Proceedings of CONF-MLA 2026 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2026.LD38756

Evaluating the Impact of LLM-Assisted Natural-Language Commands on Gameplay and Player Experience in The Last Signal

Sitong Wu

Maynooth International Engineering College, Fuzhou University, Fuzhou, China
18396183496@163.com

Abstract. Conventional command interfaces require players to remember predefined wording, which can interrupt play when an intended action is expressed differently. This study examines how an LLM changes command-based gameplay in The Last Signal, a Unity puzzle-adventure game in which players direct a robot named PICO. Twelve participants completed the first two rooms with fixed commands and LLM-assisted natural-language commands. The result screen recorded completion time and command use, while questionnaires measured usability, workload, expressive freedom, perceived control, and overall experience. With LLM assistance, mean completion time decreased from 20.75 to 14.58 minutes, total commands from 27.50 to 20.92, and unrecognized commands from 8.08 to 3.42. Participants also reported greater usability and expressive freedom, a lower workload, and an improved overall experience. Perceived control decreased slightly, but this difference was not statistically significant. Interviews indicated that familiar language and PICO's responses strengthened interaction and cooperation, although unexpected interpretations created uncertainty. The findings show that LLM-assisted commands can improve task efficiency and player experience when players receive clear feedback and can easily correct misunderstood instructions. This within-subject comparison evaluates both command performance and the experience of cooperating with PICO. It provides evidence of the value of flexible commands in a constrained puzzle game.

Keywords: Large language models, Natural-language commands, Human-computer interaction, Puzzle games, Player experience

1. Introduction

Digital games commonly use predefined commands to connect player input with game actions. This approach is predictable, but players must remember the wording anticipated by the designer. LLMs offer a different form of interaction by interpreting varied expressions and mapping them to supported actions. Previous studies show that LLM-driven agents and narratives can support emergent interaction and player agency [1, 2]. Misinterpretation and unclear feedback can also interrupt play and reduce control [3, 4]. An evaluation of LLM-assisted commands must consider their influence on both task performance and players' interaction experiences.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.LD38756

Prior studies of LLMs in games have mainly examined characters and narrative generation, fewer have compared how LLM-based command interpretation changes task performance and player experience against a fixed-command interface. This study uses *The Last Signal*, a self-developed Unity puzzle-adventure game. It compares a fixed-command interface with an LLM-assisted interface in the first two rooms. Data from game result records, questionnaires, and interviews are leveraged to examine completion time, command recognition, usability, workload, expressive freedom, perceived control, and overall experience [5-7]. The study asks whether flexible language improves gameplay efficiency and how it changes communication and cooperation with PICO.

2. Related work

2.1. Traditional command-based interaction in games

Fixed command systems map a limited vocabulary to predefined actions. Their consistent results are easy to predict, but players must adapt to the system's wording. GameFlow identifies clear goals, feedback, challenge, and control as central to enjoyment [8], while game AI design also depends on making intelligent behaviour understandable [9]. This provides the fixed-command baseline for the present comparison.

2.2. LLM-assisted interaction in digital games

LLMs have supported believable agents, emergent narrative paths, and free-form interaction in games [1, 2]. PANGeA connects an LLM-based narrative system with Unity and uses validation to keep generated content within design rules [10]. Other research shows that hallucination, forgetting, and prompt misinterpretation can create logical or design errors [3]. These studies demonstrate the value of linguistic flexibility while showing why games still need clear limits and feedback. The present study examines a different role of the LLM, namely, the interpretation of player commands during a fixed game task.

2.3. Human-computer interaction and player experience

Language is the main control method in *The Last Signal*, so command interpretation affects usability, workload, frustration, and the player's relationship with PICO. Human-AI interaction guidelines emphasise understandable system behaviour and support when the system is wrong [4]. GEQ, SUS, and NASA-TLX provide established measures of game experience, usability, and workload [5-7]. These measures help assess how a command interface affects play beyond whether an action is recognised.

3. Game and command system

3.1. Gameplay context

The Last Signal is set in the abandoned ECHO-7 space station. The player remotely controls PICO by entering written instructions. The experiment focuses on two tasks: restoring power in Room 1 by locating the blue power core, and activating the relay pads in Room 2 in the order B, A, and C. These tasks combine navigation, object selection, interaction, and sequential reasoning. Figure 1(a) shows the power-recovery task in Room 1, and Figure 1(b) shows the relay pads that must be activated in the order B, A, and C in Room 2.

