialogue options, allowing the player to choose what to say to the NPC in some non-diegetic UI menu. This was the case in 6 out of 20 games. Other methods of interaction were discovered, such as body language, where the NPC notices how hands are placed and act upon it, or proximity-based interaction. However, none of these methods made the player able to directly manipulate what they wanted their character to say to the NPC.

Based on this research, the two ways of engaging in dialogue with NPCs in VR games are either using voice commands, or dialogue options. In the (only) case of voice commands, Phasmophobia [16], they were very limited, as the NPC was only able to recognize 55 pre-defined phrases [7]. However, the advances in LLMs and speech-to-text have made it possible for fully-fledged voiced-based conversations with the NPCs. We therefore wanted to test dialogue options against speech recognition, to compare if the use of speech recognition would differ, in terms of the amount of presence experienced by the user, from the current standard of dialogue options in industry VR games.

3 METHODOLOGY

To best test how users react to LLM NPCs and how different ways of interaction change their presence in VR, we wanted to force players to interact with the NPCs as much as possible. This is why we chose to develop a murder mystery game in Unity where natural cooperation and discussion between characters are needed to solve the mystery. Game project files and code together with captured user data are available https://github.com/Julsgaard/Murder_Mystery_VR.



(a) (b) (c) (d) (e)

Figure 1: The five clues present in the game, that can be used by users to find the murderer - from left to right - doll, letter, vase, knife, and pot. The LLM-based NPCs know about the clue and can give information about them - either truth or lie

3.1 Story of the game

As previously mentioned in section 2.1.2, manually written content was found to be preferred to AI-generated content [64]. Therefore, we decided to write the story of our game ourselves. The story for the game was the following: A group of six friends —Jens, Chris, Leonard, Riley, Quinn, and Ashley— decide to explore a haunted mansion. Tensions escalate when Jens, harboring a deep love for Quinn, witnesses his closest confidant, Chris, betraying his girlfriend Ashley through an affair with Quinn. Fuelled by a sense of betrayal and rage, Jens takes a drastic turn, murdering Chris under the cover of darkness and attempting to pin the crime on Leonard.

In this narrative, players assume the role of Riley, the newest addition to the friend group, tasked with unraveling the mystery of the killer's identity before the authorities intervene. Navigating through the mansion, players can explore and interact with stationary (NPCs) to uncover vital clues. The game design incorporates five clues and four NPCs, with the fifth NPC (Chris) found deceased. The clues include a doll (Figure 1a) stolen by Jens from Leonard and strategically placed in Chris' room to incriminate Leonard, an anonymous love letter (Figure 1b) penned by Jens and planted in Quinn's room, a toppled vase (Figure 1c) from the previous night's argument between Ashley and Chris, the murder weapon (Figure 1d) discreetly stowed away in Jens' room, and a blood-stained pot (Figure 1e) in the kitchen—accidentally tarnished by Jens while cooking.

3.2 Implementation of the NPCs

We implement two ways for the player to interact with the NPCs. In one version, the player interacts with NPCs using traditional

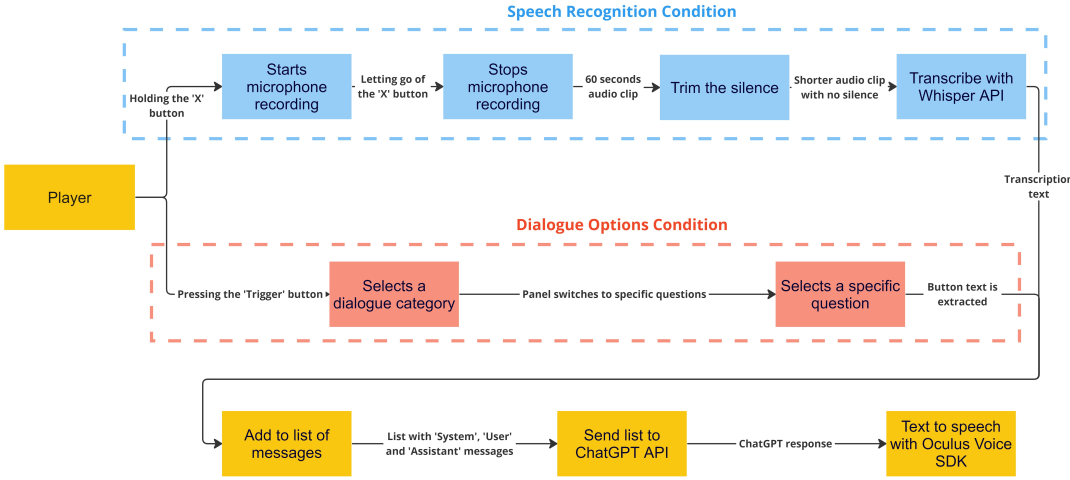
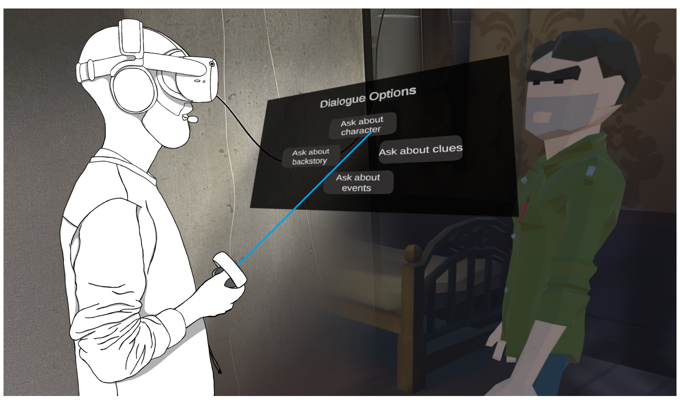
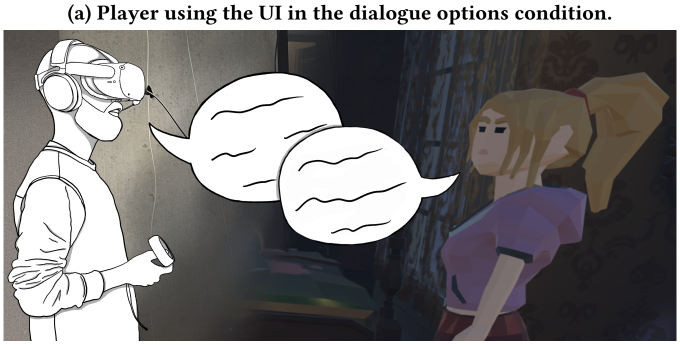


