Open-Ended NPC Dialogue Favors Casual Players: A Pilot Comparison of Three LLM-Driven Dialogue Systems

Rasmus Ploug*, Emil Rimer†, Anthon Kristian Skov Petersen‡ and Marco Scirea§
SDU Metaverse Lab, University of Southern Denmark
Odense, Denmark
{*raplo20, †emrim19, ‡antpe20}@student.sdu.dk; §msc@mmmi.sdu.dk

***Abstract*—Non-player character (NPC) dialogue plays a crucial role in shaping the player experience in narrative-driven video games, influencing agency, immersion and story engagement. Despite the recent advancements in large language models (LLMs) for dynamic dialogue generation, few empirical studies have compared their impact across different dialogue system designs.**

This pilot study explores how LLM-driven dialogue systems affect the player experience using a custom-developed role-playing game (RPG) featuring four different dialogue designs; static control (CV), rephrase (A), hybrid (B) and fully open-ended (C). Behavioral data and post-game questionnaires were collected from 64 participants.

Results indicate that fully open-ended dialogues led to significantly longer dialogue interactions and higher overall engagement, particularly among casual players, with the survey feedback highlighting its immersive and natural tone. These findings suggest that fully open-ended LLM-based dialogue in video games can enhance narrative depth and player involvement.

***Index Terms*—Large Language Models, NPC Dialogue, Casual Players, Player Engagement, Dialogue Systems, Procedural Narrative**

I. Introduction

Recent advances in large language models (LLMs) offer new possibilities for non-player character (NPC) dialogue generation in video games, but few empirical studies have compared their impact on player experience. Dialogue systems are often central to narrative-driven games, shaping how players engage with story, NPCs, and decision-making. While LLMs enable dynamic, open-ended interactions that move beyond traditional branching dialogue-trees, many of their benefits and trade-offs remain underexplored.

This paper presents a pilot study conducted within a role-playing game (RPG) framework to examine how different dialogue systems influence player interaction, engagement, and perceived immersion. Four versions of the game were developed and tested, each with its own unique dialogue system: a static control version (CV) and three LLM-driven variants; rephrase (A), hybrid (B) and fully open-ended (C). By analyzing both behavioral data and player feedback, the paper aims to inform the design of future LLM-integrated dialogue systems. The research questions for this study are as follows:

- **RQ1**: How does the integration of different LLM-driven features in NPC dialogues influence player interaction in video games?
- **RQ2**: Did any LLM-driven feature enhance player interaction compared to the Control Version?

This paper starts off by reviewing relevant background about traditional LLM-based dialogue systems, followed by a detailed explanation about the prototype design, test methodology and data collection. The results of the user tests are then presented. The implications from these results are subsequently discussed regarding LLM-driven dialogue design. Finally, potential directions for future research are outlined, followed by a conclusion that summarizes the study's key findings.

II. Traditional vs Procedural Dialogue

Traditional dialogue systems in games are typically built around branching dialogue trees, where players choose from fixed dialogue options that lead to predetermined outcomes [1]. This structure ensures consistency and authorial control, but often results in predictable and repetitive experiences across playthroughs [2]. While games like *Baldur's Gate 3* and *Disco Elysium* offer highly reactive narratives, such depth demands significant writing and design effort [3].

To address scalability and replayability issues, researchers and developers have explored procedural dialogue systems where NPC responses are generated at runtime rather than fully pre-authored [4]. With the rise of LLMs, procedural dialogue has become increasingly viable. LLMs can generate character-consistent, context-sensitive replies that adapt to player input, reducing reliance on static dialogue trees [5].

Recent examples include *AI Dungeon* [6], which offers open-ended storytelling through text prompts, and modding projects like *Mantella* for *Skyrim*, which demonstrate

the potential of LLMs in immersive, speech-driven dialogue [5]. However, challenges remain, such as inconsistent or off-tone responses [7].

Together, these trends signal a shift from static, authored conversations to flexible, player-driven interactions powered by generative systems.

