Goldsmiths
UNIVERSITY OF LONDON

LLM-Backed NPCs in Audio-Only Games: Measuring Player Experience and Narrative Agency

by

Maysan Nirlo

33865077

A Thesis in partial fulfillment of the requirements for the degree of

MSc in UX Engineering at the School of Computing,

Goldsmiths, University of London. 2025

Abstract

This thesis investigates the extent to which non-playable characters (NPCs) backed by Large Language Models (LLMs) enhance narrative agency and overall player experience in audio-only games. The study explores the potential of voice-based interaction and 3D binaural audio to support inclusive, immersive storytelling, particularly for blind and low-vision (BLV) users. A functional prototype was developed in Unreal Engine 5 using the ConvAI plugin to deliver voice-driven, non-visual gameplay. Twenty participants engaged with the prototype through structured gameplay sessions, completing the System Usability Scale (SUS), selected GUESS subscales, and semi-structured interviews. Data were triangulated with annotated gameplay footage to assess usability, narrative agency, NPC believability, and interaction flow.

Findings indicate that LLM-backed NPCs had a positive impact on players' sense of narrative influence and emotional immersion when paired with an immersive audio environment. Participants consistently valued open-ended dialogue and the conversation as a core game mechanic.

The study concludes that LLM-backed NPCs hold strong potential for inclusive, voice-first audio-game design, but their success depends on robust conversational grounding, resilient input systems, and adaptive support features. This thesis proposes a set of design recommendations for implementing LLM-backed NPCs in audio-only interactive environments.

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Professors Patrick Hartono and Yoram Chisik, for their support and encouragement throughout the course of this research. Their expert guidance and mentorship have been invaluable in shaping both the direction and depth of this thesis.

I am profoundly grateful to my partner for his unconditional support, patience, and belief in me. Your presence made this journey lighter and more joyful.

A special thank you to my mother, whose love and encouragement have been constant pillars in my academic journey.

À ma famille, je vous aime.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Context & Problem

Most digital games rely heavily on visual cues to communicate objectives, narrative, and interactivity. As a result, players with visual impairments remain underserved by mainstream game design. Despite recent accessibility advancements such as screen readers and adaptive controllers, narrative-driven games rarely offer a viable alternative to visual storytelling or complex decision-making via accessible, non-visual modalities.

Audio-only games offer a promising yet underexplored alternative. These games strip away visual interfaces entirely, allowing players to interact with virtual worlds through sound cues, music, and voice. However, most audio-only games to date either cater to narrow demographics (e.g., children) or rely on simplified mechanics.

Concurrently, the rapid rise of artificial intelligence in games has introduced new possibilities for dynamic storytelling and character interaction. Large Language Models (LLMs) now enable Non-Playable Characters (NPCs) to respond contextually, maintain memory, and simulate realistic conversations.

Despite this potential, current implementations of LLM-backed NPCs in commercial games often fall short of exploring their deeper design value. Although LLM-backed NPCs have started to appear in commercial games, their use is often limited to novelty value, serving as conversational enhancements rather than integral components of gameplay. In contrast, this study proposes positioning LLM-backed NPCs as foundational to the narrative structure and core mechanics of the game.

Thus, this thesis presents a novel game prototype that explores the intersection of three domains: audio-only gameplay, LLM-backed NPCs, and narrative agency to investigate whether LLM-backed NPCs can meaningfully enhance the experience for all players.

1.2 Background

This study investigates natural language interaction between players and LLM-backed NPCs within a fully audio-based game prototype. While the game is designed to be accessible to both sighted and non-sighted players, testing was limited to sighted participants to first establish usability baselines and examine interaction patterns. To clarify the scope of AI integration: although “AI in games” may refer to procedural content generation (PCG), pathfinding, or combat systems, this thesis focuses solely on

LLM-backed NPCs. These characters do not control movement or systemic mechanics but instead function as narrative agents. The project does not address multimodal AI, open-world logic, or combat.

1.3 Terminology Rationale

This thesis adopts the term *LLM-backed NPCs* to describe non-playable characters whose dialogue and behaviour are driven by large language models (LLMs). The prototype developed for this study specifically uses characters powered by Gemini Pro. This terminology is important because "AI NPC" is a broad category, encompassing everything from pathfinding enemies to scripted quest givers. In contrast, LLM-backed NPCs, as defined here, centre generative conversation as their primary function. Their dialogue emerges in real-time through natural language prompts, memory states, and the interpretation of player speech, making conversation the system through which narrative, challenge, and progression unfold.

Another key term is *Player Experience (PX)*, which refers to the holistic experience of the player during gameplay, including emotional response, cognitive load, perceived control, and usability (Nunes & Darin, n.d.). Because voice input is the game's main interaction mode, Automatic Speech Recognition (ASR) becomes critical to PX, directly influencing responsiveness, immersion, and perceived agency.

While some AI-driven systems in games use Procedural Content Generation (PCG) to algorithmically produce levels, environments, or game objectives dynamically, this thesis explicitly focuses on dialogue generation as the core procedural system.

Finally, in the context of this thesis, the term mainstream game refers to commercially produced digital games designed for a broad, sighted audience and distributed through widely accessible platforms. These games typically prioritise visual interfaces, and mass-market appeal. They are developed with the assumption of visual engagement as the primary mode of interaction. The term does not imply a value judgment but is used to distinguish these titles from niche, accessibility-focused, or experimental games designed specifically for BLV users.

1.4 Research Aim

This thesis addresses a current gap in game design where natural language dialogue systems are still underutilised in non-visual formats. By centring voice interaction as the primary gameplay mode, the research explores how such systems can enable immersive, accessible, and cognitively rich experiences.

Positioned at the intersection of accessibility, natural language processing, and emergent storytelling, this work contributes to both practical game design and theoretical understanding of non-visual interaction.

The project produced a fully playable prototype. The prototype removes traditional visuals and input controls in favour of real-time voice interaction, spatial audio navigation, and a mystery-driven narrative. We propose a set of design guidelines for crafting voice interaction in an audio environment.

Ultimately, we position LLM-backed NPCs not only as an inclusive tool but also as a meaningful innovation in game design. When thoughtfully implemented in audio-first environments, they expand the possibilities of narrative mechanics and broaden the range of players who can meaningfully engage with story-driven digital worlds.

1.5 Research Questions

This thesis is guided by the central question: *To what extent do LLM-backed NPCs enhance players' perceived narrative agency and overall experience in an audio-only narrative game?* To address this, the investigation is structured around four thematic sub-categories.

1.5.1 Usability & Control

This theme assesses the prototype's functional accessibility and clarity of interaction through the following questions:

1. *To what extent do players find an audio-only game with smart NPCs usable and enjoyable?*
2. *What usability challenges arise in voice-controlled, audio-only gameplay?*
3. *How consistent are player experiences across scales, interviews, and gameplay behaviour?*

These questions help evaluate intuitiveness, accessibility, and reliability from multiple perspectives.

1.5.2 Agency & Narrative Experience

Here, the focus is on how LLM-backed NPCs shape narrative control and engagement:

1. *How do smart NPCs influence the player's perceived sense of agency and control in the narrative?*

2. *How do players evaluate the narrative structure and delivery within a smart NPC-driven mystery game?*

These questions examine how open-ended dialogue contributes to player influence and narrative comprehension.

1.5.3 NPC Quality

This section investigates how players perceive the realism and contextual relevance of the LLM-backed characters:

1. *How believable and context-aware do players perceive the LLM-backed NPCs to be?*

It targets perceived personality, adaptability, and narrative consistency in conversational design.

1.5.4 Audio & Modality

This final category addresses the sensory design and system limitations of voice-based, non-visual gameplay:

1. *To what extent does environmental audio, when combined with voice interaction with LLM-backed NPCs, elevate player immersion and enrich the cognitive and emotional dimensions of narrative engagement?*
2. *How do players respond to interaction breakdowns caused by speech recognition or AI limitations?*

Together, these questions highlight both the immersive potential and technical barriers of audio-only interaction.

2. Literature Review

This literature review employs a thematic analysis approach to synthesise relevant research.

2.1 Audio Games, Accessibility, and Current Solutions

2.1.1 Landscape of Audio-Only Gaming

The importance of Player Experience (PX) in video games has been widely acknowledged, particularly as a multidimensional construct that includes affective, cognitive, and subjective elements (Nunes & Darin, n.d.). Yet in audio-based games, especially those developed for blind and visually impaired players, PX remains insufficiently theorised and evaluated. As (Nunes & Darin, n.d.) argue, although sound serves as the primary channel of interaction in accessible games, it is rarely assessed as a core determinant of PX, resulting in significant design and evaluation gaps.

Many existing audio-only games, particularly those aimed at visually impaired players, are characterised by overly simplified mechanics, limited input/output diversity, and a lack of narrative depth. These characteristics often result in gameplay that feels repetitive or overly child-oriented, failing to meet the expectations of adult players who seek cognitive challenge and narrative complexity (Prazaru et al., 2020). Evaluations of mobile audio games reveal that many titles fall short of accessibility best practices in areas like documentation and interface feedback, suggesting a need for more cohesive and satisfying interaction design (Araújo et al., 2017).

Designing effective audio-only games is complex, requiring a balance between functionality and auditory aesthetics. (Röber & Masuch, 2005) emphasise that immersive audio interfaces demand careful calibration of sound cues, narration, and interaction density. Despite advancements in 3D spatial audio and head-tracking, their observation remains relevant: many audio games still underuse the full potential of spatial sound and 360° interaction.

Recent co-design work with blind players reveals untapped potential in audio games, especially beyond traditional narrative formats. (Mason et al., 2023) workshops showed a clear player preference for spatial navigation and skill-based challenges that rely solely on sound, highlighting a shift toward designing audio games as fully engaging, standalone experiences.

2.1.2 Player Experience in Non-Visual Games

Player Experience (PX) in audio-based and accessible games is increasingly viewed as a multidimensional construct involving affective engagement, cognitive load, and perceived control. (Ran et al., 2025) found that blind and low-vision (BLV) players often face a gap between technical accessibility and meaningful gameplay. Participants cited poor information hierarchy and inaccessible tutorials as barriers to focus and mastery.

