Exploring the (Placebo) Effect of an AI-Powered Dialogue System on Player Experience

Sarita Sridharan
LMU Munich
Munich, Germany
s.sridharan@campus.lmu.de

Steeven Villa
LMU Munich
Munich, Germany
villa@posthci.com

Alena Denisova
University of York
York, United Kingdom
alena.denisova@york.ac.uk

Johanna Pirker
Graz University of Technology
Graz, Austria
johanna.pirker@tugraz.at

***Abstract*—This paper examines the impact of AI-generated versus human-written dialogue in a murder-mystery game on player experience. We manipulated both the game dialogue type (AI-generated vs. human-written dialogue) and the game type description (AI-labelled vs human-written-labelled) to assess effects on overall player experience (PX), immersion, and perceived autonomy. Results show that player expectations, modulated by the game type description, had a greater impact on overall PX than the actual game type. Players who thought that they played a game with a human-written narrative reported more positive overall PX, higher immersion and perceived autonomy regardless of which version they actually played. However, interacting with the AI-generated dialogue version led to higher perceived autonomy, suggesting that AI-driven dialogue can enhance perceived freedom without necessarily improving the overall PX. These findings show that player expectations, in response to game descriptions, play a central role in the experiences of players. This work contributes to understanding the influence of narratives and the extent to which AI features increase player engagement.**

***Index Terms*—placebo effect, artificial intelligence, AI, player experience, PX, video games, role-playing games**

I. Introduction

Designing for Player Experience (PX) is central to shaping player engagement, satisfaction, and the overall success of a game [1], [2]. One emerging technology that can transform how PX is created is the use of large language models (LLMs). While LLMs still face limitations [3], several studies have explored their application in generating more dynamic and realistic non-player characters (NPCs) [4], [5]. Instead of relying on predefined dialogue trees, AI chatbots, such as ChatGPT, can generate responses almost in real time, considering the context of the conversation and the player's actions. Recent role-playing games (RPGs) like *Baldur's Gate 3* and *Kingdom Come Deliverance 2* have reached remarkable narrative scales, with over two million and 2.2 million words of dialogue, respectively [6], [7]. In this context, LLMs offer a promising opportunity to reduce scripting complexity while preserving, and potentially enhancing, narrative depth and player engagement.

Previous research found that merely making players believe that they are playing an AI-enhanced game was enough to increase their sense of immersion [8]. However, no previous work has compared players' perceptions of a game described as using AI versus their experience with an actual AI-based system in a story-driven video game, even if prior research has emphasized the importance of narrative and game mechanics in shaping meaningful player experiences [9]. We, therefore, aim to explore the impact of AI-generated dialogue on PX and investigate whether the mere perception of interacting with an AI-enhanced system can produce a similar effect. Two versions of the same game were developed, one AI-generated (AI) and one hand-written (HW), with the latter being used as a baseline system, as it largely reflects the standard practice as of today. By comparing the expectations of players, in the following called the description type of the game (DT), with the actual version they are playing, the game type (GT), we aim to provide insights into how players respond to the integration of AI-generated dialogue in games, real or imagined.

Based on prior research [8], [10], [11] we formulated three hypotheses:

- $\boldsymbol{H_1}$: **Player Experience** will be rated higher in the *AI*-description (DT AI) condition.
- $\boldsymbol{H_2}$: The game type *AI* (GT AI) will have a higher rated **Player Experience** than game type *HW* (GT HW), regardless of description.
- $\boldsymbol{H_3}$ The *AI*-description condition (DT AI) will have a higher rated **Player Experience** than the *HW*-description condition (DT HW), in both game types.

Additionally, we analyse *Immersion*, *Perceived Autonomy*, and *Emotional Challenge* – key factors for understanding the gaming experience in this narrative-heavy game. These variables were selected for their relevance to player engagement and their ability to capture distinct dimensions of the overall experience [12], [9].

The findings of this study offer a different perspective from some previous research [8], [11]. While player expectations, influenced by the DT, clearly play a role in shaping gaming experiences, our results suggest that it was not the expectation of AI-enhanced content that contributed most to the player experience; Participants who were told that they would be playing the non-AI version reported a significantly higher *Player Experience*, *Immersion* and *Perceived Autonomy*. Not only did they have higher expectations in the game, but they also found it more engaging. Interestingly, the game type *AI* did not seem to have any significant positive effects outside of *Perceived Autonomy*, indicating that AI-generated content alone does not improve the gameplay experience.

II. Related Work

This section explores the growing use of LLM-based chatbots in games and the influence of game descriptions on player experience, positioning our contribution within this context.

A. LLM Chatbots in Games

In 2023, Nvidia released Avatar Cloud Engine (ACE), tools for game development that enable customizable LLMs for game narrative, live speech conversations and associated animations [13]. Using generative AI, Inworld (https://inworld.ai), an AI startup, offers customizable AI NPCs that dynamically respond to player actions and can be easily integrated into games. At the Game Developers Conference in 2024, Ubisoft collaborated with Nvidia and Inworld to present their so-called NEO NPC [5], which uses generative AI to create realistic NPCs. The personalities and backstories of the NPCs are written by a human writer, which then builds the basis for realistic dialogue generated by an LLM.