Figure 2: Flowchart of the NPC interaction implementation for both the speech recognition and dialogue conditions



(a) Player using the UI in the dialogue options condition.



(b) Player talking to an NPC in the speech recognition condition.

Figure 3: Hybrid sketches of the player-NPC interactions in both of the implemented test conditions. Figure a illustrates the dialogue options interaction, and figure b illustrates the speech recognition interaction.

non-diegetic UI (dialogue options), and in the other version, the player interacts with the NPCs using speech recognition. Illustrations of both conditions can be seen in Figure 3. Additionally, we implemented head tracking to ensure that NPCs gaze at the player upon entering a specific proximity, as this is known to increase presence [55]. The flowchart of the developed NPC interaction for both conditions can be seen in Figure 2.

In the dialogue option condition, the dialogue UI also appears when the player walks into said proximity. A ray interactor is simultaneously enabled, from the right in-game hand, to allow interaction with the UI. An illustration of this can be seen in Figure 3a. The player navigates the UI to choose what to ask the NPC.

In the speech condition, the player presses and holds a button on the controller and speaks to the NPC, when in proximity. An illustration of this interaction can be seen in Figure 3b. The user-uttered speech was then trimmed to remove silence and sent to be transcribed using the whisper API [43].

When interacting with NPCs in either condition, the new message, the message history, and the character backstory prompt were sent to the ChatGPT API using the model: gpt-4-1106-preview [42]. Once the response was obtained, the Oculus Voice SDK [41] was used to transform the text into speech, allowing the NPCs to talk back to the player with their answers. The messages send to ChatGPT can have three different purposes. *System* messages specifying the rules to which the LLM should adhere - what they can and cannot say, the character backstory, motivation, and information about the world. *User* messages package communication from the user to the LLM with questions or information - either by talking or by picking dialogue options. *Asistant* messages are messages that the LLM had previously generated, and which form the "memory" of each character.

3.2.1 Response time. Despite advancements in LLMs and speech recognition, delays are an inherent aspect of LLM-driven NPCs. This is evident in our analysis of response times with GPT-4 [42]. Here, we utilized 20 different questions for both conditions, measuring their average response times within the Unity environment. In the speech recognition condition, the average response time was higher, averaging 4.00 seconds, primarily due to the transcription step required before processing the input. In contrast, the dialogue

Figure 4: Figure illustrating the setting of the game, and some of the rooms within the mansion in which the game takes place. The Left shows the room of Quinn, the center shows the main hall connecting the upper and lower floor, and the right shows the kitchen area where Jens is standing.

options condition demonstrated a slightly better average response time of 2.66 seconds.

3.3 Gameplay

At the beginning of the game, the player was placed in a virtual dark room, where the backstory was read to them, and controls were explained. When ready, the player could start the actual game, by touching a *Continue* button and getting teleported to the scene of the crime. The participants then played the game for 15 minutes, walking around the rooms in the mansion, some of which can be seen in Figure 4, finding clues, and talking to the NPCs. When the time ran out, they were teleported to another dark room, with "Thank you for playing!" displayed, and then told the game was finished.

3.4 Experimental setup

A between-subjects test with 24 participants, 12 in each condition (dialogue options and speech recognition), was conducted using the following procedure: The participant was introduced to the test and asked to fill out a consent form. They were then told about the game; what would happen, the goal of the game, how they were supposed to operate the game, how to interact with the NPCs, how to interact with clues, how long the test would run for, and that they can track how much time is left through a virtual watch on their hand. They were also told they could stop at any time if they experienced any motion sickness or in any other way felt uncomfortable. Each test took approximately 30 minutes.