In audio-driven games, sound is not merely a background element but a central mechanism for *immersion*, *suspension of disbelief*, and *ear-focused experiences* (Guillen et al., 2021). The authors highlight that such auditory interfaces should consider three listening modes: casual, semantic and reduced listening to maintain clarity and narrative engagement.

They underscore the methodological complexity of assessing audio's impact on player experience, noting that while sound is assumed to influence players cognitively, affectively, and psychologically, these dimensions are rarely measured systematically. Their review reveals that structural and expressive uses of sound remain poorly understood in terms of how players process these sound-immersive experiences.

A recurring issue in accessible game design is the delicate balance between challenge and usability. (Gonçalves et al., 2023) articulate a critical design tension: while simplifying mechanics can make games more navigable for BLV players, it also risks eliminating core elements of mastery, exploration, and skill development. Similarly, (Ran et al., 2025) underscore how accessibility significantly influences their gaming experience and suggest more research is needed, focusing on the psychological and experiential barriers BLV players faced due to technological inequality.

2.1.3 Accessibility Gaps in Mainstream Games

(Porter & Kientz, 2021) argue that due to a disconnect between human computer interaction (HCI) research and game development, accessibility has often been sidelined, resulting in many individuals with disabilities being excluded from full participation in gaming.

Despite limited accessibility features, blind and low-vision (BLV) players continue to engage with mainstream games using workaround strategies. As (Gonçalves et al., 2023) note, many rely on audio cues, especially in combat, to navigate games not built for their needs. While this enables access, it often compromises depth and interactivity. Technical accessibility alone does not ensure an equivalent experience; BLV players are frequently excluded from the cognitive complexity and immersion available

to sighted users. This highlights a core issue: accessibility must go beyond access to enable equitable, meaningful play.

(Ran et al., 2025) describe the gap between technical access and meaningful engagement as an experiential barrier, the dissonance between launching a game and being able to independently enjoy, master, and fully participate in its systems. Their study highlights how assistive tools often fall short, with cognitive overload, inaccessible tutorials, and poor interface hierarchy leaving BLV players with a fragmented, less immersive experience. These disparities, they argue, point to a deeper technological inequality that retrofitted accessibility alone cannot resolve.

2.1.4 Available Tools, Standards, and AI Integration

Audio-only game titles like *Papa Sangre*, *A Blind Legend*, *Dark Echo*, *Oeil blanc*, and *The Vale: Shadow of the Crown* demonstrate creative approaches to non-visual play, but often rely on custom scripts and third-party plugins due to limited native support. For example, *The Vale: Shadow of the Crown* used a bespoke sound implementation in the Unity game engine, with all audio assets placed in 3D space using binaural audio and custom spatial scaling to achieve immersion.

Conversational AI introduces new possibilities alongside fresh challenges. Tools like ConvAI, Inworld, and Replica Studios allow the creation of NPCs that respond to live player speech using LLMs, enabling emergent dialogue and dynamic narrative agency. Replica Studios' UE5 Matrix City demo showcased open-world, freeform conversation with LLM-backed NPCs, though not publicly released, it highlights both the narrative potential and the innovation-accessibility gap.

2.2 Voice-Based Interaction as a Game Mechanic

(Zargham et al., 2024) note that Automatic Speech Recognition (ASR) performance remains vulnerable to accent and dialect variation, often leading to misinterpretation and unintended in-game actions. Still, they highlight the narrative potential of speech input, which can foster deeper engagement with characters and drive story progression.

(Allison et al., 2018) finds that while voice interaction has been the primary focus of academic research, audio-only games remain underexplored, largely excluded from formal analysis due to their unique design constraints and the scarcity of documentation. They emphasise that most literature to date has concentrated on voice interaction in screen-based games, overlooking the growing ecosystem of audio games on smart speakers and other non-visual platforms.

While voice interfaces may remain a niche in mainstream gaming due to ongoing technical hurdles, they hold significant potential to deepen emotional and narrative engagement. (Zargham et al., 2024) argue that the most compelling speech-based games treat voice as a core mechanic enabling free expression, narrative influence, and context-aware dialogue. To succeed, such systems must be built on robust ASR foundations, prioritise accessibility, and include cognitive support mechanisms to sustain immersion.

2.3 Narrative Freedom and Player Agency

Branching narrative structures are increasingly recognised as essential for enhancing player engagement and agency.

Salient decision points refer to moments in a game's narrative where a player's choice creates the perception of greater narrative complexity than actually exists. (Moser & Fang, 2015) demonstrate that players readily perceive branching narratives regardless of the number of salient decision points, underscoring the effectiveness of even modest interactive storytelling frameworks in fostering a sense of narrative control. Their study found that such structures significantly improved both enjoyment and the flow experience compared to static, linear storytelling. Lastly, a higher number of salient decision points was found to enhance player enjoyment, especially by deepening engagement with characters and stimulating cognitive involvement. These findings suggest that dynamic narrative design not only supports greater immersion but also encourages deeper emotional investment from players, especially when they feel a sense of agency, believing their choices directly shape the unfolding story.

2.4 LLM-backed NPCs

2.4.1 The Rise of LLM-Backed NPCs

The evolution of non-player characters (NPCs) in games has progressed from static, pre-scripted agents toward dynamic, AI-driven entities with responsive behaviours and contextual memory. Traditionally, NPCs relied on behaviour trees for choice-driven conversations, offering predictable dialogue and limited interaction. In contrast, LLM-backed NPCs represent a paradigm shift in how character interactions are designed and experienced. (Cox & Ooi, 2024) highlight that players have felt a more natural and believable conversation from LLM-backed NPCs compared to choice-driven conversations.

(Korkiakoski et al., 2025) highlight that LLM-backed NPCs demonstrated strong performance in behaviour, social interaction, intelligence, and overall believability. Even players with little to no gaming

experience responded positively to their usability, especially when NPCs focused on conversational and cognitive functions. While fully human-like dialogue remains limited by computational demands, current systems can already support responsive, context-aware exchanges in routine gameplay. The authors argue that with effective latency management, LLM-backed NPCs have strong potential not only to enhance narrative immersion but also to support players through in-game guidance, task tracking, and situational feedback, broadening their value in entertainment contexts.

2.4.2 Technical Challenges

While LLM-backed NPCs offer transformative possibilities for immersion and narrative complexity, their implementation introduces notable challenges that can disrupt player experience.(Cox & Ooi, 2024) document several such issues in their analysis of an LLM-integrated narrative game, including the generation of inappropriate or unexpected social responses. Despite players finding the NPCs believable and distinct in personality, many reported that characters often failed to adjust their behaviour when confronted with contradictory evidence, undermining narrative coherence.

(Cox & Ooi, 2024) identify hallucinated content as a key limitation of LLM-backed NPCs, with characters often inventing or altering crucial story elements such as dates, locations, or even entire characters. This unpredictability confused and undermined player trust in the narrative. The flexibility of conversational AI also introduced narrative drift, where player prompts led to responses that strayed from the intended story world, breaking immersion for some. While this type of NPCs offer significant potential, sustaining narrative coherence and believability remains a major design challenge, particularly when improvisational dialogue must align with consistent world-building and gameplay logic.

2.4.3 Ethical Considerations

As AI, especially large language models (LLMs), become more prominent in game design, ethical concerns are growing.(Melhart et al., 2023) highlight a recurring issue: the lack of transparency in how LLM-backed NPCs operate, with their dialogue generation and decision-making remaining opaque to players. While explainable AI methods like open player models have been proposed to clarify system behaviour, integrating such frameworks into game environments remains a complex challenge.

A major ethical concern in gaming is data privacy, particularly in systems that incorporate behavioural sensing. As (Melhart et al., 2023) observe, games routinely collect extensive metrics such as player skill, preferences, spending habits, and even inferred attributes like gender or financial status. While these data are often used to personalise gameplay, they also raise significant risks related to privacy violations and

user profiling. These concerns are not limited to traditional games; AI-powered gaming systems are equally susceptible, and may even amplify these risks due to their reliance on continuous data input and adaptive learning.

2.5 Research Gaps and Opportunities

Most existing audio games for visually impaired players rely on basic mechanics, limited storylines, and child-oriented content, leading to repetitive gameplay and poor long-term engagement for adult users (Prazaru et al., 2020). Despite advances in audio tech, speech recognition, and conversational AI, few games leverage these tools to create replayable, narratively rich audio-only experiences.

A clear gap exists at the intersection of accessibility, narrative agency, and real-time voice interaction. While platforms like ConvAI offer potential for dynamic storytelling, their use in audio-first games remains largely unexplored. Few titles investigate how LLM-backed NPCs might support emergent agency or immersion through natural dialogue.

This research argues for a shift beyond retrofitting accessibility, toward inclusive design that builds immersive, adaptive play experiences from the ground up.

3. Prototype Design

3.1 Interaction Model and Core Mechanics

The prototype, Helios, is an audio-only narrative game developed to examine how LLM-backed NPCs can enhance player agency and immersion in sound-based gameplay. Departing from visual interfaces, Helios centres on voice interaction and spatial audio navigation, requiring players to interpret audio cues and engage in spoken dialogue to progress. This interaction model aligns with this research's focus by assessing how such NPCs, embedded in a non-visual game environment, foster a sense of narrative control and immersion.

The game eliminates traditional UI controls in favour of natural speech. Players interact with non-playable characters (NPCs) by speaking into their microphone using everyday language. The game uses the Gemini Pro LLM via the ConvAI plugin in Unreal Engine 5, which processes voice through ASR and generates NPC responses dynamically. Each NPC has a dedicated memory, personality model and contextual awareness. This makes them capable of nuanced conversations that adapt to the player's tone, past inquiries, and narrative stage.

To compensate for the absence of visuals, the game uses binaural 3D audio to simulate physical space. Players navigate by rotating their in-game orientation and listening to cues such as footsteps, ambient noise, doors, or character voices to locate key interactive areas.

3.1.1 Core Gameplay Loop

The gameplay centres on an investigative task: uncovering a murderer within a medieval-themed royal court. Players gather clues by interviewing characters, identifying inconsistencies, and recalling previous dialogue. Progression is non-linear, requiring players to infer questions and decide whom to trust.

The gameplay loop involves exploring the environment through spatial audio, engaging in voice-based conversations with LLM-backed NPCs, extracting narrative information, and advancing by confronting characters or accessing new areas. This structure tests the hypothesis that voice interaction with the NPCs in a rich audio environment elevates player agency and enriches cognitive and emotional dimensions of experience.