However, the implementation of these AI-NPCs in actual games has been relatively limited so far [14], [15], [16], [17]. The cost-benefit of employing LLMs in games still seems too big of a deterrent; LLMs face severe limitations, including struggles with contextual understanding [18], lack of common sense[3], susceptibility to biases in training data [19] as well as emotional intelligence [20]. Chatbots like ChatGPT often generate responses that lack coherence or relevance when faced with ambiguous queries, and their knowledge is limited to their training data, leading to inaccuracies [3]. Specifically, LLM hallucinations, a phenomenon when an LLM generates semantically correct but factually incorrect output, pose a significant challenge in the usage of LLMs [21]. For example, reviews for the indie game *Vaudeville* (https://store.steampowered.com/app/2240920/Vaudeville/), which builds on an AI-driven narrative, revealed that while players appreciated the higher degree of freedom and thus increased immersion afforded by AI, the feedback mostly seemed to relate to problems with the technology. NPC's missing memory between interactions, LLM hallucinations as well as the missing structure in these open-ended conversations appeared to be the biggest issues for players [22]. Other experiments that measured player responses to AI-generated narratives in comparison with human-written ones concluded that while current LLM models were not far off, human-written dialogue was still considered superior concerning logical consistency and flow of the dialogue [23], [24]. Based on these findings, we decided on the mixed approach used by Ubisoft's NEO NPC [5], combining a human-written overarching narrative with AI-generated dialogue to create a realistic AI-NPC.

B. Influence of Descriptions in Player Experience

Player expectations can play a significant role in influencing player experience. Previous experiments show that these expectations can also be deliberately guided to achieve the desired result. For example, players felt significantly more immersed when they believed a game used adaptive AI, even though the feature did not exist [8]. This could possibly be explained by higher engagement elicited by the active search for adaptive elements or simply the firm belief in the adaptive nature of the game [11]. Based on this previous work, additional research revealed how varying levels of information about adaptive technology influence player experience while playing a game with or without this technology. Players felt consistently more immersed and experienced more enjoyment when adaptive technology was expected to be present in the game, regardless of whether the technology was actually present or not [11].

However, gameplay mechanics assumed or real, only make up a part of the gameplay experience, particularly in story-driven games. In these games, while both the narrative, as well as mechanics need to be appealing, mechanics alone are less likely to elicit meaningful experiences [9]. Based on related work we wanted to investigate to what extent the narrative of the game and the mechanic of dynamic AI-generated dialogue impact the player experience. Specifically how this differs between traditional and AI-generated character responses as well as how expectations in these factors can influence PX.

III. Methodology

The goal of this study was to determine the effect of AI-generated dialogue on player experience (PX) as well as to explore if a player's expectations towards a game can shape this experience. Three hypotheses were chosen to investigate whether the mere presumed perception of the AI feature would increase player engagement and whether this held true for both the actual *AI*- and the *HW* version. Further, it was examined which of the actual game versions performed better, irrespective of the game description the players received.

To test the hypotheses, two versions of the same murder-mystery RPG were developed in Unity, to be tested in a user study. One version of the game employed dialogue generated by OpenAI's GPT-4o (https://platform.openai.com/docs/models#gpt-4o), the other used handwritten, pre-scripted dialogue. The integration of GPT-4o in Unity was achieved with the unofficial Unity package "OpenAI-Unity" (https://github.com/srcnalt/OpenAI-Unity). The dialogues for the HW-version were first written and connected to dialogue trees in Twine (https://twinery.org). The resulting data was then parsed into appropriate data structures for Unity.

Both versions used the same plot, the main difference lay in the dialogue and its presentation within the game. This choice was made as a compromise to accommodate the scope of the master thesis this paper was based on. To achieve a handwritten version of approximately the same quality as the AI-Version, a significant amount of scripting would have been necessary. On the other hand, using the choice selection system of the HW-version for the AI-version would have would have diminished the core appeal of AI-chatbots. A mixed-method approach was chosen to gain further insight into what factors mainly contributed to the observed results.

We conducted a 2x2 factorial between-subjects study with two independent variables, "Game Type", (GT) and "Description Type" (DT), at two levels, "AI" and "Hand-Written"

(HW). The GT describes the actual game version participants played, meaning either with GPT-generated dialogue, GT *AI*, or with hand-written dialogue, GT *HW*. The DT denotes the game version the players were told that they would be playing with DT *AI* and DT *HW* as levels. In GT *AI* the player can ask any question they want in any way they want via free text input, whereas in the GT *HW* the player is more restricted by only being offered some responses to choose from.

Based on these experimental conditions, the participants were randomly split into four independent groups. Half of the participants played the AI version (GT AI), whereas the other half played the hand-written version (GT HW). Half of each of those groups were led to believe that they were playing the opposite version of the one they were actually playing (e.g., GT AI and DT HW; or GT HW and DT AI). For the other half of the participants the GT and the DT matched. While employing malevolent deception to entice people into buying a game that does not ship as advertised could be considered fraud, utilising this tactic to circumvent player biases when researching the effects of game features on the player experience, is within acceptable ethical standards [11].

Based on prior related research, we formulated three hypotheses on how the conditions influence the *Overall Player Experience* and each of the related factors *Immersion*, *Emotional Challenge* and *Perceived Autonomy*. $\boldsymbol{H_1}$ draws on the findings of previous work, where it was demonstrated that the mere presence of AI-enhanced content can significantly increase player immersion [8]. $\boldsymbol{H_2}$ is informed by research suggesting that AI-driven NPCs have the potential to enhance the overall gameplay experience [10]. Finally, $\boldsymbol{H_3}$ again builds on previous work, which showed that players can have an increased gameplay experience when they expect an AI-enhanced game, regardless of whether they are playing the AI-enhanced version or not [11].