III. Methodology

A. Game Prototype Design

A video game prototype has been developed to explore how different LLM-driven dialogue systems influences the player experience (Fig. 1). The game was created in Unity and all assets are from the Unity Asset Store. The prototype consists of a narrative game set in a medieval RPG setting where the player has to talk to different NPCs in order to complete a string of quests. The game is divided into four versions, each of which keeping the same identical game mechanics, narrative and visual design. The main difference between these versions comes in how the individual dialogue systems are implemented. The game serves as the test environment in which players engage with the different LLM-driven dialogue systems by interacting with NPCs, allowing for comparative analysis of player behavior and experience. The different game versions are as follows:

- **Control Version:** Uses a static, tree-like dialogue structure typical for traditional RPGs, with no LLM integration. Players progress through the conversation by selecting from predefined dialogue options (Fig. 2a). This version serves as a baseline for evaluating the other game versions.
- **Version A:** Uses an LLM to rephrase the static dialogue from the Control Version each time an NPC interaction begins, creating linguistic variation. Both NPC lines and player responses are rephrased, while the player still selects from dialogue options to continue the conversation (Fig. 2b).
- **Version B:** Uses a hybrid approach combining static dialogue options (from the Control Version) with free-form input. The latter allows players to write custom messages, which an LLM uses to generate dynamic dialogue tailored to the NPC's dialogue tree (Fig. 2c).
- **Version C:** Uses a fully open-ended dialogue system where players write custom input to NPCs. This input is processed by an LLM, which generates dynamic responses based on the player's message and the game's narrative context (Fig. 2d).

B. Participants and Procedure

A user study was conducted to examine user behavior. Data was collected passively during gameplay, allowing participants to engage naturally with the game. The study followed a three-part protocol:

1) An introductory briefing to explain the purpose of the test and how data would be collected.
2) A playthrough of the game using a randomly assigned version of the dialogue system.
3) A custom questionnaire to gather qualitative insights.

Fig. 1: Top: Town with NPCs. Bot: Player and Innkeeper

The test was conducted in a controlled environment with a facilitator present to ensure the sessions ran smoothly. Ethical concerns where considered and participants were made to tag their data so it could potentially be retracted post-test. To ensure even distribution across game versions, an algorithm assigned new participants to the least-tested version. The target profile was individuals who play video games, with requirements including general knowledge of video games and the ability to read and write in English.

C. Data Collection

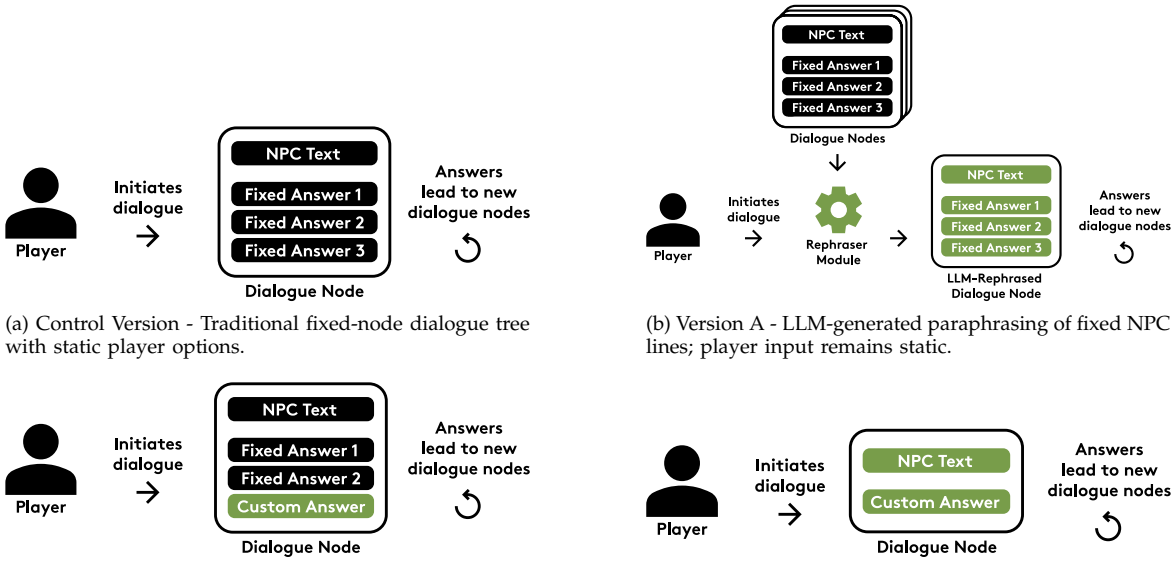
To analyze interaction differences between dialogue systems, both behavioral data and post-game feedback were collected. This allowed for a combination of quantitative and qualitative analysis.