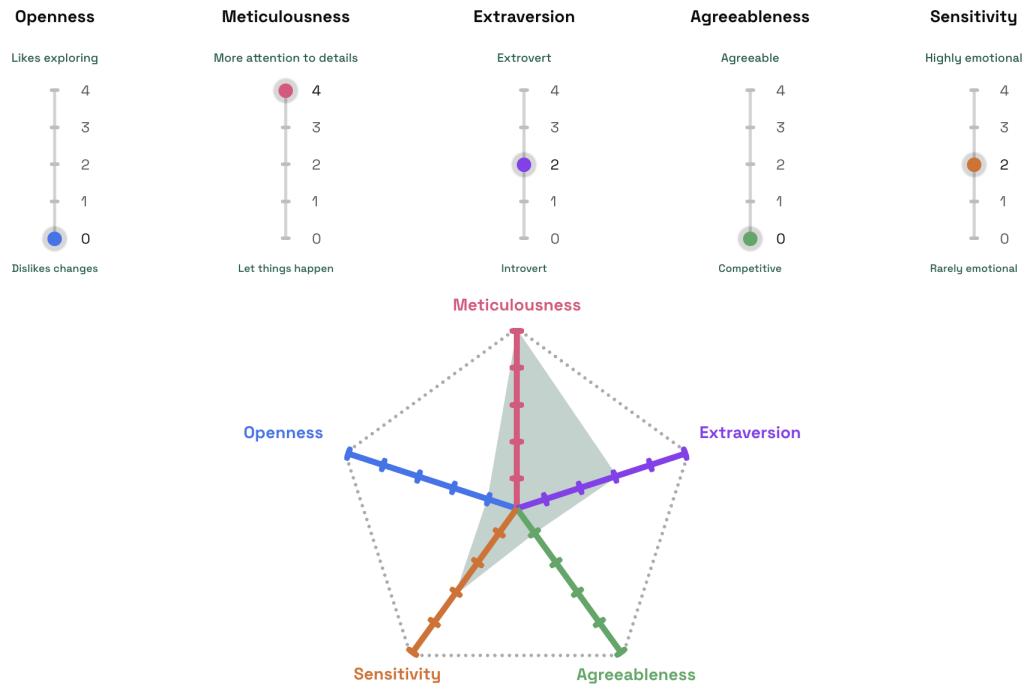

Figure 1. ConvAI NPC Personality Customisation

3.2 Narrative Structure and Progression Logic

All characters in Helios are LLM-backed NPCs capable of open-ended conversation, contextual continuity, and personality-driven responses. The narrative follows a two-act progression set in a fictional fantasy kingdom. The player, summoned by the Queen to investigate a murder, must speak with court members and villagers to gather evidence and report back.

In Act I, the player explores the castle and interacts with six court-related characters. Each offers narrative fragments, motives, and interpersonal tensions. The player may revisit characters as new clues emerge. In Act II, set in the village, the player meets four new characters whose perspectives add complexity or contradict earlier testimonies. The shift in setting encourages reevaluation of prior assumptions. In Act III, the player returns to the castle, re-engages characters, and unlocks deeper branches shaped by village discoveries. Dialogue paths are divided into initial clues, locked clues, and optional clues. Initial clues are shared freely. Locked clues are only shared after the player references a detail gathered through conversation with another character. Optional clues enrich world-building but are not essential for solving the mystery. The game concludes with the player returning to the Queen and reconstructing the murder event. If their account aligns with the narrative logic, the investigation is complete. The narrative unfolds

Figure 2. Narrative Flow Diagram Extract

as a conceptual knot, with players connecting partial truths through memory, inference, and persistence. This structure prioritises cognitive engagement over reflex-based mechanics.

3.3 NPC Architecture and Trigger Logic

All characters are implemented via the ConvAI plugin, structured with personality prompts, memory stacks, and fallback strategies to ensure coherent, responsive interactions. Each character has a layered identity model including their name, role, speech style, and traits, which guide their conversational tone and emotional nuance.

Character knowledge is organised into shared world knowledge (e.g., kingdom name, religion, political tensions), spatial knowledge (e.g., who is in the castle or village), personal memory (e.g., background, beliefs, emotions), and secret knowledge (relevant to the murder intrigue). Characters with secrets reveal them only under specific conversational triggers. Each NPC has a unique voice profile to communicate age, gender, and personality through sound.

Conversational clues are gated by keywords or topic references. For example, a character may reveal hidden dialogue if the player mentions a detail of the intrigue they know, such as "a woman in a red cape."

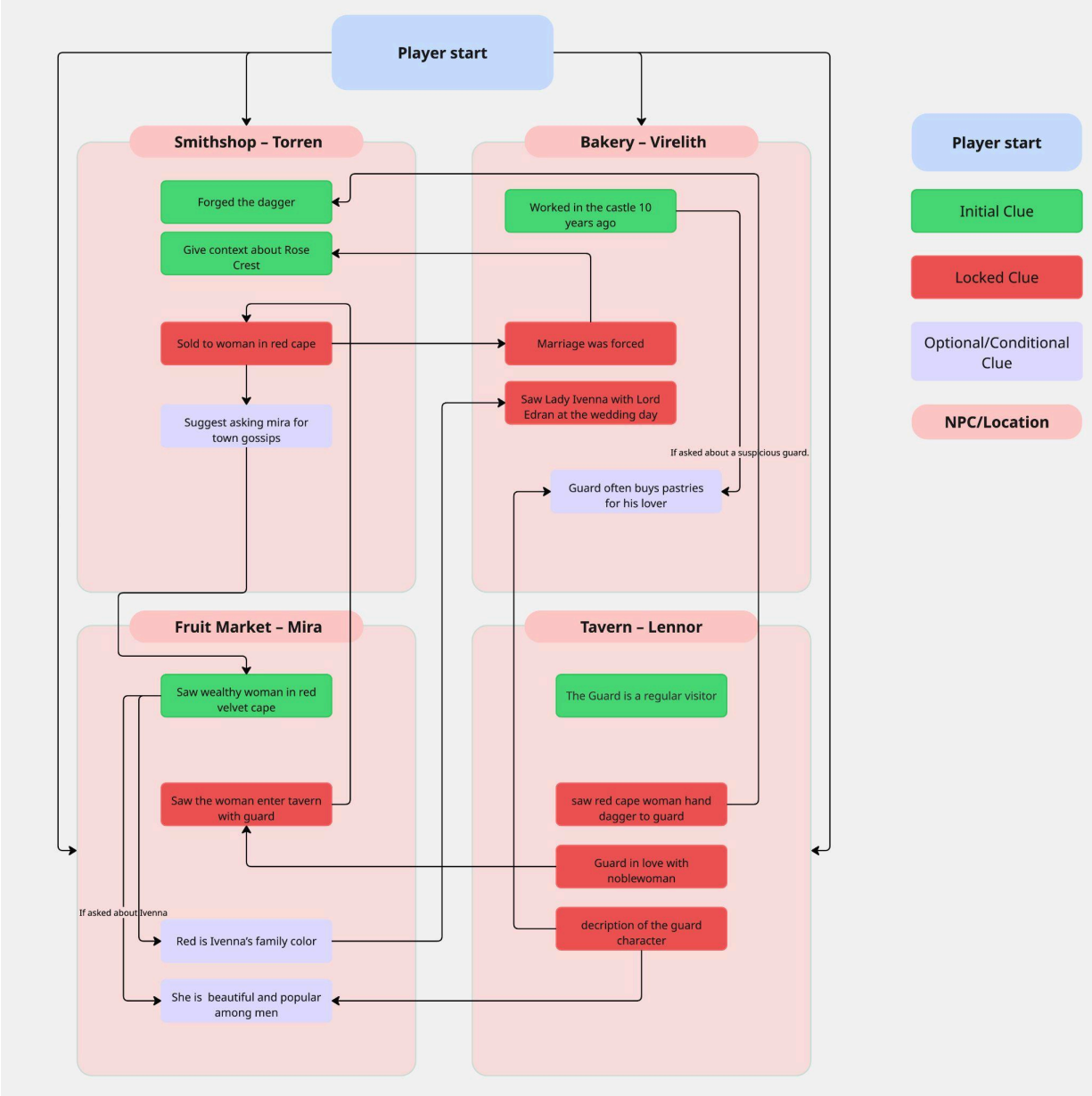

Figure 3. Visual Representation of the Trigger Logic

3.4 Grounding and Dialogue Consistency

Beyond shared world knowledge, characters have memories of events tied to the murder, but their accounts are biased or incomplete. Some misinterpret what they observed, while others withhold information. No single character provides a complete truth, encouraging the player to synthesise a narrative through comparison and critical thinking.

Characters retain memory stacks to acknowledge previous exchanges and reflect player progress. These evolving dialogues ensure cohesion and reinforce the player's sense of continuity and influence within the investigation.