These three hypotheses were applied to four key variables, *Overall Player Experience*, *Immersion*, *Perceived Autonomy* and *Emotional Challenge*, each of which is critical to understanding the overall gaming experience for this particular game. The variables were selected based on their relevance to player engagement and their ability to capture different dimensions of the gaming experience of a narrative-heavy game [1][12][25][26][9]. In the following, the previously mentioned key variables *Overall Player Experience*, *Immersion*, *Perceived Autonomy* and *Emotional Challenge* are symbolised by the placeholder **X**.

- $\boldsymbol{H_1}$: The **X** will be rated higher in the *AI*-description (DT AI) group.
- $\boldsymbol{H_2}$: The game type *AI* (GT AI) will have a higher rated **X** than game type *HW* (GT HW), regardless of description.
- $\boldsymbol{H_3}$: The *AI*-description group (DT AI) will have a higher rated **X** than the *HW*-description group (DT HW), in both game types.

A. Participants

Fifty-two participants (21 women, 31 men) were recruited via university mailing lists. The average age was 26.2 *(sd=4.57)*, ranging from 18 to 42. 69% of the participants regularly play video games. Among preferred game genres, RPG was in the top three mentioned (15 participants). Only four participants had played a game using AI for dialogue before the study. The average play time was four days a week *(sd=2.11)* for 2.8 hours a day *(sd=1.46)*.

B. Game

The game was designed as a 2D murder mystery RPG set in a small fictional village. The player takes on the role of a detective sent to the village to investigate the murder of a famous author. The objective of the game is to find the murderer only through dialogue with the three neighbours of the murder victim, one of which is the culprit. The player can move freely on the small map that only contains the NPCs and their respective houses.

Each of the NPCs was given a backstory including the relationship to all the other NPCs, as well as to the murder victim. The NPCs responses were either generated by GPT-4o or scripted, depending on the version played.

Among the tools available to the player is their cell phone, which can be used to call their boss to either receive a maximum of two hints or to report the crime. Further, the detective can make use of a notepad, which already contains the name and age of the victim, the time and cause of death as well as the location of the crime. Here the player can type in their notes to keep track of the case. To prevent the player from prematurely ending the game, certain hidden milestones were added. The player only had the option to report the murderer once all milestones had been reached.

The overarching narrative as well as the NPC's backstories were all human-written, without any input from AI. The main difference between the two versions was in the dialogue generation as well as the presentation of the dialogue within the game. The GPT version used free text input, whereas in the scripted version, players were presented with several responses to choose from, a dialogue system commonly found in RPGs.

C. GPT-version

The GPT version employed a concept previously described in related research [5] with a human-written story and an AI-generated dialogue. The main mechanic was the free text input feature, which allowed the players to ask any questions they wanted to (see Figure 1). This was implemented by sending one large initial prompt, containing the rule framework, to the GPT-API at the beginning of each game. Using prompt-engineering techniques recommended by OpenAI [27], the rule framework contained a short introduction to the game design and mechanics, followed by a short synopsis. Each NPC was described with information about them, their backstory and motives, their relationship to other NPCs as well as the murder victim. In addition, a list of independent instructions was added in the course of the development phase, which addressed undesired responses of the model.

Appended keywords or numbers at the end of user inputs and API outputs were used to monitor triggered game events and the dialogue partners of the player. For example, if the

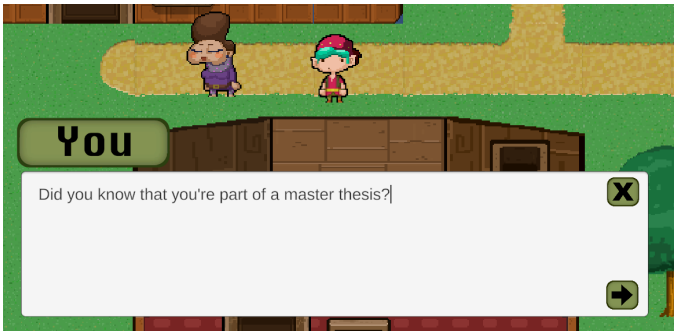


Fig. 1: Screenshot illustrating the GPT-version prompt interaction available to players in the game.

player started a chat with NPC 1, the number '1' would be appended to the user input to inform the model which NPC it should respond as.

D. Human-written version

The design of the dialogue in the *HW* version was based on the choice selection system commonly found in RPGs. Players could ask questions about what the NPCs were doing the night of the murder, what they know about the other NPCs and their relationship to the murder victim (see Figure 2). As this game was meant to be more of a demo than a full-fledged game, the choices were mostly kept to a maximum of two to avoid overly complex dialogue trees.