The tests were run on a Lenovo Legion 5 (NVIDIA RTX 3060 6 GB GPU, AMD Ryzen 7 5800H CPU & 16 GB RAM) laptop. The users were given an Oculus Quest 2 as an HMD, together with controllers and a Steel Series Arctics Pro headset to be used as a microphone and to better accommodate hearing the NPCs.

After the test, the participants were handed a questionnaire inquiring about the following:

- Questions on whom they thought the murder was, as well as the murderer's motive. This was asked to determine if the interaction modality had an impact on users' ability to follow the plot.
- A GEQ-Social Presence questionnaire [20], which we use to estimate the degree of presence the participants felt, as the module contains components that match our existing theory on presence.
- A GEQ-In-game questionnaire [20], which we use to determine the overall game experience, as it might show relevant differences between the two conditions.
- Free-form questions on what they did and did not like about interacting with the NPCs, and whether they had any additional thoughts on the experience.

4 RESULTS

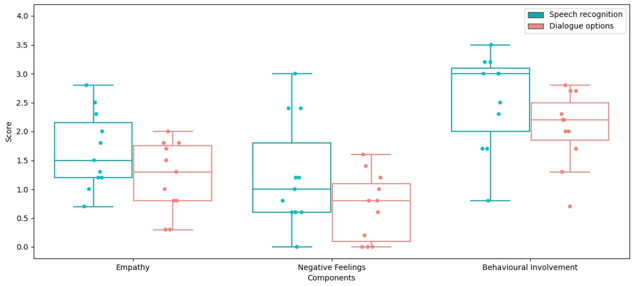
Following the completion of all tests, the data for two participants were excluded from our dataset; one was the first test, which was subsequently treated as an additional pilot test due to some adjustments needed in the test procedures, while the other had to leave due to feeling motion sick and was therefore unable to finish the test. This left us with data from 22 participants, 11 for each condition.

4.1 GEQ Social Presence results

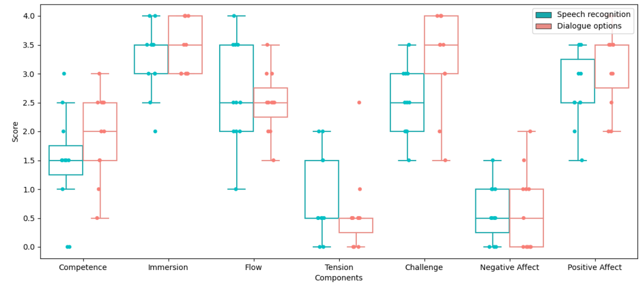
Participants in the speech recognition condition tended to score higher in all three components of the Social Presence GEQ, as shown in Figure 5a. To simplify calculations and visualization, we merged the questions into the three main social presence components - Empathy, Negative Feelings, and Behavioural Involvement. We show that this is possible by calculating Cronbach's alpha, which at 0.829 gives us good internal consistency. The component with the largest difference in scores was behavioral involvement, which translates to how well a player believes they engage with their environment in [47].

A Shapiro-Wilk test showed the data for empathy ($p = 0.92$) and behavioral involvement ($p = 0.42$) to be normally distributed, while data for negative feelings ($p = 0.03$) showed a non-normal distribution. Hence, we employed both a Two-sample T-test and a Mann-Whitney U test to determine any statistical correlation between the conditions and the component scores. No significant differences were found in either the empathy ($t = -1.66, p = 0.11$), behavioral involvement ($t = -1.52, p = 0.14$), or negative feelings ($W = 42.5, p = 0.24$) data. We also determine the effect size using Cohen's d, to see if the experiments need to be changed for future work. We get a value of 0.021, which is far from even the small relative effect size. This indicates that a more discriminate social presence questionnaire is required.

If we look into the separate questions we see much larger differences between the two methods. We run either the Mann-Whitney U or Two-sample T-test and see that four of them have a significant p-value - "**Q5**. The other(s) paid close attention to me" ($W = 5, p = 0.00064$), "**Q6**. I paid close attention to the other(s)" ($W = 22, p = 0.0124$), "**Q9**. When I was happy, the other(s) was/were happy" ($t = 2.32, p = 0.0425$) and "**Q11**. I influenced the mood of the other(s)" ($W = 28.5, p = 0.0385$). The box plots from these questions are shown in Figure 6a.

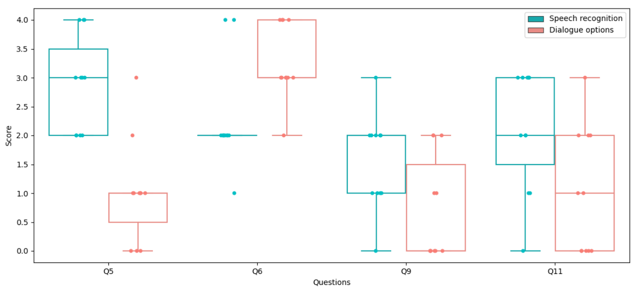


(a) GEQ social presence module.