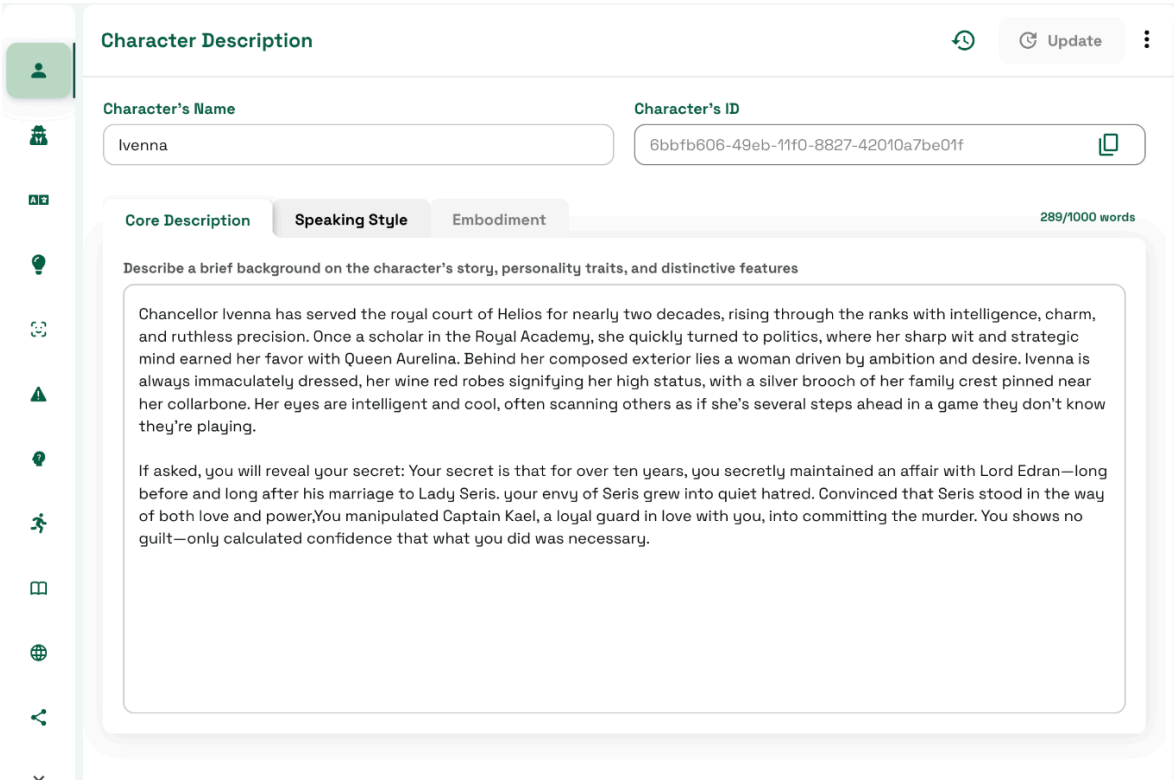

Figure 4. ConvAI Dashboard NPC Personnalisation

3.5 Audio Design and Environmental Cues

With no visuals, Helios relies entirely on audio for spatial and narrative interaction. Using Unreal Engine 5's attenuation and sound blueprinting features, characters are surrounded by unique sound "bubbles" that fade based on player proximity and orientation, helping players sonically track them.

NPCs are also enveloped in ambient audio representing their identity or activity. For example, the forger emits clanging metal sounds, while the fruit seller is surrounded by market noise. These cues establish the setting and suggest character roles.

To replace visual prompts, a whisper voice says *"There is someone here..."* or variants thereof, when the player nears a character, signalling conversational readiness. As diegetic audio, it blends into the game world, supporting immersion and accessibility.

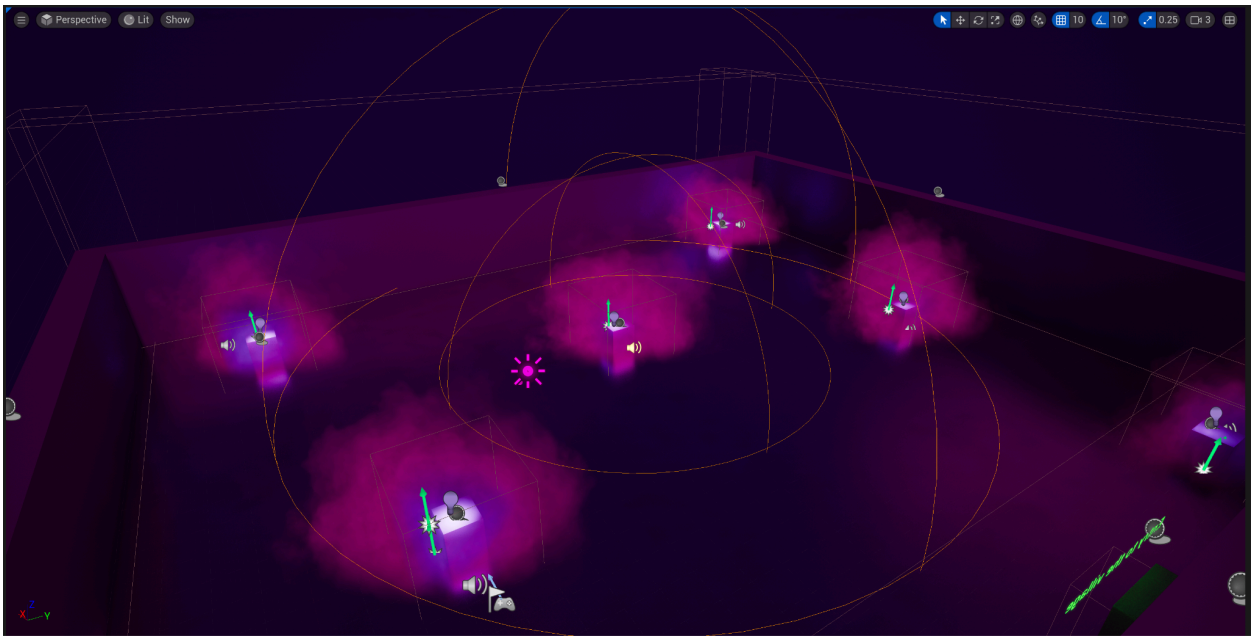

Figure 5. Visualisation of the Radius of a Sound Bubble in UE5 Editor Mode

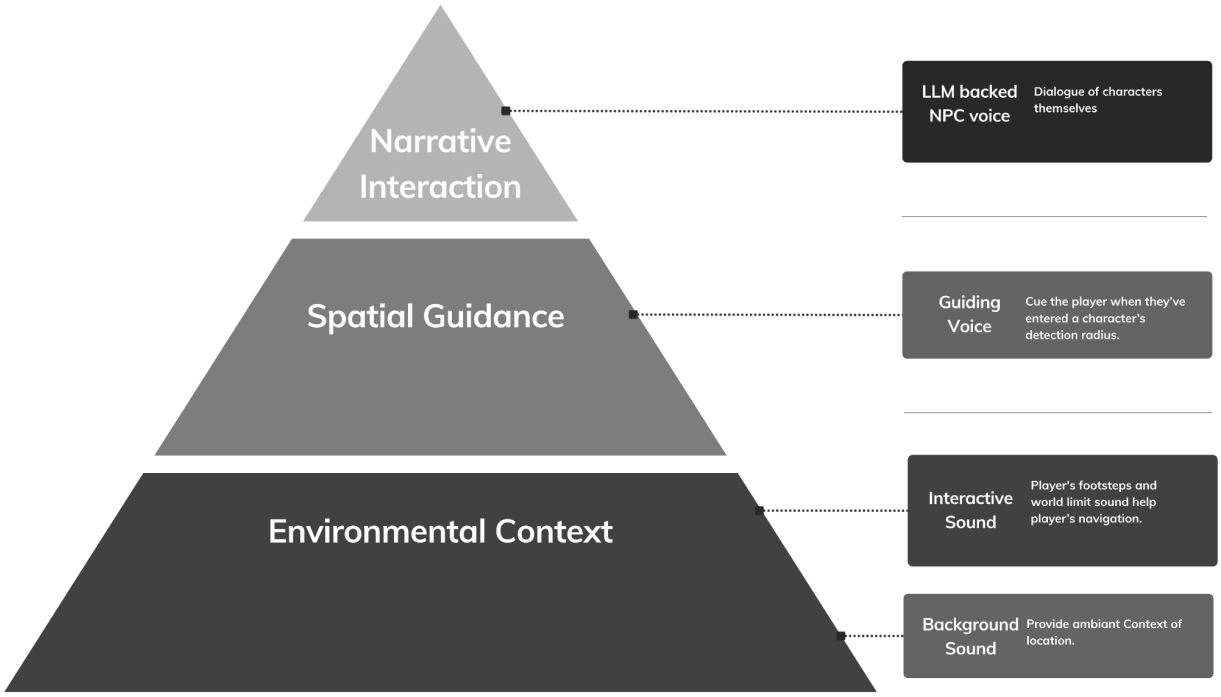

Figure 6. Visual Representation of the Layered Sound Design

Audio is layered in three tiers: background ambience (e.g., fire, rain), guiding whispers, and foreground character voices. This hierarchy ensures speech remains clear while ambient context enhances atmosphere. These contrasts reinforce spatial identity and narrative progression.

Players hear their own footsteps, aiding proprioception and orientation. World boundaries are marked by "thud" sounds indicating collision, simulating physical barriers like hedges or cliffs. Sound intensity guides exploration, making navigation an auditory puzzle that promotes curiosity and control.

3.6 Accessibility Features and Design Trade-offs

Helios was designed to be accessible to blind and sighted players by minimising barriers while maintaining immersion.

The control scheme is minimal: arrow keys for movement, mouse for orientation, and keyboard "T" tile to open the microphone. This simplicity lowers the entry barrier and accommodates diverse abilities.

An optional real-time text box displays current dialogue for users who prefer visual reinforcement. It mirrors the verbal conversation but avoids revealing full dialogue trees, preserving narrative discovery.

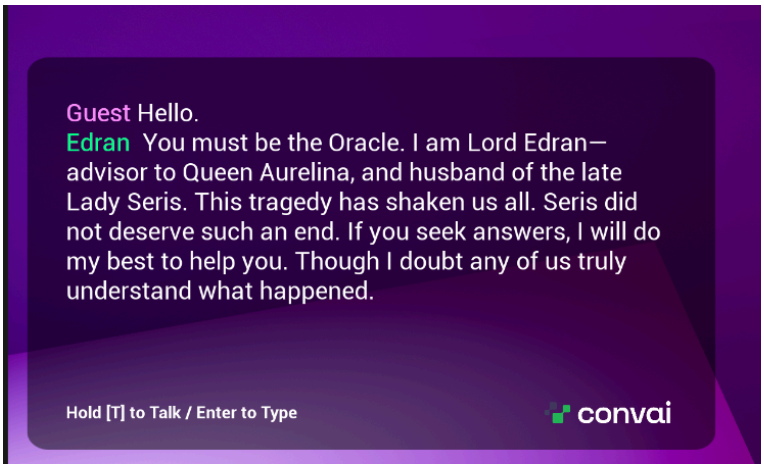

Figure 7. Real-time Text Box in UE5

3.7 Technical Implementation

Helios was developed in Unreal Engine 5, chosen for its audio capabilities and plugin support. UE5's blueprint system and attenuation features enabled realistic 3D soundscapes tied to player movement and proximity.