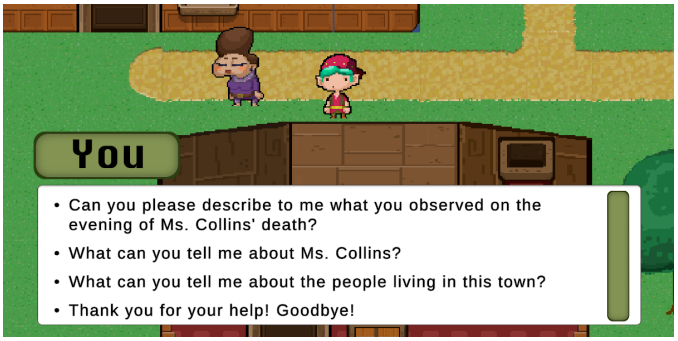


Fig. 2: Screenshot illustrating the dialogue choices available to players in the game.

However, to encourage exploration, some choices would be invisible until the player learned information from another NPC that would unlock them. Once a choice had been selected, it would be greyed out the next time the user would click through the dialogue.

E. Task

Participants were asked to play a murder mystery RPG where they needed to engage in conversation with NPCs to solve the mystery. They were free to explore the small map and interact with the NPCs in any order. They were given instruction on the objective of finding the murderer, and informed of the tools available to them (cell phone, notepad). The task was finished when the participant had entered a guess for the identity of the murderer or 20 minutes had passed.

F. Procedure

Before the study commenced, participants provided informed consent. They were told that the study aimed to investigate the effects of GPT-generated dialogue in a game and received a brief explanation of the two game versions. However, they were not informed that they might be playing a different version than the one described to them.

The experimenter told participants which game version they would play, disclosing only their assigned description type (DT). Participants then completed a questionnaire on demographics, their gaming habits, and prior experience with AI-driven dialogue in video games.

Next, they were reminded of the study's purpose and their assigned game version before answering questions about their expectations towards the game using a revised version of the iExpect scale [28].

Following this, participants received a brief explanation of the game's objective, controls and user interface. They were then instructed that they had a maximum of 20 minutes to play. If they had not finished by then, they were given one final opportunity to guess the murderer before proceeding to the main section of the questionnaire, which focused on their overall player experience.

Finally, participants took part in a short semi-structured interview to provide further insights into their experience.

G. Measures

To measure *Overall PX*, we used the miniPXI [29]. This version was shown to a reliable measure of overall PX and chosen over the regular PXI to reduce participant burden. The inventory consists of 11 Likert scale questions, each representing a subscale of the original PXI, which is used to measure an overall gaming experience. Due to the importance of this variable for the study, we also included the "immersion"-subscale of the full PXI to measure player *Immersion*.

To measure *Perceived Autonomy* we used two subscales, *Autonomy Satisfaction* and *Autonomy Frustration*, from the BANGS (Basic Needs in Games Scale) [30] – a psychometric tool that measures satisfaction and frustration of basic psychological needs (including relatedness and competence, which were not used in this study) in the context of gaming.

Lastly, for measuring *Emotional Challenge* we used one of the subscales from the CORGIS (Challenge Originating from Recent Gameplay Interaction Scale) [31] – a questionnaire measuring perceived challenge in games. The emotional challenge subscale contains nine questions.

All of the questions the dependent variables were presented on a 7-point Likert scale from 1="Strongly Disagree" to 7="Strongly Agree".

H. Data Analysis

A two-way ANOVA was used for the 2x2 factorial design. To assess normality and homogeneity of variances, the Shapiro-Wilk test and Levene's test were performed on the residuals. In one case, these assumptions were violated,

necessitating the use of the Mann-Whitney U test – a non-parametric alternative for comparing two independent groups.

A one-way ANOVA tested the hypothesis $\boldsymbol{H_3}$: *The description AI will have a higher X than description Scripted, in both game types*. As with the two-way ANOVA, normality and homogeneity of variances were tested with the Shapiro-Wilk and Levene's tests. When normality assumption was violated, a Kruskal-Wallis test with Dunn's test for the post-hoc analysis were used instead.

For statistically significant results in both two-way and one-way ANOVA tests, Tukey HSD identified pairwise differences. Spearman's Rank correlation assessed the influence of expectations from the revised iExpect scale.

Interviews were recorded and transcribed with the help of Turboscribe (https://turboscribe.ai/). The transcribed data was then analysed using content analysis.

IV. Results

A. Influence on Overall Player Experience

Although the average player experience (PX) was scored slightly higher in the GT *AI* group *(AI: M=5.6, sd=0.56; HW: M=5.4, sd=0.54)*, this difference was not significant. There was no interaction effect between GT and DT.

The type of description had a statistically significant effect on the *Overall PX*: *(F(1,48)=6.79, p=.012)*. The Tukey HSD post-hoc test indicated a significant difference in the means of the DTs *(p=.013)* (as shown in Figure 3), with participants interacting with the game in the DT *HW* condition having more positive overall PX.

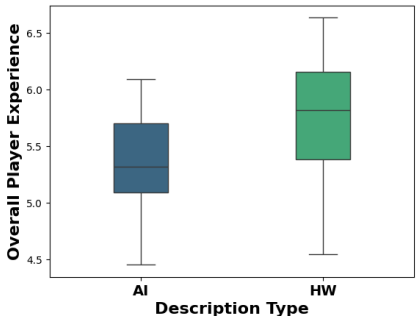


Fig. 3: Average overall PX for Description Types

The one-way ANOVA showed a significant influence of the experimental conditions on *Overall Player Experience*: *(F(1,48)=3.64, p=.019)*. The Tukey HSD post-hoc test confirmed that Group 2 (GT *AI*, DT *AI*) had a significantly higher influence than Group 1 (GT *AI*, DT *HW*) *(p=.046)*. Further, Group 2 also had a significantly higher influence than Group 3 (GT *HW*, DT *AI*) *(p=.021)* (see figure 4).