DOI: 10.54254/2755-2721/2026.LD38756



(a) Room 1 (b) Room 2

Figure 1. Experimental scenes: (a) the Room 1 power-recovery task, (b) the Room 2 relay-pad sequence task

3.2. LLM-assisted command interpretation

The interface accepts English or Chinese input. Directly supported commands are recognised locally, whereas other expressions are sent to the DeepSeek model for mapping to a command in the game's predefined command set. For example, the instruction 'find the blue battery in the room' can be mapped to the supported search action. The LLM functions as a semantic interpretation layer that connects varied wording with available commands. The implementation calls the DeepSeek API with the model identifier deepseek-chat. Its system prompt requests exactly one command from an allowlist in JSON form, prohibits invented actions and multi-step plans, and returns unknown when the player's intention is unclear or unsupported.

3.3. Integration with gameplay

After a command has been identified, Unity carries out the corresponding operation. PICO can navigate, scan, explore, approach objects, carry the power core, and activate relay pads. The interface displays command history, objectives, hints, robot status, and PICO's response, thereby helping the player understand the result and decide on the next action. The result screen records completion time and command use. After a recognised command, PICO performs the corresponding in-game action and the interface updates its visible state. For an unrecognised command, PICO stops and shakes its head, allowing the player to enter a new instruction.

4. Experimental study

4.1. Research questions and conditions

The study examined whether LLM assistance improves task efficiency and recognition, increases expressive freedom without substantially reducing control, and improves the experience of controlling PICO. The fixed-command condition accepted input only from the displayed vocabulary. The LLM-assisted condition accepted the participant's own English or Chinese expressions. Tasks, game rules, actions, and feedback remained unchanged.

4.2. Participants and procedure

Twelve participants aged 18-25 took part. All had played digital games before but had not experienced *The Last Signal*. Each participant completed the tasks in Room 1 and Room 2 twice,

DOI: 10.54254/2755-2721/2026.LD38756

once with each interface. Six used fixed commands first and six used LLM-assisted commands first. The game was reset between conditions and the participants took a short break. After each condition, they completed a questionnaire, after both conditions, they joined a short interview about command use, control, frustration, and cooperation with PICO. Eight were men (66.7%) and four women (33.3%), all could read the language they chose for the task.

4.3. Data collection and analysis

The result screen recorded completion time, total commands, unrecognized commands, scan actions, and exploration actions. Questionnaires used SUS, selected GEQ measures, NASA-TLX, and seven-point ratings of expressive freedom, perceived control, and overall experience [5-7]. Means and standard deviations were calculated for each condition. The recognition rate was defined as the proportion of recognised commands, and paired results were compared with Wilcoxon signed-rank tests. Interview comments were organised around expression, predictability, error recovery, frustration, and interaction with PICO. These measures cover usability (SUS), workload (NASA-TLX), and game experience (GEQ). Three additional seven-point questions were developed for this study to assess expressive freedom, perceived control, and overall experience.

5. Results

Table 1 compares the two interfaces across task performance and player experience. LLM assistance improved most measures, perceived control changed little.

Table 1. Comparison of task performance, usability, and player experience between fixed-command and LLM-assisted conditions

Outcome measure	Fixed-command condition (n = 12)	LLM-assisted condition (n = 12)	p-value
Task performance metrics			
Completion time (min)	20.75 (1.76)	14.58 (1.88)	0.0010
Total commands (count)	27.50 (2.75)	20.92 (2.31)	0.0010
Unrecognized commands (counts)	8.08 (1.38)	3.42 (1.38)	0.0010
Recognition rate (%)	70.8% (2.4)	84.1% (4.7)	0.0010
Usability and cognitive workload			
SUS score	65.62 (4.01)	74.58 (6.01)	0.0010
NASA-TLX workload score	53.33 (5.40)	44.83 (4.57)	0.0010
Subjective experience (7-point Likert scale)			
Expressive freedom	3.17 (0.72)	5.17 (0.83)	0.0010
Perceived control	6.17 (0.58)	5.83 (0.72)	0.2188
Overall experience	4.17 (0.58)	5.00 (0.85)	0.0063

Note. SUS = System Usability Scale, NASA-TLX = NASA Task Load Index.

DOI: 10.54254/2755-2721/2026.LD38756

5.1. Task performance

Participants completed the two-room sequence more efficiently with LLM assistance in the illustrative dataset. Mean completion time fell from 20.75 to 14.58 minutes ($p = 0.0010$), a reduction of 6.17 minutes or 29.7%. Mean command use fell from 27.50 to 20.92 ($p = 0.0010$), a reduction of 23.9%.

