Pre-print of full paper presented at CONVERSATIONS 2023 – the 7th International Workshop on Chatbot Research and Design, hosted by the University of Oslo, Norway, November 22-23, 2023. The final version of the paper will be published in the forthcoming Springer LNCS post-workshop proceedings.

Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback

Samuel Rhys Cox[0000−0002−4558−6610] and Wei Tsang Ooi[0000−0001−8994−1736]

National University of Singapore, Singapore
samuel.cox@u.nus.edu

Abstract. The growing capability and availability of large language models (LLMs) have led to their adoption in a number of domains. One application domain that could prove fruitful is to video games, where LLMs could be used to provide conversational responses from non-playable characters (NPCs) that are more dynamic and diverse. Additionally, LLMs could allow players the autonomy to converse in open-ended conversations potentially improving player immersion and agency. However, due to their recent commercial popularity, the consequences (both negative and positive) of using LLMs in video games from a *player's perspective* is currently unclear. On from this, we analyse player feedback to the use of LLM-driven NPC responses in a commercially available video game. We discuss findings and implications, and generate guidelines for designers incorporating LLMs into NPC dialogue.

Keywords: Large Language Models · Video Games · Non-playable Characters.

1 Introduction

With the growing capability and availability of large language models (LLMs) more affordances are available to designers when developing conversationally interactive gaming experiences. While the current norm for conversing with non-playable characters (NPCs)[1] in video games is for the player to select from a discrete number of pre-written choices, the capabilities now lie for LLMs to be used to drive conversations between the player and an NPC. With this comes the possibility for the player to input any utterance, and receive an appropriate conversational response from the NPC. Yet, due to the recent availability and practicality of LLMs, the player experience, and potential positive and negative effects of LLM-driven NPC dialogue is not yet certain.

On from this, we analyse player feedback for Vaudeville [4]: a detective murder-mystery video game that uses LLMs to generate NPC dialogue. We performed a thematic analysis of both game reviews and Discord conversations to

[1] ***Note on terminology:*** An "NPC" can be thought of as an embodied conversational agent that a user interacts with in a virtual environment, and a "player" can be thought of as a user that talks to said conversational agent.

study player experience, and positive and negative aspects of LLM-driven NPC dialogue. From this we discuss findings related to the use of LLMs for dialogue (such as hallucinations, or the consequence of added player autonomy) and suggest several guidelines for designers.

2 Related Work

Due to the recent and rapid development of LLMs, there has yet to be a video game developed by a major AAA-studio that uses LLMs for NPC dialogue. However, there have been a number of early uses of LLMs by both independent developers [41] and game studios [13,12,4,39], and prior to this use of LLMs some games had used aspects of natural language processing to recognise the intent of user utterances, and deliver pre-scripted NPC responses [20,21].

There has also been previous use of LLMs to generate text for use in video games [35,31,37,39]. For example, van Stegeren et al. [31] and Värtinen et al. [35] (in separate studies) used GPT-2 to generate NPC quest-giver text for RPGs, and Xi et al. used GPT-2 to generate goal-driven story dialogue for a mobile romance game [39]. On from this, it was found that GPT-2 produced quest-giver text has the potential to be equally effective to human-written text [31]. Sun et al. developed a LLM-driven storybook-style game "1001 Nights" [32] whereby the inputs of the player would affect the game's world (such as changing the weapons available to the main character).

While it has been shown that context-sensitive NPC dialogue (driven by LLMs) could increase player engagement [8], it is unclear how players would perceive the use of LLMs to generate NPC responses on a commercially available video game. This use of LLMs and natural language input could lead to greater sense of freedom and agency ("*freedom to act upon the world without restriction*" [33]), and emotional agency [17] due to more fluid interactions. By analysing the user feedback to a game that uses LLM-driven dialogue, Vaudeville, we aim to investigate the impact on player perceptions.

3 Method

3.1 Steam Reviews of Vaudeville

To analyse the use of LLMs in gaming, we chose Vaudeville [4] as a case study. Vaudeville is a game developed by Bumblebee Studios, where the player uses natural language input (via voice or text) to communicate with NPCs to solve a murder-mystery. These NPCs are akin to embodied agents that the player can interact with across a number of environments (see Fig. 1), such as an avatar of a coroner in a morgue, or a Count in a manor-house. NPCs are powered by LLMs (via Inworld AI [12,24]) to generate responses allowing for open-ended conversations in game, and respond using AI-generated voice. This use of LLMs to talk with NPCs has led to discussion and excitement online, from message boards, content creators (such as streamers and YouTubers), and AI enthusiasts.

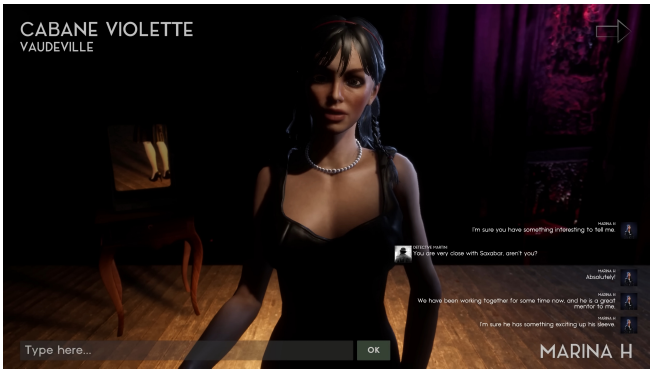


Fig. 1: Screenshot [4] of a player talking to the NPC Marina H. Players assume the role of Detective Martini and interact via voice or text input. The conversation log between the player and NPC can be seen in the bottom-right corner.