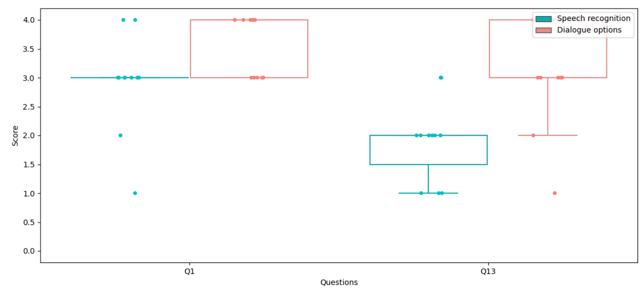


(b) GEQ in-game module.

Figure 5: Box plots displaying the scores for three components within the GEQ social and the seven components within the GEQ in-game module between the two conditions.



(a) Social presence questions with a statistical difference.



(b) In-game questions with a statistical difference.

Figure 6: Box plots displaying the scores for the four questions within the GEQ social and the two questions within the GEQ in-game module with statistical differences.

4.2 GEQ In-Game results

The GEQ-in-game results tend to be very similar between the two conditions, with identical medians being found for the immersion, flow, tension, and negative affect components (See Figure 5b). Exceptions to this are seen in the competence, challenge, and positive affect, all scoring higher in the dialogue options condition. We calculate Cronbach's alpha for the In-Game questions and see that it has a value of 0.494, which is below the threshold for poor internal consistency. Thus we will give more priority to the separate questions analysis than the combined components. After running a Shapiro-Wilks test to determine normality, we either run a T-test or a Mann-Whitney U test for the individual questions. Two questions have significant p-values - "**Q1**. I was interested in the game's story" ($t = 2.28, p = 0.0455$) and "**Q13**. I had to put a lot of effort into it" ($W = 19.5, p = 0.0078$), with their box plots shown in Figure 6b.

As we want to still see if the combined components have any significant difference, we employ Shapiro-Wilk tests to determine if the data followed a normal distribution, which was the case for the competence ($p = 0.12$), flow ($p = 0.19$), challenge ($p = 0.08$), and positive affect ($p = 0.1$) components, but not for the immersion ($p = 0.01$), tension ($p > 0.05$), and negative affect ($p > 0.05$) components. Thus, like for the social presence analysis, we used both Two-sample T-tests and Mann-Whitney U tests to determine any statistical correlation between the conditions and the component scores. A statistical difference was found for the challenge component ($t = 2.42, p = 0.02$), but not for the remaining components; competence ($t = 1.40, p = 0.17$), immersion ($W = 71.5, p = 0.47$), flow ($t = -0.43, p = 0.67$), tension ($W = 48, p = 0.39$), negative affect ($W = 59, p = 0.97$), and positive affect ($t = 1.41, p = 0.17$). When calculating Cohen's d for the GEQ In-Game results we also get a very small value of 0.032, which indicates that the questionnaire did not capture enough between-subject variance compared to the one between participants in the same group.

4.3 Participant feedback on NPC interaction and the overall experience

The delay occurring between either speaking or selecting a dialogue option and the NPC response was commonly addressed in the participant feedback. Specifically, six participants in the dialogue option condition and three individuals in the speech recognition condition mentioned it, with one expressing, "[...] sometimes I had to wait a long time for characters to respond, so I wasn't sure whether to repeat myself or wait". Both groups also mentioned the method of interaction with the NPC within their feedback, with four participants in each group praising it when asked what they liked about interacting with the NPCs. For the dialogue options condition,

Figure 7: Bar chart comparing the number of participants guessing the murderer correctly and incorrectly between the two conditions.

both the fact that you did not have to think of a question yourself: "[I liked] that I was given a set of options of what to say instead of having to find questions myself", and the variety of options were praised: "[I liked that] I had multiple options to interact with them". For the speech recognition condition, the degree of player freedom when choosing what to say was praised: "[I liked] "That it was possible to say whatever I had in mind, and that it was not set in stone".

In contrast to the praise, some participants also expressed that the speech recognition gave them too many options: "[I felt] a bit confused at the beginning of what was possible to say.", and that the dialogue options were too limiting: "It felt like my question options were lacking [. . .]". In terms of the contents of the NPC responses, 11 participants praised it and thought it to be either realistic: "They felt like real people. the dialogue was good" or useful in solving the murder mystery: "[. . .] I kept asking questions, and they kept giving me answers that I could use". On the other hand, one participant did not find the dialogue to be useful: "It felt like I wasn't [able] to conclude anything from talking to them".

4.4 Correctly guessed murderer, and overall understanding of the plot

As previously mentioned, we asked each participant to select whoever they thought the murderer was to see if there was any difference in the users' ability to understand the plot between the two conditions. As seen in Figure 7, 13 out of the 22 participants guessed correctly. Out of these 13, 8 participants were using the dialogue option condition, while the remaining 5 were using the speech recognition condition. To see if there is any significant association between the correct guesses and the used communication modality we employ a Chi-squared independence test. From the test, we get ($Chi = 0.752, p = 0.385$), which indicates that there is no statistically significant evidence that either communication modality has impacted the participant's guessing ability.