B. Influence on Immersion

The DTs also had a significant effect on *Immersion (F(1,48)=4.44, p=.040)*, while the GTs did not. There was no interaction effect. The Tukey HSD post-hoc test on the DTs showed that the difference between immersion ratings in the *HW* and *AI* conditions was statistically significant *(p=.038)*, where the human-written (*HW*) DT was perceived as more immersive (see figure 5).

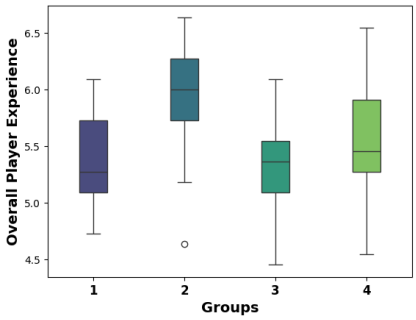


Fig. 4: Average group-wise Player Experience

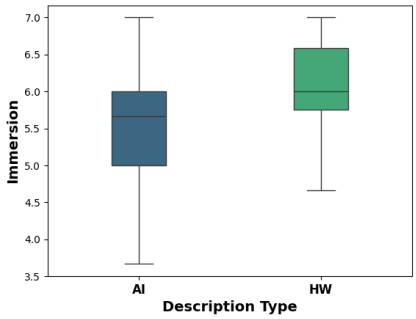


Fig. 5: Average Immersion for Description Types

None of the experimental conditions had a significant effect on *Immersion*.

C. Influence on Perceived Autonomy

Perceived Autonomy was comprised of the sub-categories *Autonomy Satisfaction* and *Autonomy Frustration*. As the two-way ANOVA assumptions were violated for *Autonomy Satisfaction*, the Mann-Whitney U test was performed on both the GTs and DTs. There was a significant difference between the two GTs *(z=3.26, p=.001)*, with GT *AI* resulting in higher *Autonomy Satisfaction* than the *HW* condition (see Figure 6).

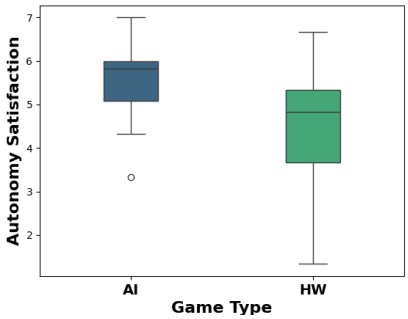


Fig. 6: Average Autonomy Satisfaction for Game Types

The effect of DTs on *Autonomy Satisfaction* was also significant: *(z=-1.69, p=.009)* with the DT *HW* having a more positive effect this time than the *AI* condition (see Figure 7).

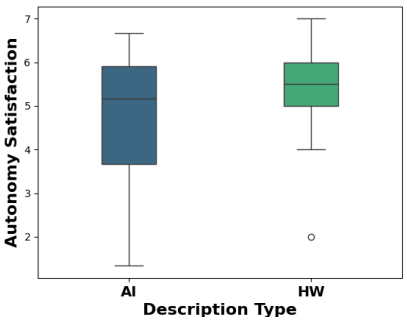


Fig. 7: Average Autonomy Satisfaction for Description Types

For the experimental condition, the Kruskal-Wallis test showed a significant influence on *Autonomy Satisfaction*: *(H(3)=15.5, p=.001)*. The Dunn's test revealed, that there was a significant difference between Group 1 (GT *AI*, DT *AI*) and 3 (GT *HW*, DT *AI*) *(p=.007)*, with Group 1 experiencing higher *Autonomy Satisfaction* as can be seen in figure 8. Further, Group 2 experienced significantly higher *Autonomy Satisfaction* than Group 3 *(p=.002)*.

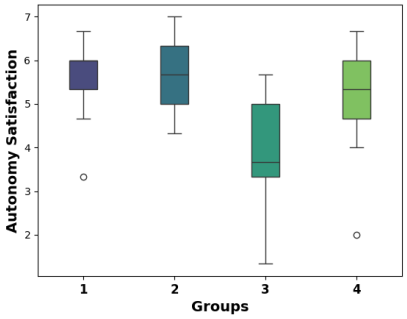


Fig. 8: Group-wise means for Autonomy Satisfaction

The results for *Autonomy Frustration* were analogous to those of *Autonomy Satisfaction*.

D. Influence of Expectations

Participants were asked to rate several statements of the *Anticipated Enjoyment* subscale of the revised iExpect scale [28] to gauge their expectations towards the game after being assigned their description type (DT). The Spearman's rank correlation coefficient showed a moderate negative correlation between *Autonomy Frustration* and expectations for the game *(ρ_s=-0.40, p=.045)*, with participants rating their expectation for the game high *(M=5.3, sd=0.84)*. For DT *HW* a near-significant moderate positive correlation was observed between participants' expectations and *Overall PX (ρ_s=0.36, p=.070)*, which indicated that participants with higher expectations tended to report better overall PXs.