1) Game Logging: Each play session logged data to one of four NoSQL databases, depending on the dialogue system. Logged collections included dialogue exchanges (with timestamps), dialogue time, movement (coordinates every 0.25s), area time and transitions, as well as other metrics like run time, quest outcomes, LLM wait time, and input type (Version B only). Sessions were linked to unique run codes for analysis.

2) Survey Responses: At the end of the session, participants completed a pre-filled questionnaire with 24 Likert-scale items and four open-ended questions. It covered consent, background, technical data, gameplay impressions, dialogue evaluation, clarity, engagement, AI perception, and task difficulty.

IV. Results

A total of 64 participants took part in the study, with 62 included in the final analysis. The mean age was 25.2



(a) Control Version - Traditional fixed-node dialogue tree with static player options.

(b) Version A - LLM-generated paraphrasing of fixed NPC lines; player input remains static.

(c) Version B - Hybrid approach with both fixed options and open-ended free-text input.

(d) Version C - Fully open-ended input interpreted by an LLM for all player messages.

Fig. 2: Comparison of dialogue system versions. Black boxes represent human-made text, while green represents LLM-generated text.

years ($SD = 4.57$), and 77.4% identified as male. Reported dyslexia occurred in 11.3% of participants. Participants were evenly distributed across the four dialogue versions. Familiarity with RPGs was high ($M = 4.34$, $SD = 1.10$), while experience with LLM-based dialogue systems was lower ($M = 2.35$, $SD = 1.33$).

A. Gameplay and Dialogue Time

Version C resulted in significantly longer total gameplay time ($p = 0.0064$, $\eta^2 = 0.152$). However, when dialogue time was subtracted, no significant difference remained ($p = 0.3739$), suggesting the added duration stemmed from extended conversations rather than slower gameplay.

Dialogue time showed a strong overall effect ($p < .0001$, $\eta^2 = 0.392$), with Version C significantly higher than all others. Pairwise tests confirmed that participants in Version C spent more time in conversation than those in the Control ($p < .0001$), A ($p = 0.0003$), or B ($p = 0.0048$) versions.

Among casual players ($\leq$ 10 hours gaming/week), dialogue time remained significantly higher in Version C, while completion time did not differ.

B. Interaction Quality and System Responsiveness

Participants in the Control Version initiated significantly more dialogue interactions than those in Version A ($p = 0.0105$, $\eta^2 = 0.135$), but Version B led to longer conversations per interaction ($p < .0001$, $\eta^2 = 0.024$). Version C balanced both volume and depth, with high dialogue time but a moderate number of turns per conversation.

Version C also had the lowest average LLM loading time ($p < 0.0001$, $\eta^2 = 0.516$), yet the highest total LLM load time over the full playthrough ($p < 0.0001$, $\eta^2 = 0.544$), reflecting the longer engagement.

Survey responses indicate that participants found the NPCs in Version C more interesting than those in Version A ($p = 0289$, $\eta^2 = 0.099$). The question "*The dialogue system made the experience enjoyable*" also reached significance ($p = 0.0402$, $\eta^2 = 0.087$), though post-hoc comparisons were not significant.

C. Qualitative Impressions

Players in Version C highlighted the freedom and expressiveness of open-ended input. One wrote, "*I liked that you could write your own options; it felt more natural*" while another noted it was "*different than most games where you choose between three options.*" The ability to engage more freely contributed to a sense of immersion and role-play.

Minor issues were also raised. One participant mentioned receiving "*a wrong answer as to how to cross the forest,*" and another noted that "*they sometimes push the conversation forward a bit unprompted*".

Overall, responses suggest that Version C supported deeper interaction and felt intuitive to many players, especially compared to more restrictive systems.

V. Discussion

The findings show that dialogue system design had a measurable impact on gameplay, engagement, and player perception. Version C stood out across both objective and subjective data, particularly for casual players, who engaged significantly longer in dialogue without

extending their total playtime. These results suggest that open-ended dialogue can support deeper, more efficient interaction, especially for players with less gaming experience.