5 DISCUSSION

While the GEQ Social Presence components tended to score higher in the speech recognition condition, no statistical differences were found between the two conditions. Despite the implication that the degree of presence remains similar in both conditions, it is still worth examining the subtle differences in participant experiences, such as the questions highlighted in section 4.1. Notably, participants in the speech recognition condition scored significantly higher on average when asked; "The other(s) paid close attention to me." This suggests an increased sense of Psi during interactions in this condition, which likely can be attributed to NPCs responding to personally formulated questions rather than selecting from a set of predetermined options in the dialogue option interface.

Participants in the dialogue options condition scored slightly higher on "I paid close attention to the other(s)," suggesting they perceived NPCs more as social actors, possibly due to the need to be close to the NPCs for dialogue to appear, which kept them focused. This may also explain why they felt more attentive to NPCs. Additionally, the shorter response time in dialogue options might have allowed more frequent questions, aligning with Jakob Nielsen's [39] guideline that a 1-second delay maintains user engagement. Although participants were informed about delays, the higher response time in speech recognition ($M = 4.00, SD = 0.30$), compared to dialogue options ($M = 2.66, SD = 0.45$) likely caused them to lose interest, reducing the sense of interacting with a sentient being.

While the only component within the GEQ In-game module producing a significant difference between the two conditions was "Challenge", there are still some small variations in others, namely "Competence" and "Positive Affect". Both of these components score lower in the speech recognition condition, which suggests that participants using this interaction method felt slightly less competent, and positively affected by the experience. Among the questions within these two components, the question "I felt skillful" had the largest difference between the two conditions, with dialogue option participants scoring higher.

A possible explanation for the heightened sense of "skill" when employing dialogue options may be attributed to the additional task of navigating a layered UI. Our dialogue options UI involves multiple 'layers,' allowing users to first select a category and subsequently choose a question within the selected category. It is possible that navigating these layers and locating the desired question demanded more skill than verbally formulating responses. The potentially more difficult nature of interacting through the dialogue options can also be seen by looking at the questions within the challenge component. The question, 'I had to put a lot of effort into it', was scored over one higher on average by participants using dialogue options. This observation once again suggests that navigating the dialogue options interface to locate the desired response may demand more effort compared to verbally articulating it.

6 LESSONS LEARNED AND FUTURE WORK

We want to highlight some lessons learned from our experiments that could be useful for other researchers working in bringing better LLM-based NPC communication. We also give some future work directions.

6.1 Referring to In-game Items

Five participants in the speech recognition condition referred to clues held in their hands by saying "this" when talking to the NPCs. This way of referring to clues was not implemented in a way where NPCs would recognize the item, and would in all cases lead to the NPCs guessing for which clue was referred to by the player. In three cases, the guess was wrong, which might have negatively impacted the presence felt towards the NPCs. It is worth pointing out that when participants referred indirectly to a held-in-hand clue, it suggests that they perceive the NPCs as being aware of their surroundings, and therefore likely perceived the NPCs as social actors within the medium. This method of interaction is more challenging to implement naturally using dialogue options, and further research should look into how a successful implementation of this might have created a stronger sense of presence in the speech recognition condition.

6.2 Minimizing Variance in Communication Delay

In our study, nine of the 22 participants commented on the delay between their interactions with NPCs and the generation of responses within our game. Interestingly, three participants from the speech recognition condition and six from the dialogue option condition specifically mentioned this delay as a negative feature. Despite this being expected, it brought about an intriguing observation when comparing the two conditions; a higher proportion of participants from the dialogue option group mentioned the delay, despite this modality showing a shorter average delay (2.66 seconds) than speech recognition (average delay of 4.0 seconds).

This discrepancy might be attributed to the expectations associated with interacting through dialogue options versus speech recognition. Participants accustomed to previous gaming experiences may expect instantaneous responses when utilizing dialogue options, whereas the novel nature of speech recognition interactions with larger freedom within a gaming environment might be accompanied by fewer expectations. The potential impact of this discrepancy on the study results remains unclear. Nevertheless, the results would likely have been different had the delay been eliminated, or at least made equal, for both conditions. As it stands now the delay can be seen as a confounding variable, which could have polluted the results, but the relatively small difference in the delay time, should have minimized the possible effects.

6.3 Pursuing a Hybrid Approach

As the overall game experience seemed very close for both conditions (See section 4.2), we can potentially look at both the positive and negative feedback specific to each interaction type to understand what shaped this experience. As mentioned in section 4.3, speech recognition was both praised for its degree of freedom in choosing what to say, but also criticized for not guiding the user in what to say. On the other hand, dialogue options were praised for providing a variety of options to the user but also criticized for being too limiting in what to say. This indicates that a third, hybrid interaction type, incorporating positive aspects from both conditions while also addressing the negative feedback, could potentially produce improved user experience scores. An implementation of this could see the players use speech recognition to communicate with NPCs while having an in-game object (such as a sheet of paper) containing suggestions on what to say. Ideally, this would allow the player the total freedom that the speech-recognition participants praised, while also giving them some sort of guidance on what to say.

7 CONCLUSION

In this paper, we established the important role of presence in enhancing VR experiences, especially with the evolving VR technology. As system constraints on presence decrease, we explored other factors influencing presence, focusing on integrating generative AI and Large Language Models (LLMs) for realistic NPC dialogues. Our study compared user interactions with NPCs through standard dialogue options and microphone-based voice input to assess differences in perceived presence.