E. Qualitative Results

Qualitative results were obtained from short, semi-structured post-questionnaire interviews. While most of the data gathered with this method was inconclusive, some important findings could be extracted, which will be presented in the following. The reported factors that contributed to immersion as well as major factors that contributed to enjoyment indicated that the narrative played a more deciding role for players of GT *HW* (*62%*) than those of GT *AI* (*27%*). Participants in the latter group felt the narrative and realistic NPC responses contributed to their feeling of immersion in approximately equal measure (*31%* and *23%*, respectively). More participants in GT *HW* (*73%*) and DT *HW* (*77%*) were motivated to keep playing because they wanted to solve the riddle when compared to their *AI* counterparts (GT *AI*: *62%*, DT *AI*: *58%*). The rated difficulty deviated only by 0.1 points in a 3-point Likert scale rating across all DTs and GTs *(sd≤0.7)*. The only noticeable difference between the groups with regards to the reasoning behind this result was a higher number of participants citing the lack of more clues or more definitive clues in DT *HW* (*27%*) compared to DT *AI* (*4%*). Concerning the influence on play-time, *35%* of all the participants stated that 20 minutes was not sufficient time to solve the riddle, *46%* of them said that they spent more time because they wanted to be sure to report the right subject. *46%* of the participants in GT *AI* added that they would have liked to play longer, as opposed to none in GT *HW*.

V. Discussion

Our findings show that the description type (DT) *HW* group reported significantly higher *Overall PX*, *Immersion* as well as *Perceived Autonomy* compared to the DT *AI* group, suggesting a strong influence of player expectations. The GT *AI* was associated with significantly higher *Autonomy Satisfaction* and lower *Autonomy Frustration*.

A. Overall Player Experience

Contrary to what was hypothesised in $\boldsymbol{H_1}$, *Overall PX* was rated higher in the DT *HW* group, which meant that the hypothesis for *Overall PX* could not be supported.

When comparing the average expectations between the DT groups, participants with DT *HW* also had significantly higher expectations than those with DT *AI*. The moderate positive correlation between the participant's expectations towards the game and their *Overall PX* in the DT *HW* group was, however, not significant. This meant that participants not only had a more positive playing experience in the DT *HW* group but also had correlating higher expectations once they were told their description type, regardless of which version they ended up playing. This is in contrast to the previous research, where players reported higher immersion and enjoyment when they expected adaptive AI technology to be present [11]. The DT *HW* group in this study had higher expectations and a more positive experience without the expectation of AI. This suggests, that the placebo effect observed in previous studies may not apply uniformly across all contexts of games.

Possible reasons for higher expectations in the DT *HW* group include a preference for familiar game formats, the belief that a human-written narrative would be of higher quality or simply scepticism towards the use of AI.

The results of the qualitative analysis suggest that participants in DT *HW* were more engaged because they were actively searching for additional clues, which would be in alignment with prior research which emphasised the importance of challenge to enable flow [9]. This is further supported by the fewer number of participants in DT *HW* struggling with the time constraint and stating that they would have preferred to play longer.

Contrary to expectations of $\boldsymbol{H_2}$ the GT *AI* group did not experience significantly higher PX. The qualitative data suggests that the free text input feature might not have had as strong of an impact on the player experience as expected. Moreover, participants may have perceived the level of autonomy as more overwhelming than desired, which could have diminished their enjoyment due to the lack of guidance. These findings portray a somewhat more positive outlook than previous work that showed that while the technology showed potential, it was not yet comparable to human-written narrative [32], [33]. However, for *Overall PX* $\boldsymbol{H_2}$ was also not supported.

$\boldsymbol{H_3}$ could not be supported as well. The combination of DT *AI* and GT *AI* (Group 1) and DT *AI* and GT *HW* (Group 3) had a significantly lower average than Group 2 (GT *AI*, DT *HW*). These findings further support the possible explanation for why $\boldsymbol{H_1}$ was not supported.

B. Immersion

Participants reported on average a high level of immersion, however, only those in the DT *HW* group experienced significantly higher *Immersion* compared to DT *AI* ($p<.001$). These findings again contradict the original expectations, therefore, $\boldsymbol{H_1}$ for *Immersion* is rejected. As mentioned in the previous section, an increased feeling of immersion could be attributed to the slightly higher rated difficulty and the possibly resulting engagement in the DT *HW* group, which could have enabled flow [9], which in turn is related to immersion [34]. Previous research offered two explanations for the increased immersion in their studies. The first one was that the search for adaptive elements caused increased engagement and, therefore, immersion. The second was that the players' belief in the supposed AI-technology was enough to cause an increase in immersion. The findings of this paper provide further evidence that while the belief in a certain game feature might be enough to elicit immersion, the differentiating factor appears to be the system in question. As with *Overall PX*, only a small disparity in *Immersion* could be observed between GT *AI* and *HW*, causing $\boldsymbol{H_2}$ to be rejected for *Immersion* as well. Again, most likely the free text input feature of the GT *AI* version had less of an impact on *Immersion* than the narrative. $\boldsymbol{H_3}$ is rejected as well, as none of the experimental conditions influenced *Immersion* in a significant way.