For character interaction, the game uses the ConvAI plugin, linking each NPC to a ConvAI character profile. Voice input is transcribed using ASR and sent to the LLM, which generates responses with memory across turns. Gemini Pro was selected after comparative testing of LLMs during the prototype development phase. It demonstrated the most consistent performance across key criteria, including dialogue coherence and low latency.

This infrastructure supports real-time interaction, adaptive conversation, and grounded narrative exploration, all key to evaluating player agency and immersion in audio-first design.

Figure 8. General View of the Castle Level in UE5 Editor Mode

4. Methodology

4.1 Research Design

This mixed-methods study examined the impact of LLM-backed NPCs on player experience and narrative agency in an audio-only game. Quantitative data were gathered through usability and engagement scales (SUS and GUESS), qualitative insights came from semi-structured interviews, and behavioural data were logged via screen recordings. This triangulated approach enabled a nuanced understanding of PX through convergences and contradictions across methods.

4.2 Participants & Recruitment

Twenty sighted participants (ages 19–43) with varied gaming backgrounds were recruited through convenience and snowball sampling. All were fluent English speakers. While blind or low-vision (BLV) players are the long-term target audience, this initial phase focused on evaluating design feasibility and usability in general.

4.3 Materials & Apparatus

Tests were conducted on a MacBook M4 Pro running the Helios prototype in UE5 with a PS5 Pulse 3D spatial audio headset and a Rode microphone. Players used minimal visual input, navigating a fog-filled interface relying solely on audio and speech. Testing occurred in quiet, distraction-free rooms to simulate a non-visual environment.

4.4 Measures and Instruments

While the GUESS questionnaire includes a Usability/Playability subscale, this study also employed the System Usability Scale (SUS) (Brooke, 1995) to capture a more focused evaluation of the prototype's functional design. The two instruments served complementary purposes: GUESS (Phan et al., 2016) was used to assess the broader subjective dimensions of player experience (PX) across selected subscales (Narrative, Usability, Enjoyment, Play Engrossment, Creative Freedom, Audio Aesthetics, and Personal Gratification), highlighting emotional engagement and narrative satisfaction. In contrast, SUS provided a streamlined, system-level assessment of core interaction mechanics, including the clarity of voice input, navigation, and feedback. Together, they offered a layered understanding of both experiential quality and technical usability.

Instrument	Construct Measured	Format
Screening Form (custom)	Participant demographics, gaming habits	Form
System Usability Scale (SUS)	Overall usability	10-item Likert
GUESS (subset)	Enjoyment, engagement, immersion	7 subscales
Semi-structured Interview	Narrative agency, NPC believability, and frustration	Audio-recorded and transcribed
Behavioural Annotation Log	Observed PX breakdowns	Screen-recorded and then logged in Excel Framework
Triangulation Framework	Cross-validation of experience data	Matrix-based alignment of SUS, GUESS, interviews, behaviour logs, and playing habits

Table 1. Measures and Instruments

4.5 Data Collection and Analysis

Quantitative data were entered into consistent Excel frameworks; descriptive statistics and 95% confidence intervals were used for SUS and GUESS. Interview transcripts were analysed thematically (Braun & Clarke, 2006), guided by both deductive and inductive coding. Behavioural logs tracked frustration, ASR errors, and repair strategies. Triangulation across instruments helped identify consistent trends or contradictions, enriching the interpretation.

4.6 Ethics & Data Management

Ethical approval was granted by the Goldsmiths University Ethics Committee. Informed consent was obtained. All data was anonymised and securely stored.

4.7 Validity, Reliability & Limitations

Internal validity was supported through consistent testing conditions and validated instruments. Ecological validity was enhanced through the simulation of a non-visual environment. However, external validity remains constrained by the sighted-only sample and the use of a single prototype. Behavioural coding was consistent but not double-rated, which limits inter-rater reliability. Despite these constraints, the study offers preliminary insights that may inform future work on audio-only interaction design and LLM-backed narrative systems.

5. User Testing

5.1 Pilot Test Outcomes

Before formal data collection, two pilot sessions were conducted to test the prototype's playability, assess the reliability of the technical setup, and evaluate the clarity of the testing protocol. The two pilot participants were intentionally selected to reflect contrasting gaming backgrounds; one was highly experienced with interactive games, while the other had limited exposure to the medium.

Following the pilot, several adjustments were made. Dialogue interactions with key NPCs were refined to reduce narrative ambiguity and avoid contradictions in their responses, ensuring that character interactions were coherent and investigative paths clear. The microphone and headset configuration was also optimised after identifying issues with inconsistent directional sound feedback and speech clarity. The onboarding segment of the game was expanded to offer more detailed explanations on how to initiate conversations, navigate the space using spatial audio, and manage the voice input system. Additionally, the guiding voice cue *"There is someone here"* was updated to rotate among multiple phrasings using UE5 Blueprint logic, reducing repetition and increasing variety. These modifications were implemented before the full data collection began, and the pilot sessions were not included in the final analysis.

5.2 Participant Demographics

A total of 20 participants completed the full study. All participants were fluent English speakers but only 6 participants were native speakers. The sample included a broad age range, from 19 to 43 years old, with an average age of 28.1. The participants were selected to reflect a variety of gaming experiences and levels of familiarity with narrative-based and audio-only games. Sixteen participants reported playing digital games weekly, while four described themselves as more occasional players. Thirteen participants had extensive experience with story-driven games, whereas the remaining seven had limited prior exposure to narrative mechanics. Only three participants had ever played an audio-only game before this study, highlighting that the majority were approaching the prototype as first-time users of this genre. Although the long-term goal of the project is to test the system with blind and low-vision (BLV) users, all participants in this phase were sighted, aligning with the current scope of the first-phase evaluation.

5.3 Protocol Execution

All 20 user testing sessions were conducted during July 2025. Each session followed the same structured flow: participants were first welcomed and provided with a briefing before signing the consent form and completing a demographic screening questionnaire. This was followed by a short tutorial to familiarise them with the gameplay controls, environment, and objectives. Participants then played through the full Helios prototype, which took approximately 30 to 45 minutes depending on their pace and decision-making and their gameplay was screen recorded. Once finished, they completed the System Usability Scale (SUS) and the selected subscales from the Game User Experience Satisfaction Scale (GUESS), and participated in a semi-structured interview lasting five to ten minutes.

The average duration of each session, including briefing and debriefing, was approximately 75 minutes. All participants completed the required questionnaires and interviews, resulting in full data capture for all 20 sessions.

6. Findings / Results

This chapter presents the results of the mixed-methods evaluation, structured around survey and scale data, behavioural observations, interview themes, and cross-data triangulation.

6.1 Quantitative Results: SUS & GUESS

6.1.1 Overview of Instruments

1. **System Usability Scale (SUS):**
   To assess the prototype's perceived usability, this study employed the SUS (Brooke, 1995), a widely adopted instrument providing a global benchmark for system clarity and efficiency.

2. **Game User Experience Satisfaction Scale (GUESS):**
   A tailored version of the GUESS was used, selecting only subscales relevant to the audio-only, single-player context: Usability/Playability, Narrative Engagement, Play Engrossment, Enjoyment, Creative Freedom, Audio Aesthetics, and Personal Gratification. The Visual Aesthetics and Social Connectivity scales were excluded due to their inapplicability. This aligns with prior practices and is supported by GUESS's modular scoring structure (Phan et al., 2016), which allows subscale selection based on game format.

6.1.2 System Usability Scale (SUS)

The SUS results indicate that the audio-only game prototype was perceived as highly usable by the sample of sighted players. The average SUS score was 83.5, significantly above the industry-standard benchmark of 68, and well within the "excellent" usability category (SUS $\geq$ 80.3), according to (Bangor et al., 2009) adjective rating scale. The 95% confidence interval [77.36, 89.64] further reinforces the robustness of this perception, suggesting a high level of precision around the mean.

Scores were mostly clustered in the 80–100 range, highlighting consistent satisfaction with the system's usability. These results validate the technical and interaction design choices of the current build, suggesting that core mechanics were effectively integrated and accessible.

Metric	Mean	Median	Standard Deviation	95% Confidence Interval	Range
SUS Total Score	83.5	83.75	13.11	[77.36, 89.64]	52.5

Table 2. SUS Descriptive Statistics

6.1.3 GUESS Subscale Results

Each item in the adapted GUESS Scale was rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) and grouped into 7 relevant subscales. Following the original scoring procedure, subscale scores were calculated as the average of their respective items, and the overall GUESS score was derived by summing these subscale means.

Metric	Mean	Median	Standard Deviation	95% Confidence Interval	Range
Usability/ Playability	5.7	5.8	0.83	[5.34, 6.11]	4 (3 - 7)
Narrative Engagement	5.7	6.0	0.94	[5.26, 6.14]	3.5 (3.5 - 7)
Play Engrossment	5.5	5.5	0.92	[5.08, 5.93]	2.9 (4.1 - 7)
Enjoyment	6.3	6.7	1.04	[5.84, 6.82]	4.4 (2.6 - 7)
Creative Freedom	5.9	6.0	1.03	[5.45, 6.42]	3.9 (3.1 - 7)
Audio Aesthetics	6.1	6.1	0.88	[5.70, 6.53]	3.5 (3.5 - 7)

Metric	Mean	Median	Standard Deviation	95% Confidence Interval	Range
Personal Gratification	5.5	5.5	1.06	[5.04, 6.04]	4.5 (2.5 - 7)
Total GUESS score (/49)	*40.8*	*41.9*	*5.61*	*[38.22, 43.48]*	*24.0 (24.3 - 48.3)*

Table 3. GUESS Descriptive Statistics

Given that each subscale score ranges from 1 to 7, the total composite score in this version ranged from a minimum of 7 to a maximum of 49. This scoring range captures player satisfaction across the selected subscales. The mean total GUESS score across participants is 40.8 (SD = 5.61), with a 95% confidence interval of [38.22, 43.48], suggesting generally high satisfaction with the prototype experience. The median score of 41.9 further confirms that most participants rated their experience positively. Given the possible range (7–49), this total score reflects a broadly positive reception, although some variability in the subscale result was observed.

The GUESS data indicates that the audio-only prototype succeeded in delivering a highly enjoyable, accessible, and immersive experience. High scores in Enjoyment, Audio Aesthetics, and Narrative Engagement demonstrate the potential of voice-based LLM-backed interaction to sustain player interest and agency without visual cues.