We built a murder mystery game to push users to interact more with the LLM-based NPCs and tested their social presence and game experience through GEQ questionnaire. Even though we did not see a statistically significant difference in many of the tested components we showed that speech interaction with the NPCs gave a stronger behavioral involvement in users and was seen as easier and more immersive. The added caveat was that users also felt lost and overwhelmed by the freedom to ask whatever they wanted. On the other hand, the traditional dialog option was seen as more complicated but led people to interact with the NPCs more closely and to guess the correct murderer, showing a better understanding of the VR environment. This was also shown in the qualitative data, where multiple users did not like the restrictive nature of the dialogue options, while others were left confused by what to ask the NPCs and how to start the conversation in the speech recognition option.

We therefore propose further research into a hybrid approach between the two, utilizing the nonrestrictive capabilities of the speech recognition modality, while still providing some guidance for what to ask the NPCs through a more diagetic menus.

Finally, we also saw that the problems of slow response generation times are still present in the current version of the used ChatGPT-4, as well as the tendency to hallucinate answers when not enough information is given to it. Additional testing of different LLM implementations is needed, with a focus on on-computer offline versions. In the next iteration of the testing, we would also like to focus on non-verbal communication with the NPCs through testing user posture, eye gaze, and facial expression to determine how different interaction modalities change users' involvement.

REFERENCES

[1] IO Interactive A/S. 2023. HITMAN World of Assassination. [Playstation Store]. https://store.playstation.com/da-dk/product/EP3969-PPSA01769_00-0000000000000WOA

[2] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-Based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441

[3] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-Based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg,

Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441
[4] Michelle Barker, Neil P Chue Hong, Daniel S Katz, Anna-Lena Lamprecht, Carlos Martinez-Ortiz, Fotis Psomopoulos, Jennifer Harrow, Leyla Jael Castro, Morane Gruenpeter, Paula Andrea Martinez, et al. 2022. Introducing the FAIR Principles for research software. *Scientific Data* 9, 1 (2022), 622.
[5] Miguel Barreda-Ángeles and Tilo Hartmann. 2022. Psychological benefits of using social virtual reality platforms during the covid-19 pandemic: The role of social and spatial presence. *Computers in Human Behavior* 127 (2022), 107047.
[6] Doug A Bowman and Ryan P McMahan. 2007. Virtual reality: how much immersion is enough? *Computer* 40, 7 (2007), 36–43.
[7] Community. 2023. Voice chat | Phasmophobia Wiki | Fandom. https://phasmophobia.fandom.com/wiki/Voice_chat#Voice_recognition
[8] Samuel Rhys Cox and Wei Tsang Ooi. 2023. Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. In *International Workshop on Chatbot Research and Design*. Springer, 167–184.
[9] Feral Interactive (Linux) Creative Assembly, Feral Interactive (Mac). 2014. Alien: Isolation. [Steam]. https://store.steampowered.com/app/214490/Alien_Isolation/
[10] James J Cummings and Jeremy N Bailenson. 2016. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media psychology* 19, 2 (2016), 272–309.
[11] Frontier Developments. 2015. Elite Dangerous. [Steam]. https://store.steampowered.com/app/359320/Elite_Dangerous/
[12] Sony Interactive Entertainment Europe. 2019. Blood & Truth™. [Playstation Store]. https://store.playstation.com/da-dk/product/EP9000-CUSA11098_00-0000000000000000
[13] Fireproof Games. 2020. The Room VR: A Dark Matters. [Steam]. https://store.steampowered.com/app/1104380/The_Room_VR_A_Dark_Matter/
[14] Hello Games. 2016. No Man's Sky. [Steam]. https://store.steampowered.com/app/275850/No_Mans_Sky/
[15] Insomniac Games. 2019. Stormland. [Meta Store]. https://www.meta.com/da-dk/experiences/pcvr/1360938750683878/?utm_source=www.vrgamerankings.com&utm_medium=oculusredirect
[16] Kinetic Games. 2020. Phasmophobia. [Steam]. https://store.steampowered.com/app/739630/Phasmophobia/
[17] Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, and Jianfeng Gao. 2023. MindAgent: Emergent Gaming Interaction. arXiv:2309.09971 [cs.AI] https://arxiv.org/abs/2309.09971
[18] Alessandra Gorini, Claret S Capideville, Gianluca De Leo, Fabrizia Mantovani, and Giuseppe Riva. 2011. The role of immersion and narrative in mediated presence: the virtual hospital experience. *Cyberpsychology, behavior, and social networking* 14, 3 (2011), 99–105.
[19] Insook Han, Hyoung Seok Shin, Yujung Ko, and Won Sug Shin. 2022. Immersive virtual reality for increasing presence and empathy. *Journal of Computer Assisted Learning* 38, 4 (2022), 1115–1126.
[20] Wijnand A IJsselsteijn, Yvonne AW De Kort, and Karolien Poels. 2013. The game experience questionnaire. (2013).
[21] ILMxLAB. 2019. Vader Immortal: Episode I. [Meta Store]. https://www.meta.com/da-dk/experiences/2108775495884888/?utm_source=www.vrgamerankings.com&utm_medium=oculusredirect
[22] Skydance Interactive. 2020. The Walking Dead: Saints & Sinners. [Steam]. https://store.steampowered.com/app/916840/The_Walking_Dead_Saints__Sinners/
[23] Gary Kacmarcik. 2021. Using Natural Language to Manage NPC Dialog. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 2, 1 (2021), 115–117.
[24] Clark R Karr, Douglas Reece, and Robert Franceschini. 1997. Synthetic soldiers [military training simulators]. *IEEE spectrum* 34, 3 (1997), 39–45.
[25] Taeyong Kim and Frank Biocca. 1997. Telepresence via Television: Two Dimensions of Telepresence May Have Different Connections to Memory and Persuasion.[1]. *Journal of computer-mediated communication* 3, 2 (1997), 0–0.
[26] Lawrence J Klinkert, Stephanie Buongiorno, and Corey Clark. 2024. Driving generative agents with their personality. *arXiv preprint arXiv:2402.14879* (2024).
[27] Jeroen S Lemmens, Monika Simon, and Sindy R Sumter. 2022. Fear and loathing in VR: the emotional and physiological effects of immersive games. *Virtual reality* 26, 1 (2022), 223–234.
[28] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *Comput. Surveys* 56, 9 (2024), 1–39.
[29] Jiping Li and Xiaohui Yang. 2021. Game-Based 3D Virtual Environment for Learning Japanese Language and Culture. In *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering* (Xiamen, China) *(EITCE '20)*. Association for Computing Machinery, New York, NY, USA, 1147–1152. https://doi.org/10.1145/3443467.3443920
[30] Zhongyang Li, Xiao Ding, and Ting Liu. 2021. TransBERT: A Three-Stage Pre-Training Technology for Story-Ending Prediction. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 20, 1, Article 16 (mar 2021), 20 pages. https://doi.org/10.1145/3427669
[31] CE EUROPE LIMITED. 2022. RESIDENT EVIL 7 biohazard. [Playstation Store]. https://store.playstation.com/da-dk/product/EP0102-PPSA04401_00-BH70000000000001
[32] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication* 3, 2 (09 1997), JCMC321. https://doi.org/10.1111/j.1083-6101.1997.tb00072.x
[33] Firesprite Ltd. 2020. The Persistence. [Steam]. https://store.steampowered.com/app/1140100/The_Persistence/
[34] Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A. Santos, João Dias, and Ana Paiva. 2022. FAtiMA Toolkit: Toward an Accessible Tool for the Development of Socio-Emotional Agents. *ACM Trans. Interact. Intell. Syst.* 12, 1, Article 8 (mar 2022), 30 pages. https://doi.org/10.1145/3510822
[35] Grant McClure, Maiga Chang, and Fuhua Lin. 2013. MAS controlled NPCs in 3D virtual learning environment. In *2013 International Conference on Signal-Image Technology & Internet-Based Systems*. IEEE, IEEE, 1026–1033.
[36] Daniel R Mestre. 2015. On the usefulness of the concept of presence in virtual reality applications, Vol. 9392. SPIE, 93920J–93920J–9.
[37] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems* 77, 12 (1994), 1321–1329.
[38] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. https://doi.org/10.1145/3544548.3581225
[39] Jakob Nielsen. 1993. Response Times: The 3 Important Limits. https://www.nngroup.com/articles/response-times-3-important-limits/. Accessed: 2023-12-12.
[40] Niels Chr. Nilsson, Rolf Nordahl, and Stefania Serafin. 2016. Immersion revisited: A review of existing definitions of immersion and their relation to different theories of presence. *Human Technology* 12, 2 (2016), 108–134.
[41] Oculus. n.d.. Voice SDK Overview | Oculus Developers. https://developer.oculus.com/documentation/unity/voice-sdk-overview/
[42] OpenAI. 2022. ChatGPT. https://chat.openai.com/
[43] OpenAI. 2023. *Introducing ChatGPT and Whisper APIs*. https://openai.com/blog/introducing-chatgpt-and-whisper-apis Accessed: 2023-12-07.
[44] David Ota, Bowen Loftin, Tim Saito, Robert Lea, and James Keller. 1995. Virtual reality in surgical education. *Computers in biology and medicine* 25, 2 (1995), 127–137.
[45] Federica Pallavicini, Alessandro Pepe, and Maria Eleonora Minissi. 2019. Gaming in virtual reality: What changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simulation & Gaming* 50, 2 (2019), 136–159.
[46] Tommaso Palombi, Federica Galli, Francesco Giancamilli, Monica D'Amico, Fabio Alivernini, Luigi Gallo, Pietro Neroni, Marco Predazzi, Giuseppe De Pietro, Fabio Lucidi, et al. 2023. The role of sense of presence in expressing cognitive abilities in a virtual reality task: an initial validation study. *Scientific Reports* 13, 1 (2023), 13396.
[47] K Poels, Y de Kort, and WA IJsselsteijn. 2008. FUGA-the fun of gaming: measuring the human experience of media enjoyment. *Deliverable D3* 3 (2008).
[48] Polyarc. 2018. Moss. [Steam]. https://store.steampowered.com/app/846470/Moss/
[49] Sudha Rao, Weijia Xu, Michael Xu, Jorge Leandro, Ken Lobb, Gabriel DesGarennes, Chris Brockett, and Bill Dolan. 2024. Collaborative Quest Completion with LLM-driven Non-Player Characters in Minecraft. *arXiv preprint arXiv:2407.03460* (2024).
[50] Vertical Robot. 2018. Red Matter. [Steam]. https://store.steampowered.com/app/966680/Red_Matter/
[51] Aitor Rovira, David Swapp, Bernhard Spanlang, and Mel Slater. 2009. The Use of Virtual Reality in the Study of People's Responses to Violent Incidents. *Frontiers in behavioral neuroscience* 3 (2009), 59–59.
[52] Sanzaru. 2019. Asgard's Wrath. [Meta Store]. https://meta.com/da-dk/experiences/pcvr/1180401875303371/?utm_source=www.vrgamerankings.com&utm_medium=oculusredirect
[53] Guilherme Silva and Marcos Souza Ribeiro. 2021. Development of Non-Player Character with Believable Behavior: a systematic literature review. In *Anais Estendidos do XX Simpósio Brasileiro de Jogos e Entretenimento Digital* (Online). SBC, Porto Alegre, RS, Brasil, 319–323. https://doi.org/10.5753/sbgames_estendido.2021.19660
[54] Mel Slater. 2003. A note on presence terminology. *Presence connect* 3, 3 (2003), 1–5.
[55] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (2009), 3549–3557.
[56] Mel Slater. 2018. Immersion and the illusion of presence in virtual reality. *The British journal of psychology* 109, 3 (2018), 431–433.