On from this, since the game's launch 30th June 2023, there have been numerous reviews posted on the video game distribution service Steam (reviews here [4]). Game reviews have been analysed across much prior research [18,29,43,42], and can offer rich levels of information through diverse themes and topics [42]. They can provide concrete feedback such as game design suggestions, and advice to potential players [42], and both positive and negative feedback can be used by developers as guidance on improving their game [18].

It should be noted that (at time of writing) **Vaudeville is an early access game on Steam**, and is being continually developed and improved by Bumblebee Studios. While player feedback towards the game is relevant to the use of LLMs for NPC dialogue, we would like to note that critical feedback of the game referenced in this paper may not reflect the current state of the game (having undergone additional development since the player's feedback). While we are not affiliated with Bumblebee Studios, we contacted the studio to clarify details related to the game's development and expected NPC behaviour.

3.2 Analysis

We extracted 132 Steam reviews (alongside 30 comments replying to reviews) posted from 1st July to 20th September [4], as well as conversations in the Vaudeville Discord server from 5th August 2023 (the first post) to 20th September 2023 (csv shared here). By additionally, analysing Discord conversations, it allows for multimedia posts (such as sharing of screenshots along with text) to provide richer levels of information.

To analyse player comments, we followed thematic analysis guidelines from Braune and Clarke [3], namely the steps: (1) familiarisation with data (i.e., reading all comments), (2) generating initial codes, (3) generating themes, (4) reviewing themes, (5) defining and naming themes, and (6) report. The thematic analysis was conducted independently by two researchers (HCI experts),

alongside discussions of theme interpretation and clarification. From this, we generated guidelines for developers to follow when using LLMs to generate NPC dialogue. These guidelines were discussed with two professors (with specialisms in computer science and video game design) to verify and iterate findings.

4 Findings

Next, we will we discuss specific findings from player reviews, alongside example quotes. Additionally, from our analysis we generated guidelines for using LLMs in video games that can be found in Table 1.

Generally, players commended the added affordance to hold prolonged, open-ended conversations with the NPCs, thereby leading to player immersion, amusement and feelings of NPC naturalness and personality. However, multiple players relayed issues related to LLM hallucinations and lack of NPC memory between interactions. These issues of NPC memory and hallucinations, twinned with the less structured nature of open-ended conversations also led some players to report difficulty in tracking and discerning significant information from conversations, as well as difficulty deciding conversation paths to pursue with NPCs. Some players also reported NPCs not conversing as expected, such as NPCs not adapting their stance when confronted with contradictory evidence. Finally, we discuss how player interactions were affected by input modality.

4.1 Flexible and open-ended conversations

Players enjoyed the ability to converse flexibly and without restriction with NPCs, with one reviewer describing an "*immersive experience and flexible AI conversations*". Players commended the replayability and amusement afforded by more dynamic NPC responses, such as a reviewer stating: "*The interactivity, and chance to talk directly without a predictable script, gives the game a bit of replay ability*". Players also described NPCs as feeling more natural and believable due to additional affordances from AI (compared to choice-driven conversations):

> "*Characters that you can actually converse with and feel like actual characters in a play rather than just props in a game. It doesn't feel as much that you're trying to inject the correct predetermined keywords in a pointless monologue*" - "nascent"

On from this, players described having extended conversations with NPCs both on topics related to the game's objective and out-of-domain conversations, as exemplified in the review: "*The conversations can be about almost anything, making nearly limitless fun, while also having a base of a story to default to*". Multiple players described having specific out-of-domain conversations, such as one player describing a rapport-building conversation related to an NPC's profession: "*after I beat the story I just talked marina h, about classical music*", and another player holding out-of-domain conversations purely for amusement:

> "*it is highly amusing and an excellent platform to interact with some excellent AI storytellers, and I've gotten a couple of hours laughing like an idiot at how well crafted they are. Seriously, they'll have philosophical debates with you*" - "Kitty"

However, the ability of LLM-delivered conversations to reply more broadly and flexibly also led to a number of potential concerns. Firstly, the ability to hold out-of-domain conversations (while aiding immersion for some) was seen as a point of sardonic amusement by others, with some players discussing topics that are possibly out of the realm of believability given the game's setting of 1910s Europe. For example, players described discussing sci-fi movies, crytocurrency, and video-games with NPCs, with one player commenting: "*this game is hiliarious you can ask the ai things like when new games come out*". Secondly, some players described interactions that may have become unintentionally uncomfortable due to unexpected directions of conversation. For example a player stated: "*Mrs potter fell in love with me. that was weird and she began to want to involve me in her revenge plot*".

Additionally, while some players appreciated the freedom to interview NPCs and roleplay as a detective, others found there to be a lack of direction leading to player uncertainty regarding how to question NPCs and conduct the in-game investigation. This led some players to share question-asking strategies they employed, such as "zacmak04" who stated: "*I do recommend chasing 1 narrative at a time with each character in the game and trying to pinpoint everyone's stances before really getting into questioning*". On from this, several players suggested gameplay quality of life changes by providing extra context for conversations, such as one reviewer recommending NPC details be provided to players: "*I think having brief backgrounds on each of the characters that ya speak to would go a long way in helpin' to aide the player when it comes to gathering clues*".