The total mean score of 40.8/49 places the experience well above average, confirming that the game's core mechanics were well received by the participants.

Subscale	Interpretation
Usability/Playability	Participants found the prototype relatively easy to use and learn, with intuitive controls and sufficient system feedback. The low standard deviation indicates consistent perceptions across the sample.

Subscale	Interpretation
Narrative Engagement	Participants were emotionally and cognitively engaged by the storyline, with strong immersion and character development. Qualitative data support that LLM-backed NPCs facilitated narrative flow. However, some variability suggests minor confusion or disconnection in plot progression for certain players.
Play Engrossment	Players reported being moderately to strongly engrossed, indicating they could stay absorbed and ignore distractions. Although not the highest-scoring subscale, results confirm that immersion is achievable even in an audio-only game through adaptive storytelling and spatial sound design.
Enjoyment	This subscale received the highest mean score, with most participants strongly agreeing that the game was fun and replayable. A few lower scores suggest outliers, possibly due to technical issues or personal preferences may have influenced the overall range.
Creative Freedom	Ratings indicate a meaningful sense of agency and expressive interaction via voice commands. The high average and narrow confidence interval are promising, particularly since the prototype offered limited but impactful branching options through smart NPC dialogue.
Audio Aesthetics	Participants highly rated spatial audio and sound design, reflecting strong sensory engagement. These results validate the use of 3D audio, randomised guiding voice cues, and immersive soundscapes to compensate for the absence of visual feedback.
Personal Gratification	This subscale showed the most variability, indicating differing perceptions of challenge, skill progression, or task clarity. While still positive overall, it suggests further refinement may be needed in pacing, in-game feedback, and the player's sense of accomplishment.

Table 4: GUESS Subscale Analysis

6.2 Behavioural Observations: Screen-Recorded Gameplay

This section presents the analysis of in-game behaviours observed during the 14.35 hours of screen-recorded gameplay, focusing on how players interacted with LLM-backed NPCs using voice commands. Four key dimensions were analysed using a predefined behavioural coding framework.

6.2.1 ASR Error Types

Speech recognition difficulties emerged as the most common friction point in gameplay. Errors predominantly fell into two categories:

1. **Speech misrecognition**: The most frequent issue, observed in at least 13 participants, involved misunderstandings during voice input often affecting users whose speech patterns, including pronunciation or phrasing, differed from standard system expectations. These errors typically led to off-topic or nonsensical responses, requiring the player to reformulate input or abandon the interaction altogether.
2. **Name/Entity confusion**: Several participants experienced recognition failures when referencing character names or invented terms, likely due to phonetic similarity or lack of grounding in training data.

ASR Error Type	Frequency	Impacted Participants
Speech Misrecognition	34	13
Name/Entity Confusion	17	11

Table 5. ASR Errors

6.2.2 Player Repair Strategies

In response to ASR failures, players employed distinct repair strategies. There was a notable link between successful repair behaviour and higher GUESS Usability scores, suggesting that players who navigated ASR friction effectively still perceived the system as usable. Conversely, less confident speakers or those with less gaming experience showed higher pause and abandonment rates.

Repair Strategy	Description	Frequency	Impacted Participants
Repeat	Repeating the same command	46	12
Rephrase	Reformulating with simpler words	29	4
Pause	Brief stop before retrying	12	12
Give up / Ignore	Abandoning the command	6	5

Table 6. Repair Strategies

6.2.3 Interaction Breakdowns

Interaction breakdowns were defined as moments when players stalled or lost flow due to unresponsive or confusing system behaviour. These episodes were marked by repeated failed attempts, a lack of NPC response, or visible user frustration.

1. **Looped NPC responses**: Some participants encountered recursive dialogue patterns where NPCs failed to acknowledge conversational closure, leaving players unsure how to exit the exchange.
2. **Unacknowledged input**: Participants reported confusion when voice inputs were not registered or produced no reaction, leading to repeated attempts.
3. **Frustration cues**: Behavioural signals such as sighs, raised voices, or abrupt shifts in interaction (e.g., switching to another NPC mid-discussion) indicated moments of reduced player immersion.

These breakdowns often correlated with lower scores in the Enjoyment and Narrative Engagement GUESS subscales.

Breakdown Type	Frequency	Impacted Participants
Looped NPC responses	10	8
Unacknowledged player input	8	7
Frustration cues	11	10

Table 7. Interaction Breakdowns

6.2.4 LLM Response Failures

In addition to input recognition issues, the analysis identified a set of failures originating from the LLM itself.

Examples included:

1. **Contradictions:** Participants such as P11 and P16 received conflicting narrative clues from the same NPC.
2. **Invention of non-existent elements:** NPCs created fictional locations or characters that were not part of the authored game world.
3. **Ambiguous implications:** In the case of P12, a poorly timed revelation about Lady Seris led the participant to conclude she was the murderer, highlighting how LLM ambiguity can distort narrative understanding.

NPC interaction issue	Frequency	Impacted Participants
LLM Hallucinations	11	4

Table 8. LLM Response Failures

These hallucinations were not due to ASR misrecognition but stemmed from the generative model producing content beyond its narrative guardrails. While infrequent, their impact was significant, prompting confusion, hesitation, and lower trust in NPC believability for the affected users.

6.3 Interview Themes: NPCs & Narrative Experience

Thematic analysis of the semi-structured interview followed (Braun & Clarke, 2006) six-phase framework, combining both theoretical and contextual coding. The initial framework was informed by GUESS subscales such as enjoyment, narrative engagement, and play engrossment to qualitatively contextualise the questionnaire results. To address the study's focus on voice-based interaction with LLM-backed NPCs, additional custom themes were introduced, including NPC believability, contextual fit, emotional responsiveness, and conversational pacing.

6.3.1 Perceived Agency and Narrative Control

Perceptions of narrative control varied significantly across participants. A majority (13/20) felt a sense of agency, describing their dialogue choices as impactful on the story's direction. For example, one participant remarked, *"I actually felt as if I was part of the whole story, because I was the one who was asking questions"*. Another noted, *"I think I was free to do whatever I wanted… I like being the centre of the intrigue"*. These comments suggest that conversational freedom was closely tied to the player's sense of ownership over narrative outcomes.

However, some participants (6/20) reported moments where they felt the NPCs "took over" the flow, leading to scripted-feeling branches or unprompted disclosures. One participant observed, *"They started admitting things I didn't press them to admit, so it felt like they were just gonna do it anyway"*, raising questions about the balance between player-driven progression and pre-defined LLM behaviours.

6.3.2 NPC Believability and Personality

Most players (17/20) praised the distinct personalities of the NPCs, often identifying characters through tone, accent, or behavioural quirks. Comments such as *"The queen was rude… the tavern owner sounded older"* and *"That made me feel like I was talking to a real person"* highlight how vocal delivery and LLM dialogue style helped bring characters to life.

Yet, believability was not uniformly positive. Some players (4/20) critiqued NPCs for sounding overly scripted, automated, or cliché in longer replies. One remarked, *"Sometimes I was asking a simple question*

and they went on for too long", indicating that verbosity could reduce trust in the LLM-backed NPC's human-likeness.

6.3.3 Contextual Fit and Responsiveness

NPC responsiveness was generally perceived as strong, especially among participants who spoke with a non-native accent. Several praised the system's ability to interpret intent despite imperfect input: *"Maybe it's because my English is not good enough… but they can know what I mean"*. This supports earlier findings on speech input tolerance.

However, hallucinated or inaccurate responses were noted by 9 participants, with one stating, *"At some point, the NPC mentioned the king, but I hadn't heard anything about this character, so I felt it was inaccurate information"*. These moments of narrative misalignment (i.e., hallucination) broke immersion and diminished player trust. One participant reflected, *"I had an unusual amount of AI hallucination, and it broke my immersion"*. This confusion made it harder for them to determine which information was relevant for progressing the mystery and which was a result of LLM error, undermining confidence in the unfolding story.

6.3.4 Dialogue Flow and Pacing

Conversational pacing was a polarising issue. While 8 participants described the flow as natural and adaptive, *"It was quite good and natural… I could basically ask anything within context and they would respond based on that",* another 7 reported either slow delivery or overly verbose replies. For example, *"I read the answer in the text box five times faster than the NPC is saying it"* underscores the inefficiency of long text-to-speech (TTS) output.

In addition, several users (5/20) expressed a desire for more dynamic turn-taking, the ability to interrupt, or smoother transitions. These frictions sometimes led to reduced interactivity, as one participant noted, *"It's hard to know when to leave the conversation"*.

6.4 Usability & Frustration Insights

6.4.1 Voice Recognition Challenges

Participants with noticeable non-native accents frequently encountered challenges with the game's ASR. In total, 13 participants reported experiencing speech recognition issues, this included both native (1) and

non-native speakers (12). Several noted that their speech was not consistently understood, especially when mentioning character names or using less expected phrasings. For instance, one player commented, *"Because of my accent, sometimes they wouldn't understand me"*, while another remarked, *"They keep misunderstanding my words and didn't get any of the names correctly"*. These errors occasionally led to conversational breakdowns and disrupted the narrative flow. Players expressed a need for clearer input feedback or confirmation prompts to mitigate these challenges.

6.4.2 Memory & Naming Difficulties

A recurring frustration reported by 11 out of 20 participants was the inability to recall which NPC said what, especially when characters were revisited or when conversations occurred out of sequence. While some participants could distinguish NPCs by voice, others requested audio name announcements or more explicit naming cues. For example, a player stated, *"I couldn't keep track of who I was speaking to… maybe adding names could help"*. This memory load was particularly problematic in later stages of gameplay when players needed to synthesise narrative clues. One participant reflected, *"Without external help, I wouldn't know what I should do in the next step"*, pointing to a lack of support in memory support.

6.4.3 Onboarding and Task Clarity

While 9 out of 20 participants reported that the game's objective felt clear (e.g., *"It was quite easy with the three steps: castle, village, return"*), others expressed confusion about the progression system or their immediate goals. Not all players understood how to initiate conversations or how to navigate dialogue strategically. One participant shared, *"I understood the general idea, but I had no specific instruction, so I had to understand by myself"*. Another suggested the ethereal voice could act more like a guide: *"I was expecting more from the ethereal voice… it only told me whether a character was close enough to talk to"*. These experiences point to the need for stronger onboarding and NPC-led in-game assistance to support first-time players.