C. Perceived Autonomy

Participants in GT *AI* and DT *HW* experienced both significantly higher *Autonomy Satisfaction* and significantly lower *Autonomy Frustration* than their respective counterparts. The qualitative results did not provide any insights as to why DT *HW* influenced *Perceived Autonomy* at this level. The most likely assumption is that participants, not expecting AI, but then playing a game with high autonomy, would rate *Perceived Autonomy* higher than participants who expected a high autonomy in the first place. This leads to the rejection of $\boldsymbol{H_1}$ for *Perceived Autonomy*. However, in alignment with the expectations of $\boldsymbol{H_2}$, GT *AI* rated *Perceived Autonomy* significantly higher than GT *HW*, therefore, the hypothesis holds. Lastly, $\boldsymbol{H_3}$ is rejected, Group 1 (GT *AI*, DT *AI*) could not be observed to have a significantly higher *Perceived Autonomy* than Group 2 (GT *AI*, DT *HW*). In fact, Group 3 (GT *HW*, DT *AI*) experienced significantly lower autonomy, which was again, probably due to the unmet expectations for autonomy.

D. Limitations

To obtain a higher reliability of the findings, further research with a higher sample size should be considered. Future research could add additional pre-game questions pertaining to a possible bias towards AI or their thoughts on the use of AI in games. It might also be of interest to learn whether participants have used ChatGPT previously and are possibly more skilled in writing high-quality prompts. Additionally, questions could be asked about the value participants place on high-quality dialogue in games. The pre-game questionnaire should also include questions about the expected autonomy, to enable a comparison with the perceived autonomy and identify any discrepancies. In general, a more detailed examination of the participants' expectations could greatly improve the understanding of DT results. Further, more clarifying questions in the post-game interview might shed more light on the unexpected non-significant results for GT *AI* in *Overall PX* and *Immersion*.

VI. Conclusion

This paper highlights the strong influence of player expectations on gaming experiences. It suggests that in story-focused games, human-written narratives may shape expectations more strongly than the promise of dynamic AI features. Future work should explore potential biases towards the quality of human-written content or scepticism of AI-generated narratives to validate these findings. Additionally, it should be investigated whether these findings generalise to other game genres.

The results further indicate that the narrative, as well as the difficulty level, could be the main factors for immersion in these types of games. While the higher level of autonomy in the AI version led to an increase in player experience, future research needs to investigate the evolving role of AI in gaming. Designers should carefully evaluate whether AI-driven features align with the game's goals and add meaningful

value to the gameplay experience. Understanding how expectations influence player experience could provide insights for integrating AI in ways that resonate with players. This way, game designers can ensure that AI enhances, rather than detracts from, the overall gaming experience.

REFERENCES

[1] J. Wiemeyer, L. Nacke, C. Moser, and F. 'Floyd' Mueller, "Player Experience," in *Serious Games: Foundations, Concepts and Practice*, R. Dörner, S. Göbel, W. Effelsberg, and J. Wiemeyer, Eds. Springer International Publishing, pp. 243–271. [Online]. Available: https://doi.org/10.1007/978-3-319-40612-1_9

[2] K. Chu, C. Y. Wong, and C. W. Khong, "Methodologies for Evaluating Player Experience in Game Play," in *HCI International 2011 – Posters' Extended Abstracts*, C. Stephanidis, Ed. Springer, pp. 118–122.

[3] S. AlZu'bi, A. Mughaid, F. Quiam, and S. Hendawi, "Exploring the capabilities and limitations of chatgpt and alternative big language models," *Artificial Intelligence and Applications*, vol. 2, no. 1, p. 28–37, April 2023.

[4] "Enable dynamic AI-powered gameplay with player-facing AI." [Online]. Available: https://inworld.ai/engine#ai-npcs

[5] L. O'Brien, "How Ubisoft's New Generative AI Prototype Changes the Narrative for NPCs," Mar. 2024. [Online]. Available: https://news.ubisoft.com/en-us/article/5qXdxhshJBXoanFZApdG3L/how-ubisofts-new-generative-ai-prototype-changes-the-narrative-for-npcs

[6] Longest script for a videogame. Guinness World Records. [Online]. Available: https://www.guinnessworldrecords.com/world-records/764696-longest-script-for-a-videogame.html

[7] D. Bailey. Kingdom Come Deliverance 2 devs boast of a 2.2 million word script that would break Baldur's Gate 3's world record, and Larian's publishing director says "I hope they do!". gamesradar. [Online]. Available: https://www.gamesradar.com/games/rpg/kingdom-come-deliverance-2-devs-boast-of-a-2-2-million-word-script-\protect\penalty-\@Mthat-would-break-baldurs-gate-3s-world-record-\protect\penalty-\@Mand-larians-publishing-director-says-i-hope-they-do/

[8] A. Denisova and P. Cairns, "The Placebo Effect in Digital Games: Phantom Perception of Adaptive Artificial Intelligence," in *Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, ser. CHI PLAY '15. Association for Computing Machinery, pp. 23–33.

[9] M. Elson, J. Breuer, J. D. Ivory, and T. Quandt, "More than stories with buttons: Narrative, mechanics, and context as determinants of player experience in digital games," vol. 64, no. 3, pp. 521–542.

[10] "Survey: Nearly 3-in-4 game devs are excited about AI NPCs," Mar. 2024. [Online]. Available: https://inworld.ai/blog/future-of-game-development-with-ai-npcs-report?utm_source=www.boteatbrain.com&utm_medium=referral&utm_campaign=inworld-s-study-shows-that-3-in-4-developers-want-ai-npcs

[11] A. Denisova and P. Cairns, "Player experience and deceptive expectations of difficulty adaptation in digital games," vol. 29, pp. 56–68.