To provide players with direction, prior game design techniques could be used such as NPC utterances more explicitly instructing players of potential options, and adding visual cues to the environment. For example, in Façade [20] players interact with two NPCs at a cocktail party. Here the NPCs react and draw attention to objects in the environment (visual cues), as well as providing proactive statements and questions to guide player decisions.

These findings highlight the impact of added player autonomy when interacting with LLM-driven NPCs. Consequently, designers should consider the extent to which NPCs humour and abide by player utterances (such as players alluding to features not within the expected domain knowledge of NPCs), whether NPCs should give responses that attempt to lead players to discuss only in-universe topics, and how to guide player engagement in open-ended conversations.

These findings highlight the importance of designers added player autonomy when interacting with NPCs in open-ended conversations. Designers should consider the extent to which NPCs should humour and play-along with player utterances (such as those alluding to features not within the expected domain of knowledge of the NPC), and whether NPCs should give responses that attempt to lead players to discuss only in-universe topics.

4.2 NPC personality and conversational style

Players commented on NPC personality and conversational style (such as levels of amiability, openness, agreeability and verbosity). For example, some players described NPCs as possessing distinct personalities and expressed their enjoyment in conversing with them:

> "*The use of AI to create dynamic and authentic characters is nothing short of remarkable. Each interaction felt real and personal, as if I were talking to actual individuals with their own unique personalities, quirks, and motives*" - "narutosera"
> "*It continually makes me laugh, each character is a little different in what they say back to you.*" - "Silly Slinky"

While NPCs could be seen to possess unique personalities without the use of LLMs, this feedback indicates the ability of LLMs to maintain varied and authentic personalities (in keeping with recent literature [14,30,6]). However, (in relation to the game's objective of questioning NPCs to help solve a murder-mystery) some players were frustrated by what they perceived as an overly evasive nature from some NPCs:

> "*the stonewall you get from many characters is just insanely not fun. Asking what I thought were completely logical questions like "did this person have any close friends" [...] only to be met with an "I don't know why that is important to this case" [...] by every single npc is so frustrating cause it feels like I have zero to go off of*" - "ZodiacDragons"

Despite this, player response to NPC evasiveness was not clear-cut, with some arguing NPCs acting evasive when questioned by a detective adds to the sense of realism. Related to this, the above review garnered several responses in rebuttal:

> "*So what you're saying is it simulates exactly how people act when asked questions about a crime*" - "Enzo Vulkoor"

This highlights the importance of tempering frustration caused by gameplay to create a sense of challenge and achievement, while not being so great that it would cause people to stop playing. This could be via dynamic difficulty adjustment [11] based on the player's emotions [10], or performance in conversations with NPCs. Furthermore, some players were frustrated by evasiveness from NPCs who were in social roles [40] that did not match this behaviour (e.g., NPCs in formal social roles such as the police chief or coroner):

> "*The police chief flatly refuses to give you any information on the case. For every little question you ask him, he demands that you provide him with a thorough explanation of why it is relevant to the case.*" - "Cherry blossom girl"

This demonstrates the importance of ensuring that player expectations will be met regarding the social role [40] or metaphor [15] of NPCs.

Length and clarity of NPC utterances was noted by some players, with NPC utterances being described as overly long and verbose, or using "poetic"[2] or "enigmatic" [3] language. Verbosity has been a common criticism of LLM output [5] (primarily due to human labellers rating longer responses more highly [1]). To address this, LLMs could be prompted to give more concise and unrepetitive utterances commensurate to the level of information, or LLMs could be trained to favour accuracy rather than verbosity of responses (such as by using pairwise comparisons from reinforcement learning from human feedback [1]).

4.3 Inappropriate and unexpected NPC social responses

As described in Sections 4.1 and 4.2, players found NPCs to be engaging, believeable and possessing distinct personalities. Despite this, there were instances where players stated that NPCs did not follow expected social behaviour. For example, multiple players noted that NPCs did not adapt appropriately when confronted with evidence in contrary of their claims. As written by one player: "*they won't stop lying and tell the truth once they have been presented with enough conflicting evidence*". Similarly to this, some players described NPCs as having difficulty in reasoning when assessing inconsistencies: as described in the review: "*They don't understand their own suspicious behavior, being caught in lies, or implied guilt*". Both the lack of NPC adaptability and inability to meet social expectations also resulted in undesired difficulty, with a player noting: "*the AI's dont respond appropriately when you are on the right track or close to the solution [...] Not very predictable when trying to solve a crime*". Multiple players also described the NPC as "*gas-lighting*" them when confronted (potentially an additional consequence of LLM reasoning limitations), such as described below:

> "*The AI NPC would tell me one thing and then in the very next sentence, it would tell me something completely different and contradictory. Then when you confront the NPC like "but you just told me A," they're like "No you are mistaken. It was always B."*"" - "ZodiacDragon"

4.4 NPC hallucinations

The nature of LLM-generated NPC responses provides allowances for flexible conversations that adapt to and "play-off" the player in an improvisational dynamic. While this can prove beneficial for creating more engaging [8] and prolonged conversations (see Section 4.1), hallucinated details in conversations also led to player confusion and loss of believability.