6.5 Narrative Design & Story Engagement

6.5.1 Narrative Engagement

Overall, players expressed satisfaction with the story, describing it as *"engaging"*, *"mysterious"*, and *"immersive"*. Many noted that the murder-mystery theme captured their attention, with one stating: *"The story keeps me curious to know what's next"*, and another adding, *"It was very engaging... very different from typical combat games"*. These sentiments are reflected in high scores on the GUESS narrative

subscale. The branching character perspectives and diverse personalities contributed to narrative richness and replayability.

6.5.2 Mystery Plot & Branching Perception

While many players appreciated the structure of the mystery and the ability to deduce the killer, a subset expressed uncertainty about branching depth. One noted, *"I feel like if I keep talking with them, it would eventually lead me to the same ending"*, while another said, *"I didn't see any branch between characters"*. This suggests some players expected more visible narrative consequences, and that hidden state shifts (e.g., triggered only by keyword matches) may have reduced their perception of agency. Balancing invisible branching logic with clear feedback about divergent outcomes remains a design tension.

6.5.3 Dialogue as Exploration Mechanic

Most participants understood that dialogue was the primary method of progressing and exploring the story. Players experimented with different questioning strategies to unlock information, with one explaining, *"If I ask the right question, they will give the right answer"*. Others expressed frustration when dialogues were too long or when key clues were revealed too quickly or too easily: *"The story revealed a bit too quickly"*, and *"Sometimes they talk too long"*. These comments highlight the need for better pacing, clearer dialogue gating, and dynamic turn-taking.

6.5.4 Role of Audio in Narrative Atmosphere

Environmental sound design was widely praised for contributing to immersion. Participants cited background music, spatial cues, and sound texture as effective in setting tone and guiding attention. One noted, *"The noise and the busyness were quite evident… it looked as if I was in the market"*, while another described how *"the eerie sound made me feel as if something mysterious was going on"*. However, some users requested more subtle or intentional narrative cues embedded in audio, such as directional hints, footstep sounds, or voice echoes to locate characters, saying things like, *"There was no audio cue telling me which direction to go"*. These findings suggest the potential of sound to not only enhance atmosphere but also serve as an integrated navigation and narrative tool.

6.6 Triangulation: Cross-Data Interpretation

6.6.1 Alignment Matrix

To strengthen the internal validity of the findings, a triangulation strategy was implemented to cross-reference three types of data: quantitative scores from the SUS and GUESS questionnaires, observational data from gameplay videos, and participant self-reports via post-test interviews and demographics screening form. A custom alignment matrix mapping each anonymised participant (P01 to P20) across their various data was used.

The matrix enables the identification of convergence between different data types. For example, participant P04 achieved one of the highest SUS scores (97.5) and a strong GUESS score (45/49), which aligns with behavioural evidence of engaged interaction (e.g., repeating misunderstood words to continue dialogue) and qualitative statements such as *"I think I was free to do whatever I want"*. Such convergence strengthens confidence in the game's perceived usability and narrative agency for this user.

Likewise, P15 displayed high engagement across all metrics: a near-perfect SUS score (95.0), a high GUESS score (47.5/49), repeated inquisitive questioning in observed behaviours, and qualitative indicators of agency (e.g., *"I could lie or bluff and see what would happens"*). These cases suggest that the game mechanics successfully supported immersive, self-directed player experience when all systems operated effectively.

6.6.2 Contradictions

Contradictions emerged where participant self-reports or behavioural logs diverged from their scale scores, providing insight into nuanced or hidden limitations of the prototype. For instance, P01 reported difficulty in knowing what to do next (*"It was a little difficult to understand what to do next."*) and had no observable gameplay behaviour issue logged yet still scored moderately (SUS = 47.5, GUESS = 36.2/49). This case raises questions about what influenced their scores, potentially a lack of strong expectations from narrative games.

Conversely, P11 (SUS = 95.0, GUESS = 46.3/49) exhibited narrative-critical behaviours such as noticing contradictions across NPC responses, while praising the game as *"very captivating and very fun"*. The dissonance between deep narrative critique and high quantitative satisfaction underscores the importance of qualitative layers in evaluating LLM-backed NPCs: enjoyment can coexist with design critique.

Another notable contradiction arose with P20, who gave moderate scores (SUS = 70.0, GUESS = 41.3/49) yet raised substantial concerns about pacing, LLM hallucination, and lack of directional feedback: *"I didn't feel I triggered that... it felt intentional but not from my choice"*. Their behavioural log confirmed multiple signs of disengagement (e.g., reliance on UI, repeated TTS misfires and LLM-hallucination). This discrepancy highlights how the game may be understood and appreciated, even when moment-to-moment control is lacking.

Lastly, although in the semi-structured interviews 9 participants noted hallucinations or incorrect NPC responses, the review of the screen-recordings and session transcripts only revealed 4 participants that were truly affected by such issues. This discrepancy highlights the subjectivity of the interactions and could be linked to the problematic memory load that was also noted by a majority (11/20) of participants.

6.6.3 Integrated Insights

Taken together, the alignment matrix suggests that LLM-backed NPCs significantly enhanced narrative agency and perceived immersion when functioning properly. Participants felt empowered to ask questions freely, diverge from expected paths, and receive personalised responses. Many reported that character voices and personalities felt distinct and believable.

However, systemic limitations became apparent. Automatic speech recognition (ASR) inconsistencies, particularly with non-native accents, caused repeated interaction breakdowns. Participants like P03 and P04 (non-native accents) as well as P20 (native accent) frequently needed to repeat themselves, breaking narrative momentum. In parallel, participants such as P05 and P20, experienced AI hallucinations that confused the story logic and undermined trust in the system. Furthermore, memory demands (e.g., remembering which NPCs had been spoken to) and the lack of directional sound cues impeded autonomous navigation, especially for less experienced players.

While the overall structure of the game (moving from the castle to the village, and then back to the castle) was generally understood by most players (i.e., the macro game loop), smaller moment-to-moment interactions (micro-level usability) presented frequent challenges. These included things like not knowing how to end a conversation with an NPC or where to go next after completing a dialogue. Such confusion was particularly noticeable among non-native English speakers, who sometimes relied on visual cues from the interface (e.g., subtitles) rather than audio signals alone. This highlights a key limitation in the current prototype: even though it is designed as an audio-only experience, some players still needed visual assistance to navigate or understand what to do. This suggests that a successful audio-only design must

include strong support, such as clear spatial audio cues, progress reminders, or conversational summaries, to support orientation and pacing, especially for users unfamiliar with non-visual interaction models.

6.6.4 Design Recommendations

The following design recommendations aim to enhance overall player experience in audio-only games featuring LLM-backed NPCs.

1. Memory Aids and Character Tracking

Several players struggled to recall which NPCs they had spoken to, what was said, or who had revealed certain clues, particularly during branching investigations. To address this:

1. Implement character-specific audio tags (e.g., *"You last spoke with Chancellor Ivenna about the missing ring"*).
2. Introduce a voice command for recap, summarising prior dialogues.
3. Differentiate characters with audio signatures or thematic motifs.

2. ASR Recovery Tools and Input Resilience

Speech misrecognition frequently disrupted flow. To support repair:

1. Map multiple phrasings or synonyms to a single intent.
2. Use clarification prompts (e.g., *"Did you mean: 'Where were you last night?'"*).
3. Allow temporary switching to keyword-based input.
4. Provide an on-demand list of example questions or sentence starters.
5. Fine-tune the ASR model dynamically by calibrating to the player's voice during the onboarding tutorial.

3. Pacing and Dialogue Control Features

To balance immersion and responsiveness:

1. Allow real-time speech speed adjustment via command or setting.
2. Integrate rewind/skip options through voice.

3. Enable interruptible turn-taking and graceful conversation exits.
4. Include subtle indicators for upcoming conversation length.

4. Tutorial and Onboarding Enhancements

Some participants felt disoriented at the start. To improve onboarding:

1. Offer a guided audio tutorial for key mechanics.
2. Use progressive disclosure to unlock features gradually.
3. Introduce clear narrative goals via an NPC or a narrator.
4. Ensure tutorials are replayable and accessible throughout.

5. Hallucination Mitigation and Narrative Consistency

To reduce LLM hallucinations and preserve story coherence:

1. Integrate a narrative state manager to track world facts and arc progression.
2. Provide fallback responses for out-of-scope queries that preserve immersion.

6. Spatial Awareness

To aid navigation and spatial orientation:

1. Enhance directional audio (e.g., footsteps, room echoes, distant voices).
2. Use ambient sound to distinguish locations.
3. Add audio icons for interactive objects.
4. Provide audio calibration at game start.

7. Conclusion

7.1 Main Research Focus

This thesis investigated the extent to which LLM-backed NPCs drive players' perceived narrative agency and overall experience in an audio-only narrative game. Findings indicate that such NPCs meaningfully support both, primarily by enabling natural, open-ended dialogue. Participants reported a strong sense of influence over the story, reflected in high Narrative Engagement, Enjoyment, and Creative Freedom scores on the GUESS scale. Interview data reinforced this, describing the experience as *"mysterious"*, *"natural"*, and *"empowering"*.

However, this enhancement was tempered by limitations in speech recognition, memory load, and narrative coherence. ASR errors, especially among non-native speakers, disrupted the flow while difficulty recalling prior clues impaired the strategy. Occasional hallucinations from the LLM also led to confusion or mistrust in the story logic.

Overall, LLM-backed NPCs gave a positive amount of agency and immersion, but their effectiveness depends on reliable speech systems, narrative consistency, and player support mechanisms. With thoughtful design, they show strong potential as a core mechanic in audio-first interactive storytelling.

7.2 Sub-Question Insights

7.2.1 Usability & Control

Most participants found the prototype usable and enjoyable, reflected in a strong average SUS score (83.5) and positive GUESS ratings for Usability/Playability and Enjoyment. Voice-controlled mechanics were generally intuitive, especially after the tutorial and with repeated use. However, non-native speakers faced recurring ASR misrecognition and often relied on visual UI cues, disrupting the intended audio-only experience. These difficulties led to repair behaviours such as repeated commands or task abandonment. Triangulated findings, behavioural logs, questionnaire scores, and interview data revealed consistent challenges around speech input clarity, navigation, and feedback. Overall, the study highlights the need for improved ASR support, redundant input strategies, and adaptive guidance to make voice-only interaction more reliable and inclusive.