[12] R. M. Ryan, C. S. Rigby, and A. Przybylski, "The Motivational Pull of Video Games: A Self-Determination Theory Approach," vol. 30, no. 4, pp. 344–360.

[13] "NVIDIA ACE for Games Sparks Life Into Virtual Characters With Generative AI | NVIDIA Newsroom," May 2023. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-ace-for-games-sparks-life-into-virtual-characters-with-generative-ai

[14] J. Loynds, "ChatGPT is being used in games to generate AI NPC dialogue," Feb. 2023. [Online]. Available: https://www.dexerto.com/tech/chatgpt-games-to-generate-ai-npc-dialogue-2063163/

[15] "How NetEase integrated Inworld AI in Cygnus Enterprises, Part 1," Nov. 2023. [Online]. Available: https://inworld.ai/blog/netease-cygnus-enterprise-integrated-ai-npcs-inworld

[16] "How AI NPCs helped MageCosmos increase playtime by 5%." [Online]. Available: https://inworld.ai/case-study/how-ai-npcs-helped-magecosmos-increase-playtime-by-5-percent

[17] D. Yang, E. Kleinman, and C. Harteveld, "Gpt for games: A scoping review (2020-2023)," in *2024 IEEE Conference on Games (CoG)*, 2024, pp. 1–8.

[18] A. Matarazzo and R. Torlone. A Survey on Large Language Models with some Insights on their Capabilities and Limitations. [Online]. Available: http://arxiv.org/abs/2501.04040

[19] N. F. Ayoub, K. Balakrishnan, M. S. Ayoub, T. F. Barrett, A. P. David, and S. T. Gray, "Inherent Bias in Large Language Models: A Random Sampling Analysis," vol. 2, no. 2, pp. 186–191. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2949761224000208

[20] S. Sabour, S. Liu, Z. Zhang, J. M. Liu, J. Zhou, A. S. Sunaryo, J. Li, T. M. C. Lee, R. Mihalcea, and M. Huang. EmoBench: Evaluating the Emotional Intelligence of Large Language Models. [Online]. Available: http://arxiv.org/abs/2402.12071

[21] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, p. 1–55, Jan. 2025. [Online]. Available: http://dx.doi.org/10.1145/3703155

[22] S. R. Cox and W. T. Ooi, "Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback," in *Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, November 22–23, 2023, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, Mar. 2024, pp. 167–184. [Online]. Available: https://doi.org/10.1007/978-3-031-54975-5_10

[23] T. Ashby, B. K. Webb, G. Knapp, J. Searle, and N. Fulda, "Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-based Approach," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–20. [Online]. Available: https://doi.org/10.1145/3544548.3581441

[24] N. Akoury, Q. Yang, and M. Iyyer, "A Framework for Exploring Player Perceptions of LLM-Generated Dialogue in Commercial Video Games," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 2295–2311. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.151

[25] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny, "The development of the game engagement questionnaire: A measure of engagement in video game-playing," *Journal of Experimental Social Psychology*, vol. 45, no. 4, pp. 624–634, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022103109000444

[26] W. A. IJsselsteijn, Y. A. De Kort, and K. Poels, *The Game Experience Questionnaire*. Technische Universiteit Eindhoven.

[27] OpenAI Platform. [Online]. Available: https://platform.openai.com

[28] S. Villa, T. Kosch, A. M. Kloft, J. Quinn, S. Kujala, and R. Welsch, "Inventory of user expectations for technology (iexpect)," *Submitted to, CHB 2025*, Sep. 2025.

[29] A. Haider, C. Harteveld, D. Johnson, M. V. Birk, R. L. Mandryk, M. Seif El-Nasr, L. E. Nacke, K. Gerling, and V. Vanden Abeele, "miniPXI: Development and Validation of an Eleven-Item Measure of the Player Experience Inventory," vol. 6, pp. 244:1–244:26.

[30] N. Ballou, A. Denisova, R. Ryan, C. S. Rigby, and S. Deterding, "The basic needs in games scale (bangs): A new tool for investigating positive and negative video game experiences," *International Journal of Human-Computer Studies*, vol. 188, p. 103289, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1071581924000739

[31] A. Denisova, P. Cairns, C. Guckelsberger, and D. Zendle, "Measuring perceived challenge in digital games: Development & validation of the challenge originating from recent gameplay interaction scale (CORGIS)," vol. 137, p. 102383.

[32] S. R. Cox and W. T. Ooi, "Conversational Interactions with NPCs in LLM-Driven Gaming: Guidelines from a Content Analysis of Player Feedback," in *Chatbot Research and Design: 7th International Workshop, CONVERSATIONS 2023, Oslo, Norway, November 22–23, 2023, Revised Selected Papers*. Springer-Verlag, pp. 167–184.

[33] T. Ashby, B. K. Webb, G. Knapp, J. Searle, and N. Fulda, "Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-based Approach," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, ser. CHI '23. Association for Computing Machinery, pp. 1–20. [Online]. Available: https://doi.org/10.1145/3544548.3581441

[34] E. Kannegieser and D. Atorf, "A study to further understand the link between immersion and flow," in *Interactivity and the Future of the Human-Computer Interface*. IGI Global, 2020, pp. 114–122.