A. System Comparison

Table I summarizes each version's performance. While the Control Version was stable and fast, it lacked flexibility and depth. Some players felt it did not reflect the personality of the NPCs, which reduced immersion. Version A introduced paraphrasing, but the added load time and minimal variation diminished its impact. Its core content rarely changed, and many players never revisited NPCs to experience the system's intended diversity. However, it maintained strong narrative control, making it suitable for tightly authored games.

Version B offered a promising hybrid model that allowed players to ask custom questions. This encouraged exploration and helped players navigate difficult moments. Still, it showed the most unpredictable behavior, occasionally producing confusing or unnatural responses. It received high marks for interest and engagement, but the inconsistencies reduced perceived control.

Version C produced the longest dialogues and was rated as the most engaging. It supported spontaneous, in-character responses and fostered immersion. Notably, most of its player inputs stayed grounded in the game world, signaling effective role-play. This system may particularly favor casual gamers, who are less familiar with rigid dialogue trees and more accustomed to natural conversation. By mimicking real-life dialogue patterns, Version C likely reduced the barrier to interaction for less experienced players, allowing them to express intentions more intuitively. However, it required manual typing and had the longest LLM response time. While load times were rarely noted as a problem, they could become more significant in longer games. Additionally, its effectiveness may depend on the player's willingness to actively engage with the system.

Metric	CV	A	B	C
Engagement	+	−	++	+++
Flexibility	+	+	++	+++
Stability	+	+	− −	−
Performance	+	−	− −	−
Player Agency	+	+	++	+++
Naturalness	+	+	±	++
Dev Control	+	±	±	− −

TABLE I: Comparison of dialogue systems across key design metrics.

VI. Future Work

Several directions for future research remain. The results warrant further exploration into how open-ended dialogue systems (Version C) affect larger and more diverse population samples, as well as across different game genres, to assess generalizability.

Secondly, the hybrid design (Version B) shows promise as a middle ground between structure and freedom, but suffers from stability issues and occasional inconsistencies in dialogue generation. Future work should focus on improving the technical reliability of this system, particularly through improved prompt engineering and parameter tuning to reduce hallucinations and ensure consistent NPC behavior.

Finally, although the automated rephrase system (Version A) yielded minimal impact in this study, its underlying functionality may offer value in other use cases or applications. Therefore, future research should explore the effectiveness of rephrasing dialogue in alternative contexts to explore the field even further.

VII. Conclusion

This study presents a pilot study of four different dialogue systems in a custom-developed role-playing game (RPG); a static control version (CV) and three variants (A, B and C) powered by large language models (LLMs). The findings demonstrate that LLM-driven features meaningfully shape player interaction, with each version offering distinct strengths and trade-offs. Furthermore, the results showed that the most open-ended dialogue system (Version C) led to significantly longer and more engaging dialogue interactions, particularly among casual players. These findings suggest that different LLM-driven dialogue systems have the potential to enhance player experiences, although further research and improvements must be conducted to adopt these systems in commercial games.

References

[1] M. Mateas and A. Stern, "Façade: An experiment in building a fully-realized interactive drama," in *Game developers conference*, vol. 2, pp. 4–8, Citeseer, 2003.

[2] C. Crawford, *Chris Crawford on interactive storytelling*. Pearson Education, 2004.

[3] M. O. Riedl and R. M. Young, "Narrative planning: Balancing plot and character," *Journal of Artificial Intelligence Research*, vol. 39, pp. 217–268, 2010.

[4] M. O. Riedl and V. Bulitko, "Interactive narrative: An intelligent systems approach," *Ai Magazine*, vol. 34, no. 1, pp. 67–67, 2013.

[5] R. Gallotta, G. Todd, M. Zammit, S. Earle, A. Liapis, J. Togelius, and G. N. Yannakakis, "Large language models and games: A survey and roadmap," *IEEE Transactions on Games*, 2024.

[6] M. Hua and R. Raley, "Playing with unicorns: Ai dungeon and citizen nlp.," *DHQ: Digital Humanities Quarterly*, vol. 14, no. 4, 2020.

[7] 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*, pp. 2295–2311, 2023.