7.2.2 Agency & Narrative Experience

LLM-backed NPCs enhanced players' sense of narrative control by enabling open-ended interaction. Thirteen participants reported feeling they could shape the story, supported by high GUESS scores for Creative Freedom and Narrative Engagement. However, this agency was sometimes undermined by NPCs revealing key information unprompted or by unclear branching structures. Some players expressed uncertainty about the consequences of their choices or desired a more visible narrative impact. While most participants experienced a strong sense of influence, sustaining agency requires clearer pacing control, consistent dialogue logic, and more transparent branching design.

7.2.2 NPC Quality

NPCs were widely perceived as believable, with seventeen participants praising their voice acting, tone, and contextual responsiveness. These traits contributed to immersion and trust in the game world. However, a few participants noted overly scripted or repetitive responses, and 4 encountered hallucinations such as invented characters or contradictions, which disrupted credibility. These instances negatively impacted narrative coherence. Overall, LLM-backed NPCs were successful in creating compelling, human-like interactions, though maintaining consistency and managing hallucinations remains essential for narrative integrity.

7.2.3 Audio & Modality

Spatial and environmental audio proved central to immersion, helping players distinguish zones and build atmosphere. Participants responded positively to ambient cues and character voices, which reinforced narrative tone and world-building. High Audio Aesthetics scores support this. Yet, several players struggled with orientation due to vague directional cues or inconsistent feedback. While audio-only design holds strong immersive potential, it must be underpinned by reliable spatial markers and consistent auditory feedback to prevent disorientation and reduce cognitive strain.

7.3 Limitations

These findings should be considered in light of several limitations. All 20 participants were sighted, which limits generalisability to the intended user base of blind or low-vision (BLV) players. While some avoided the UI to simulate non-visual play, others occasionally relied on visual cues, particularly during

ASR failures or navigation difficulties, introducing potential bias in assessing immersion and usability under fully audio-only conditions.

The prototype was an early-stage vertical slice with technical constraints, including looping dialogue, minor bugs, limited NPC diversity, and a compressed narrative. These factors may have restricted perceived agency and depth. Each participant engaged in only one 30-45 minute session, limiting insights into learning curves, long-term engagement, and player adaptation over time.

Finally, although the sample included non-native English speakers, most participants were digitally literate university students in their 20s and 30s. The results may not extend to older users, individuals with cognitive impairments, or those less familiar with technology.

7.4 Future Research

To advance the design of LLM-backed, audio-only narrative games, future research should prioritise testing with blind and visually impaired (BLV) participants to better assess accessibility, immersion, and the effectiveness of spatial audio and memory aids under truly non-visual conditions.

Improving automatic speech recognition (ASR) for accented and non-standard speech remains essential, alongside implementing stronger grounding and narrative state tracking to reduce hallucinations and maintain coherence in NPC dialogue.

As single-session testing limits insight into adaptation over time, longitudinal studies could reveal how players develop conversational strategies and how perceptions of agency and usability evolve with extended exposure.

Broader participant demographics, including users with motor impairments, neurodivergent individuals, and those unfamiliar with digital games, are also needed to understand varied cognitive and physical interactions with voice-based systems. These insights can guide the development of adaptive difficulty, multimodal input options, and more inclusive onboarding.

By addressing these areas, future studies can more fully evaluate the potential of LLM-backed NPCs for accessible, immersive storytelling beyond visual-first design norms.

7.5 Discussion

This study shows that LLM-backed NPCs can significantly enhance narrative agency and player experience in audio-only games, provided system reliability and interaction design are carefully managed. While the prototype was designed with accessibility in mind, testing with sighted users revealed strong mainstream appeal. Participants valued the freedom of open-ended dialogue and the immersion created by spatial audio and naturalistic NPCs, suggesting wider relevance beyond assistive contexts.

However, limitations such as ASR errors, memory strain, and hallucinations underscored the fragility of voice-based systems. Players desired both exploratory freedom and structural support, clear feedback, conversational grounding, and narrative coherence. These findings highlight the importance of robust scaffolding and responsive design in LLM-driven interactions.

Importantly, this research aligns with prior findings by (Cox & Ooi, 2024), who identified similar challenges in LLM-backed NPC design, particularly around hallucinated content and narrative drift. Their work offered initial design guidelines to address these issues, mainly in visually supported gaming environments. The current study complements their research by applying and extending these design principles within an *audio-only* context, where the absence of visual cues magnifies the consequences of narrative inconsistency and misunderstanding. This thesis contributes a tailored set of design recommendations specifically suited to audio-based gameplay, addressing not just hallucination mitigation but also pacing, memory aids, and spatial audio design elements that are critical to sustaining immersion when voice and sound are the sole modalities.

Beyond its accessibility benefits, this approach offers a shift in narrative delivery from visual instruction to vocal interaction. With further refinement, LLM-backed NPCs could support slower, dialogue-centric gameplay that appeals to both inclusive and mainstream audiences seeking deeper emotional engagement.

8. Bibliography

Allison, F., Carter, M., Gibbs, M., & Smith, W. (2018, 10 23). Design Patterns for Voice Interaction in Games. *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play.* 10.1145/3242671.3242712

Araújo, M. C. C., Façanha, A. R., Darin, T. G. R., Sánchez, J., Andrade, R. M. C., & Viana, W. (2017). Mobile Audio Games Accessibility Evaluation for Users Who Are Blind. *Universal Access in Human–Computer Interaction. Designing Novel Interactions*, 242-259. 10.1007/978-3-319-58703-5_18

Bangor, A., Kortum, P., & Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies*, *4*, 114-123.

Braun, V., & Clarke, V. (2006, 1). Using Thematic Analysis in Psychology. *Qualitative Research In Psychology*. https://www.researchgate.net/publication/235356393_Using_thematic_analysis_in_psychology

Brooke, J. (1995, 11). SUS: a Quick and Dirty Usability Scale. *ResearchGate*. https://www.researchgate.net/publication/228593520_SUS_A_quick_and_dirty_usability_scale

Cox, S. R., & Ooi, W. T. (2024, 1 1). Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback. *Lecture notes in computer science*, 167-184. 10.1007/978-3-031-54975-5_10

Gonçalves, D., Piçarra, M., Pais, P., Guerreiro, J., & Rodrigues, A. (2023, 4 19). "My Zelda Cane": Strategies Used by Blind Players to Play Visual-Centric Digital Games. *arXiv (Cornell University)*. 10.1145/3544548.3580702

Guillen, G., Jylhä, H., & Hassan, L. (2021, 6). The Role Sound Plays in Games:. *Academic Mindtrek 2021*. 10.1145/3464327.3464365

Korkiakoski, M., Sheikhi, S., Tapio, K., Nyman, J., Saariniemi, J., & Kostakos, P. (2025, 1 1). An Empirical Evaluation of Ai-Powered Non-Player Characters' Perceived Realism and Performance in Virtual Reality Environments. 10.2139/ssrn.5148461

Mason, Z., Green, D., Lindley, J., & Coulton, P. (2023, 6 20). Improving Digital Accessibility Through Audio-Game Co-Design. *Conference Proceedings of DiGRA 2023 Conference: Limits and Margins of Games Settings*. 10.26503/dl.v2023i1.1922

Melhart, D., Togelius, J., Mikkelsen, B., Holmgård, C., & Yannakakis, G. N. (2023). The Ethics of AI in Games. *IEEE Transactions on Affective Computing*, *15*, 1-14. 10.1109/TAFFC.2023.3276425

Moser, C., & Fang, X. (2015, 12 2). Narrative Structure and Player Experience in Role-Playing Games. *International Journal of Human-Computer Interaction*, 146-156. 10.1080/10447318.2014.986639

Nunes, C., & Darin, T. (n.d.). Echoes of Player Experience: A Literature Review on Audio Assessment and Player Experience in Games. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CHI PLAY), 1-27. 10.1145/3677069

Phan, M., Keebler, J. R., & Chaparro, B. S. (2016, 9 19). The Development and Validation of the Game User Experience Satisfaction Scale (GUESS). *SAGE Publications*. https://www.researchgate.net/publication/308343588_The_Development_and_Validation_of_the_Game_User_Experience_Satisfaction_Scale_GUESS

Porter, J. R., & Kientz, J. A. (2021, 3 22). An empirical study of issues and barriers to mainstream video game accessibility. *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*. 10.1145/2513383.2513444

Prazaru, A., Balan, O., Moldoveanu, A., Moldoveanu, F., Morar, A., & Ivascu, S. (2020, 7 1). OVERVIEW ON VISUALLY IMPAIRED GAMERS AND GAME ACCESSIBILITY. *EDULEARN proceedings*. 10.21125/edulearn.2020.1439

Ran, Z., Li, X., Xiao, Q., Fan, X., Li, F. M., Wang, Y., & Lu, Z. (2025, 4 24). How Users Who are Blind or Low Vision Play Mobile Games: Perceptions, Challenges, and Strategies. 1-18. 10.1145/3706598.3714205

Röber, N., & Masuch, M. (2005, 1 1). Playing Audio-only Games: A compendium of interacting with virtual, auditory Worlds. *Digital Games Research Conference 2005, Changing Views: Worlds in Play, June 16-20, 2005, Vancouver, British Columbia, Canada*. 10.26503/dl.v2005i1.129

Zargham, N., Friehs, M. A., Tonini, L., Alexandrovsky, D., Ruthven, E. G., Nacke, L. E., & Malaka, R. (2024, 4 19). Let's Talk Games: An Expert Exploration of Speech Interaction with NPCs. *International journal of human-computer interaction*, 1-21. 10.1080/10447318.2024.2338666

9. Appendix

Appendix A: Prototype Gameplay Video

To provide additional context, a recorded "happy path" gameplay session of the *Helios* prototype is available at the following link:

Gameplay Video Access

This recording presents a representative walkthrough of the prototype, illustrating the core interaction model, narrative structure, and audio-only mechanics explored in this study. While it does not encompass the full range of narrative variability, it reflects one coherent and uninterrupted route through the game, as designed.