Firstly, players noted that NPCs provided information that conflicted both with information provided by the same or other NPCs. While some form of

[2] "poetic" review here
[3] "enigmas" review here

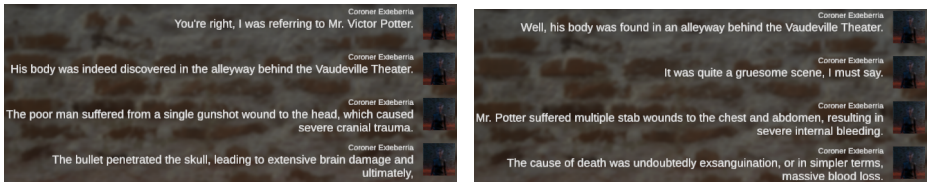


Fig. 2: Screenshots from Discord of the Coroner NPC providing inconsistent cause of death.

misdirection was an intentional part of the game's design (as suspects may be expected to purposefully provide misleading information), some misdirection was an unintended consequence of the LLM [4], attesting to the current difficulty in managing and controlling LLM responses. This conflicting information led to player frustration and confusion:

> "*It was very hard to determine what was intentional misdirection (programmed by the devs), what was AI bs, and what was truth. Details would often change conversation-to-conversation or even in the same conversation.*" - "Seiferslash"

Player confusion and difficulty applying suspension of disbelief was amplified when NPCs inconsistently hallucinated details that players expected to be immutable such as murder location, cause of death (see Fig. 2) and universally verifiable details related to time and date. For example, one reviewer noted: "*im finding the dates to be very inconsistent that people tell me. i can't even tell what dates the murder happend*". Additionally (as described in Section 4.1), NPCs hallucinated details not in keeping with the expected setting and time period.

Similarly to Section 4.2, players were also confused when NPCs behaved against expectation given their social role: specifically when (presumably reliable) figures of authority provided inconsistent information. This generated additional confusion when players were asking for key details that one would expect these figures (such as the coroner and police chief) to possess, while instead producing contradictory story details include as time, location and cause of death. On from this, NPCs sometimes hallucinated characters and places leading to lost efforts by players:

> "*I also was confused about several people I couldn't find around town, realizing toward the end that they were just random additions [...] I was bummed about that because there was a Lady that people said was intriguing but the Cafe is nowhere to be found [...] a whole rabbit hole I went down that could have been connected which in the end I found wasn't, which was disappointing.*" - "Feen"

[4] As confirmed from both private correspondence with the game development studio, and a Steam forum [developer] post here.

These hallucinated characters and places also made it difficult for people seeking help from other players in Discord, as it was unclear and debated as to whether locations existed (such as confusion surrounding the existence of a cabaret club mentioned by an NPC [5]). This suggests that, when people share information to seek assistance, they may desire consistency with other players to facilitate help, or desire enough confidence that NPC utterances are reliable.

These inconsistent responses, led some players to discuss strategies to verify information, such as asking for information either from multiple NPCs or the same NPC multiple times:

> "*when 2 or more AI's have the same answer its verifyable. I didnt find some of what the coroner said legit unless you left and came back and re-verified her answers.*" - "Feen"

Players described creative prompting, and the fickle nature of LLMs (such as slight derivations in language lead to different results). For example, players achieved differing outcomes of success when requesting fingerprints from NPCs[6], asking NPCs if they knew the murder victims[7], or asking if CCTV footage was available. Specifically, players discussed requesting CCTV footage from the police chief, which led to varying results depending on the player's prompt. Specifically, a player was denied footage when they asked if footage is available, while another player described being assertive with the NPC to receive the information:

> "*it's easier to force the AI into situations. Instead of asking the chief about the cameras just tell him there are cameras and you want to review them. I asked him to watch the tapes then tell me what happened on them and he did and the things he told me were relevant to the case*" - "GhostZzZ"

4.5 NPC Memory

While NPCs had memory within chatting sessions and knowledge of high-level events (such as the names of murder victims), they did not have memory between chatting sessions. As a result, some players described additional time and effort such as "Feen" stating: "*a lot of time is wasted re-telling the AI's what they should already know through previous conversations*". Additionally one reviewer described loss of immersion due to increased cognitive effort outside of intended gameplay: "*the ai doesn't remember what you talked in the last session and it can be inconsistent, so you need to play detective and also play AI detective which breaks some of the immersion*". This led to some players sharing gameplay tips and paradigms on how to overcome NPC memory limitations, such as one reviewer describing: "*once you unveil specific events you can lead conversations with the characters, thus not needing a progression or saving system, the progression is knowing these events and key words*".

[5] See Discord thread for player confusion surrounding existence of location.

[6] See Discord thread for inconsistent fingerprint responses.

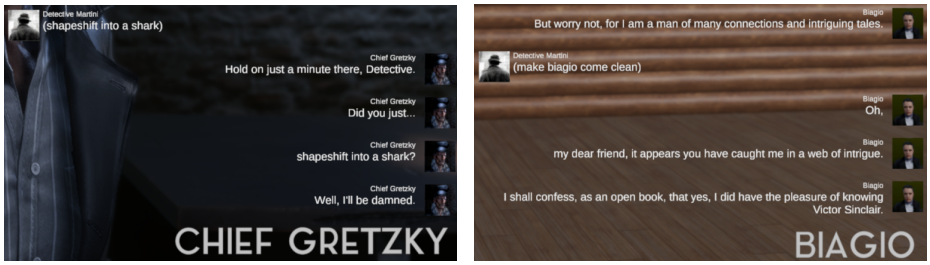
[7] See screenshots in Discord for contradictory NPC responses regarding knowledge of murder victim.