[57] Bradley Standen, John Anderson, Alexander Sumich, and Nadja Heym. 2023. Effects of system-and media-driven immersive capabilities on presence and affective experience. *Virtual Reality* (2023), 1–14.
[58] Bethesda Game Studios. 2017. Fallout 4 VR. [Steam]. https://store.steampowered.com/app/611660/Fallout_4_VR/
[59] Bethesda Game Studios. 2018. The Elder Scrolls V: Skyrim VR. [Steam]. https://store.steampowered.com/app/611670/The_Elder_Scrolls_V_Skyrim_VR/
[60] Ayoung Suh and Jane Prophet. 2018. The state of immersive technology research: A literature analysis. *Computers in Human Behavior* 86 (2018), 77–90. https://doi.org/10.1016/j.chb.2018.04.019
[61] Nordic Trolls. 2020. Karnage Chronicles. [Steam]. https://store.steampowered.com/app/611160/Karnage_Chronicles/
[62] Anders Tychsen, Susana Tosca, and Thea Brolund. 2006. Personalizing the player experience in MMORPGs. In *Technologies for Interactive Digital Storytelling and Entertainment: Third International Conference, TIDSE 2006, Darmstadt, Germany, December 4-6, 2006. Proceedings 3*. Springer, Springer-Verlag Berlin Heidelberg, 253–264.
[63] Valve. 2020. Half-Life: Alyx. [Steam]. https://store.steampowered.com/app/546560/HalfLife_Alyx/
[64] Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-Tuning GPT-2 on Annotated RPG Quests for NPC Dialogue Generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games* (Montreal, QC, Canada) *(FDG '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3472538.3472595
[65] Valentijn T Visch, Ed S Tan, and Dylan Molenaar. 2010. The emotional and cognitive effect of immersion in film viewing. *Cognition and Emotion* 24, 8 (2010), 1439–1445.
[66] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 65, 7 pages. https://doi.org/10.1145/3613905.3651026
[67] Hongyu Wan, Jinda Zhang, Abdulaziz Arif Suria, Bingsheng Yao, Dakuo Wang, Yvonne Coady, and Mirjana Prpa. 2024. Building LLM-based AI Agents in Social Virtual Reality. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
[68] Weilai Xu, Fred Charles, and Charlie Hargood. 2023. Generating Stylistic and Personalized Dialogues for Virtual Agents in Narratives. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* (London, United Kingdom) *(AAMAS '23)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 737–746.
[69] Pinar Yanardag, Manuel Cebrian, and Iyad Rahwan. 2021. Shelley: A Crowd-Sourced Collaborative Horror Writer. In *Proceedings of the 13th Conference on Creativity and Cognition* (Virtual Event, Italy) *(C&C '21)*. Association for Computing Machinery, New York, NY, USA, Article 11, 8 pages. https://doi.org/10.1145/3450741.3465251
[70] Gabriel Zamora. 2024. The Best VR Games for 2024. https://www.pcmag.com/picks/the-best-vr-games
[71] Stress Level Zero. 2019. BONEWORKS. [Steam]. https://store.steampowered.com/app/823500/BONEWORKS/
[72] Shijie Zheng, Keith He, Le Yang, and Jie Xiong. 2024. MemoryRepository for AI NPC. *IEEE Access* (2024).