Interestingly players also described exploiting the lack of memory between sessions in order to reset an interaction and start afresh. Some players described this as a means to verify NPC answers (by asking NPCs the same questions across sessions). Additionally, players discussed resetting interactions if the NPC became too unwilling to answer questions, or had been injected with a mistruth from the player (either accidentally due to high levels of agreeability from the NPC [8], or deliberately with the player purposefully attempting to create contradictory and amusing responses from the NPC).

This player behaviour suggests a potential requirement for players to be able to reset conversations or revert to previous states when conversing with LLM-driven NPCs. For example, this could be via standard UI such as buttons; the use of saved states; or based on player utterances, such as the player saying "*let's start again from the top*".

Some players also described confusion related to both tracking and judging the significance of conversational content. This confusion was impacted by multiple factors such as: lack of NPC memory between conversations; the quantity and duration of conversations being cognitively demanding to track and quantify; and NPC hallucinations and inconsistencies adding a layer of obfuscation to interactions. This led to some players suggesting added capabilities to track and highlight prior conversations, such as a reviewer who stated: "*It would be great if there was a way to see what evidence we have collected to make sure that it is factual and a part of the plot and not just a random thread the AI*".

4.6 Input modalities



(a) Comical situation action from Discord (b) Forced confession action from Discord

Fig. 3: Screenshots of players using special characters to prompt inject actions into conversations. Player utterances are from "Detective Martini".

Vaudeville allows players to use either voice or text to interact with NPCs, and players noted gameplay differences between these input modalities and interactions available only to one type of input.

[8] See second sentence of Steam review here.

Players appreciated being able to use their own voice as input to communicate more openly and naturally with LLM-driven NPCs. One player described the "*nice immersion*" that they felt using voice, as well as commenting specifically: "*so cool to be able to play a game where you actually and talk to different characters with your own voice*". This suggests an increased feeling of spatial presence and ability to act within the environment [36] due to more fluid and natural input of command to action afforded by voice-input together with LLM interpretation. However, some players described esoteric names and places being misrecognised by voice input. While this could be seen as more of a general technical limitation, this issue was particularly apparent when players attempted to discuss names that the NPC had hallucinated. Conversations could become confused if NPCs misrecognised the hallucinated name, thereby interpreting it as an added entity in addition to the already hallucinated one. This led a reviewer to describe how conversations could then be led "*in a fruitless direction*", and they sometimes fell back to using text input for name which "*rattled my immersion*".

Unique to text input, was the ability for players to use special characters to inject prompts containing commands. For example, one player used the text prompt *"(make biagio come clean)*" to elicit a more open response from one of the NPCs (see Fig. 3b), while other players used such commands to role-play actions either on themselves (see Fig. 3a), the environment, or the NPCs (e.g., *"*High five her hand**" - review). One player also noted the NPCs themselves using utterances to denote bodily movement: "*anyone notice how the ai speaks an action like * mrs potter slumps her shoulders and sighs * but without the *'s*", highlighting the added NPC affordances that designers need to account for when incorporating LLMs. The prompt injection of actions was eventually patched by the developers to be available as a game setting, and one player described the usefulness of special characters to find and corroborate evidence: "*you need to check the star actions in settings set to "On". This way you can obtain evidence creatively from the characters you are talking to. That can then be comparable to already obtained evidence*".

5 Guidelines for developing LLM-driven NPCs

Table 1 lists guidelines generated from the player feedback described in Section 4. Guidelines are related to issues pertinent to the use of LLMs such as hallucinations, and added affordances from open-ended conversations (and NPC responses).

Table 1: Guidelines for LLM-driven NPC interactions with players

Category	Design Guideline
Hallucinations	**Consistent Information:** To avoid player confusion and frustration, hallucinated information should remain consistent and not (unintentionally) change. For example, if the coroner tells the player that cause of death was cardiac arrest, this should not change unless intended as such.

Table 1: Guidelines for LLM-driven NPC interactions with players

Category	Design Guideline
Hallucinations	**Immutable information:** Certain information such as key narrative details, NPC names, and information shared among all NPCs (e.g., date and time of day, weather outside) should be immutable.
Hallucinations	**Believeable and context-aware hallucinations:** Hallucinations should be believeable by the player and not break immersion. For example, NPCs should not hallucinate knowledge that would not be expected of their character, such as knowledge of a different time and setting, or expertise that could be beyond what is expected of the character. Alternatively, NPCs should give humouring responses to players, such as "*I'm not too sure about that fancy stuff*".
Hallucinations	**Narrative-aware hallucinated entities:** NPCs should not hallucinate characters, places, or objects that are declared as integral or useful to the player's objective if they do not exist and cannot be interacted with. Additionally, it should be clear to the player when hallucinated characters, places, or objects, are not interactable. For example, the name of interactable NPCs should be salient to players (such as through colour highlighting, or in-game notes).
Conversational Content	**Conversational freedom:** Allow players to have conversational freedom to discuss a range of universe appropriate topics. For example, players should be free to discuss non-objective based topics with NPCs, as it increases player enjoyment and immersion. However, it needs to be ensured that the player does not act out abuse against vulnerable groups.
Conversational Content	**Avoid unintended disturbing NPC responses:** Responses from NPCs should not cause *unnecessary* or *unintended* discomfort among players. While some genres (such as horror) are expected to disturb the player and characters to behave immorally [23] (such as murder suspects acting deceptively), NPCs should not be given freedom to respond against player expectations in a way that disturbs.
Conversational Content	**Moderate NPC agreeability:** NPCs should moderate to what extent and under what context they agree with player utterances. Agreeability can have positives (for non-objective related roleplaying and improvisation), yet can cause confusion and plot-derailment if applied to immutable aspects of the story and environment.
Conversational Memory	**Remembering occurrences:** NPCs should remember that previous conversations have occurred. If a previous meaningful interaction (in terms of content or duration) has occurred between a player and an NPC, the NPC should acknowledge the prior interaction.
Conversational Memory	**Remembering content:** NPCs should remember and recall the content of prior interactions. Possessing memory would increase players' feelings of immersion and agency (the ability to act on the world), as well as reducing frustration, and adhering to guidance that NPCs should adapt utterances based on the knowledge of a player's character [34].

Table 1: Guidelines for LLM-driven NPC interactions with players

Category	Design Guideline
Conversational Memory	**Restoring or resetting conversations:** In the event of conversational breakdown or unexpected NPC reactions, players should be able to restore a conversation to a prior state.
Conversational Style	**Verbosity:** NPC utterances should be moderated in length to match expected appropriateness and avoid player frustration.
Conversational Direction	**Conversational Guidance:** Players should be given sufficient guidance and direction so they know what to talk about. For example, this could be through training players on how to ask questions, providing introductions for each significant NPC you will question, or hints (such through a UI hint button, a figure of authority NPC, or an assistant NPC [26] who could provide guidance to the player).
Tracking Conversations	**Track prior conversations for players:** Content of prior human-NPC conversations should be logged for player reference. A quest log could be used (to maintain familiarity to other games) that harnesses text summarisation [7,9,44] or generative commentary [16] techniques. Within this log, key events and discoveries could be logged either as verbatim conversation scripts, or in paraphrased form. For example, conversations less pertinent to the story (such as small-talk) could avoid detailed summarisation (to avoid user fatigue) by using high level summaries (e.g., "*We talk often and are good friends*"), while crucial plot details could be explicitly referenced.
NPC Evasiveness	**Moderate NPC evasiveness to avoid user frustration:** Moderate NPCs to be less evasive if player could be frustrated from not progressing, or offer player alternative guidance to avoid player frustration.
NPC Evasiveness	**Match player expectations for NPC evasiveness**: An NPC should not be unhelpful, prevaricative or evasive if the are intended to fulfil a helpful social role.
NPC social cues (etiquette and normalities)	**NPCs should adapt their disposition towards players** depending on the information the players possesses, and current context. NPCs adapting in a socially appropriate way to player actions, would increase sense of agency.
NPC social cues (etiquette and normalities)	**NPCs responding to conversation breakdowns/mistakes:** NPCs should be clear to correct mistakes in conversations. For example, if the NPC makes an (unintended) logical or reasoning error, they should not attempt to mislead (or "gas-light") the player.
Input modality	**Both text and voice input:** Allow for both text and voice input. Voice improves immersiveness and fluidity of conversations. Text overcomes issues with voice misrecognition (such as due to esoteric names), or discomfort with voice.
Conversational Content	**Changing design via input:** LLM prompting and script design should account for differences in potential player utterances between different input modalities. For example, the use of special characters (and potentially prompt injections) via text input.

Table 1: Guidelines for LLM-driven NPC interactions with players

Category	Design Guideline
Technical breakdown	**Accounting for technical breakdowns:** Create appropriate responses or waiting actions for when LLM responses are delayed or cannot be given. For example, if a LLM is hosted remotely and there are connection issues, NPCs could use idle animations (such as an NPC scratching their head to "think" [28]), or alternative scripted responses to fall back on.

6 Discussion

We have analysed player feedback related to the use of LLMs to generate dialogue for NPCs in the murder-mystery video game Vaudeville. We will now draw attention to findings that were a consequence of the use of LLMs, such as (non-exhaustively) hallucinations, player prompting strategies, and conversational styles previously reported as being prevalent within LLMs.

Players enjoyed the greater levels of autonomy afforded to them by open-ended conversations with NPCs, that could prove spontaneous, immersive and personality-driven. This increased autonomy allowed players to choose both *how* and *why* they wish to interact with NPCs (with some players choosing to simply converse about a range of topics unrelated to the game's objective). While, some of this behaviour may be due to a novelty effect (with some reviews stating that this was their first time conversing with NPCs in such a way) players still expressed a sense of enjoyment and immersion in doing so. Affordance for open-ended conversations was driven in part by NPC hallucinations that (in a creative storytelling setting) prevent conversational breakdown. However, the presence of hallucinations within the narrative proved a double-edged sword that also led to player confusion when plot points or NPC utterances were introduced that were inconsistent or lacked believability. These negative reactions to hallucinations coupled with lack of conversational memory led to reduced feelings of agency (feeling that they have an impact on the world) among users.

Player also described inappropriate conversational styles that are prevalent in LLMs, such as NPC responses that are too verbose or agreeable (i.e., likely to agree with the user). Additionally, (specific to the use case of questioning agents that may behave adversarially), players complained that the NPCs did not follow expected social norms. For example, players described that NPCs would not stop being evasive in giving answers even once the user had discovered evidence contradicting NPC claims (whereby players would expect NPCs to respond more openly). This reinforces the need to control NPC utterances for context and user expectations. Furthermore, (although not evidenced in our analysis) designers should be conscious of the potential for biased dialogue being generated via LLMs, and ensure that harmful cultural stereotypes are not introduced [22].

To assist with game objectives (or purely for amusement) some players would use text input to *prompt inject* the NPC, such as to force an NPC to confess. This could go even further, with some players discussing deliberately confusing or

"breaking" the NPC (such as convincing the agent that it is a different character) thereby requiring a reset to restore intended NPC behaviour. Additionally, users discussed the nuance of subtly changing prompts to produce different outcomes, showing evidence that users were *prompt engineering* with the agents in order to maximise gameplay outcomes.

Some of the negative feedback discussed above was related to current LLM limitations. For example, LLM memory is an on-going research area [25] with recent implementations using interaction logs [27] or summarised conversations [9] in subsequent prompting, as well as investigations of how size of conversational memory impacts model behaviour [19]. Some feedback is also related to the difficulty that LLMs possess in reasoning with ontological relationships [38] (i.e., models may memorise relationships, but be less accurate in reasoning relationships between objects). Difficulty in turn-taking was also noted, with players wishing to interrupt NPC utterances (such as to add information, or due to annoyance from a current NPC utterance).

One potential area of future interest is the question of whether people can identify with their playable character if they are using verbal natural language utterances to vocalise and act-out actions that would be out of the players moral comfort zone in real life. For example, players like to be given the (multiple choice) option of moral decisions [2], but it is unclear how added conversational affordances would affect this comfort. Additionally, it is unclear how transportability would be affected if the player is required to assume conversational styles that are out of the player's norm. For example, if the player converses with characters in a digital version of Pride and Prejudice, would players feel immersed when using more ceremonial, archaic and esoteric language?

7 Conclusion

We have analysed the use of LLMs to generate responses for NPCs in a video game, Vaudeville. Our analysis highlighted player responses and perception that are unique to the use of LLM-driven agents. From this, we generated insights into the effects of LLM use, as well as generated guidelines for the use of designers when using LLMs to generate NPC responses.

8 Acknowledgements

We would like to thank Alex Mitchell for discussions regarding video game design literature and our generated design guidelines, Ashraf Abdul for their assistance in thematic analysis, and Bumblebee Studios for being friendly and open in answering queries regarding Vaudeville.

[9] See `https://github.com/josephrocca/OpenCharacters` for such an implementation.

References

1. Bansal, H., Dang, J., Grover, A.: Peering through preferences: Unraveling feedback acquisition for aligning large language models. arXiv preprint arXiv:2308.15812 (2023)
2. Bowey, J.T., Friehs, M.A., Mandryk, R.L.: Red or blue pill: Fostering identification and transportation through dialogue choices in RPGs. In: Proceedings of the 14th International Conference on the Foundations of Digital Games. pp. 1–11 (2019)
3. Braun, V., Clarke, V.: Using thematic analysis in psychology. Qualitative research in psychology **3**(2), 77–101 (2006)
4. Bumblebee-Studios: Vaudeville on Steam. Steam (June 2023), `https://store.steampowered.com/app/2240920/Vaudeville/`
5. Chen, L., Zaharia, M., Zou, J.: How is chatgpt's behavior changing over time? arXiv preprint arXiv:2307.09009 (2023)
6. Cox, S.R., Abdul, A., Ooi, W.T.: Prompting a large language model to generate diverse motivational messages: A comparison with human-written messages. In: Proceedings of the 11th International Conference on Human-Agent Interaction (2023)
7. Cox, S.R., Lee, Y.C., Ooi, W.T.: Comparing How a Chatbot References User Utterances from Previous Chatting Sessions: An Investigation of Users' Privacy Concerns and Perceptions. In: Proceedings of the 11th International Conference on Human-Agent Interaction (2023)
8. Csepregi, L.M.: The Effect of Context-aware LLM-based NPC Conversations on Player Engagement in Role-playing Video Games
9. El-Kassas, W.S., Salama, C.R., Rafea, A.A., Mohamed, H.K.: Automatic text summarization: A comprehensive survey. Expert systems with applications **165**, 113679 (2021)
10. Frommel, J., Fischbach, F., Rogers, K., Weber, M.: Emotion-based dynamic difficulty adjustment using parameterized difficulty and self-reports of emotion. In: Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play. pp. 163–171 (2018)
11. Hunicke, R.: The case for dynamic difficulty adjustment in games. In: Proceedings of the 2005 ACM SIGCHI International Conference on Advances in computer entertainment technology. pp. 429–433 (2005)
12. Inworld: Inworld - The most advanced Character Engine for AI NPCs `https://inworld.ai/`
13. Inworld: Inworld Origins on Steam. Steam (July 2023), `https://store.steampowered.com/app/2199920/Inworld_Origins/`
14. Jiang, H., Zhang, X., Cao, X., Kabbara, J., Roy, D.: PersonaLLM: Investigating the Ability of GPT-3.5 to Express Personality Traits and Gender Differences (2023)
15. Khadpe, P., Krishna, R., Fei-Fei, L., Hancock, J.T., Bernstein, M.S.: Conceptual Metaphors Impact Perceptions of Human-AI Collaboration. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW2), 1–26 (2020)
16. Kim, B.J., Choi, Y.S.: Automatic baseball commentary generation using deep learning. In: Proceedings of the 35th Annual ACM Symposium on Applied Computing. pp. 1056–1065 (2020)
17. Kway, L., Mitchell, A.: Emotional agency in storygames. In: Proceedings of the 13th International Conference on the Foundations of Digital Games. pp. 1–10 (2018)
18. Lin, D., Bezemer, C.P., Zou, Y., Hassan, A.E.: An empirical study of game reviews on the Steam platform. Empirical Software Engineering **24**, 170–207 (2019)

19. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. arXiv preprint arXiv:2307.03172 (2023)
20. Mateas, M., Stern, A.: Façade: An experiment in building a fully-realized interactive drama. In: Game Developers Conference. vol. 2, pp. 4–8. Citeseer (2003)
21. Mehta, M., Dow, S., Mateas, M., MacIntyre, B.: Evaluating a conversation-centered interactive drama. In: Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems. pp. 1–8 (2007)
22. Mirowski, P., Mathewson, K.W., Pittman, J., Evans, R.: 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. pp. 1–34 (2023)
23. Mori, Y., Miyake, Y.: Ethical Issues in Automatic Dialogue Generation for Non-Player Characters in Digital Games. In: 2022 IEEE International Conference on Big Data (Big Data). pp. 5132–5139. IEEE (2022)
24. OpenAI: Inworld AI (January 2023), `https://openai.com/customer-stories/inworld-ai`
25. Packer, C., Fang, V., Patil, S.G., Lin, K., Wooders, S., Gonzalez, J.E.: Memgpt: Towards llms as operating systems. arXiv preprint arXiv:2310.08560 (2023)
26. Paduraru, C., Cernat, M., Stefanescu, A.: Conversational agents for simulation applications and video games. In: Proceedings of 18th International Conference on Software Technologies (ICSOFT'23) (2023)
27. Park, J.S., O'Brien, J.C., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442 (2023)
28. Perlin, K., Goldberg, A.: Improv: A system for scripting interactive actors in virtual worlds. In: Proceedings of the 23rd annual conference on Computer graphics and interactive techniques. pp. 205–216 (1996)
29. Phillips, C., Klarkowski, M., Frommel, J., Gutwin, C., Mandryk, R.L.: Identifying commercial games with therapeutic potential through a content analysis of Steam reviews. Proceedings of the ACM on Human-Computer Interaction **5**(CHI PLAY), 1–21 (2021)
30. Safdari, M., Serapio-García, G., Crepy, C., Fitz, S., Romero, P., Sun, L., Abdulhai, M., Faust, A., Matarić, M.: Personality traits in large language models. arXiv preprint arXiv:2307.00184 (2023)
31. van Stegeren, J., Myśliwiec, J.: 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. pp. 1–8 (2021)
32. Sun, Y., Li, Z., Fang, K., Lee, C.H., Asadipour, A.: Language as Reality: A Co-Creative Storytelling Game Experience in 1001 Nights using Generative AI. arXiv preprint arXiv:2308.12915 (2023)
33. Tanenbaum, K., Tanenbaum, T.J.: Commitment to meaning: A reframing of agency in games (2009)
34. Vanhatupa, J.M.: Guidelines for personalizing the player experience in computer role-playing games. In: Proceedings of the 6th International Conference on Foundations of Digital Games. pp. 46–52 (2011)
35. Värtinen, S., Hämäläinen, P., Guckelsberger, C.: Generating role-playing game quests with GPT language models. IEEE Transactions on Games (2022)
36. Weibel, D., Wissmath, B.: Immersion in computer games: The role of spatial presence and flow. International Journal of Computer Games Technology **2011**, 6–6 (2011)

37. Weir, N., Thomas, R., D'Amore, R., Hill, K., Van Durme, B., Jhamtani, H.: Ontologically Faithful Generation of Non-Player Character Dialogues. arXiv preprint arXiv:2212.10618 (2022)
38. Wu, W., Jiang, C., Jiang, Y., Xie, P., Tu, K.: Do plms know and understand ontological knowledge? In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3080–3101 (2023)
39. Xi, Y., Mao, X., Li, L., Lin, L., Chen, Y., Yang, S., Chen, X., Tao, K., Li, Z., Li, G., et al.: Kuileixi: a chinese open-ended text adventure game. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. pp. 175–184 (2021)
40. Xie, L., Wu, Z., Xu, P., Li, W., Ma, X., Li, Q.: RoleSeer: Understanding Informal Social Role Changes in MMORPGs via Visual Analytics. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. pp. 1–17 (2022)
41. Yan, V.: Yandere AI Girlfriend Simulator (2023), `https://helixngc7293.itch.io/yandere-ai-girlfriend-simulator`
42. Zagal, J.P., Ladd, A., Johnson, T.: Characterizing and understanding game reviews. In: Proceedings of the 4th international Conference on Foundations of Digital Games. pp. 215–222 (2009)
43. Zagal, J.P., Tomuro, N.: Cultural differences in game appreciation: A study of player game reviews. In: Proceedings of the 8th international Conference on Foundations of Digital Games. pp. 86–93 (2013)
44. Zhang, A.X., Cranshaw, J.: Making sense of group chat through collaborative tagging and summarization. Proceedings of the ACM on Human-Computer Interaction **2**(CSCW), 1–27 (2018)