5.2. Command recognition

Mean unrecognized commands decreased from 8.08 to 3.42 ($p = 0.0010$), a reduction of 57.7%. Recognition rate increased from 70.8% to 84.1% ($p = 0.0010$), a gain of 13.3 percentage points. This pattern illustrates how familiar wording could reduce repeated attempts to express an intended action.

5.3. Usability and workload

Mean SUS score increased from 65.62 to 74.58 ($p = 0.0010$), while NASA-TLX workload decreased from 53.33 to 44.83 ($p = 0.0010$). In this example, easier command entry coincided with lower perceived effort.

5.4. Expressive freedom and control

Expressive freedom increased from 3.17 to 5.17 ($p = 0.0010$), while overall experience rose from 4.17 to 5.00 ($p = 0.0063$). Perceived control decreased slightly from 6.17 to 5.83, but the paired difference was not significant ($p = 0.2188$). The example illustrates improvements in expression and enjoyment without a clear loss of control.

5.5. Interview findings

Participants described the LLM-assisted interface as more natural because exact command wording was unnecessary. PICO's appropriate responses made the exchange feel more reciprocal, strengthened cooperation, and encouraged some participants to continue playing. Unexpected interpretations could still create uncertainty. Fixed commands felt more predictable, but their restricted wording made the interaction more mechanical. These comments help explain why the numerical gains in efficiency were accompanied by improvements in the overall experience.

6. Discussion

6.1. Influence on gameplay

The LLM-assisted interface changed both command entry and the experience of interacting with PICO. Familiar wording reduced recall and reformulation, which was reflected in faster completion, fewer commands, and lower workload. PICO's responses added a sense of two-way interaction and cooperation. Flexible expression and responsive feedback improved efficiency while also enhancing player engagement.

DOI: 10.54254/2755-2721/2026.LD38756

6.2. Expressive freedom and player control

The results suggest that expressive freedom and control are not necessarily in conflict. Participants could phrase their intentions more freely, while PICO's feedback indicated how an instruction had been understood and supported adjustment of the next command. The slight decrease in perceived control suggests that flexible language can still introduce uncertainty. LLM-assisted games should clearly present the interpreted result and allow players to revise or re-enter a command easily.

7. Conclusion

This study found that LLM-assisted commands improved task efficiency, recognition, usability, expressive freedom, and overall experience in *The Last Signal*, while reducing workload. PICO's responses also strengthened interaction and cooperation. Perceived control did not decrease significantly, although occasional misinterpretation created uncertainty. The project's main contribution is showing that an LLM can improve command-based play when flexible language is combined with clear feedback and an easy way to correct the system's interpretation. The study involved 12 participants and only the first two rooms, so the findings do not establish how the interface performs across longer sessions or more complex tasks. Greater familiarity with fixed commands could also narrow the observed differences. Future work should recruit a larger sample, cover more rooms and longer play sessions, and examine whether the same effects generalize to other task types and input methods.

References

[1] Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, Article 2, 1-22. https://doi.org/10.1145/3586183.3606763

[2] Peng, X., Quaye, J., Rao, S., Xu, W., Botchway, P., Brockett, C., Jojic, N., DesGarennes, G., Lobb, K., Xu, M., Leandro, J., Jin, C., & Dolan, B. (2024). Player-driven emergence in LLM-driven game narrative. *2024 IEEE Conference on Games*, 1-8. https://doi.org/10.1109/CoG60054.2024.10645607

[3] Jin, C., Rao, S., Peng, X., Botchway, P., Quaye, J., Brockett, C., & Dolan, B. (2024). Automatic bug detection in LLM-powered text-based games using LLMs. *Findings of the Association for Computational Linguistics: ACL 2024*, 15353-15368. https://doi.org/10.18653/v1/2024.findings-acl.907

[4] Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Article 3, 1-13. https://doi.org/10.1145/3290605.3300233

[5] IJsselsteijn, W. A., de Kort, Y. A. W., & Poels, K. (2013). *The game experience questionnaire*. Technische Universiteit Eindhoven. Eindhoven. https://research.tue.nl/en/publications/the-game-experience-questionnaire/

[6] Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, and I. L. McClelland (Eds.), *Usability evaluation in industry*, 189-194. Taylor & Francis, London. https://doi.org/10.1201/9781498710411-35

[7] Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), 904-908. https://doi.org/10.1177/154193120605000909

[8] Sweetser, P., & Wyeth, P. (2005). GameFlow: A model for evaluating player enjoyment in games. *Computers in Entertainment, 3*(3), 1-24. https://doi.org/10.1145/1077246.1077253

[9] Yannakakis, G. N., & Togelius, J. (2018). *Artificial intelligence and games*. Springer, Cham. https://doi.org/10.1007/978-3-319-63519-4

[10] Buongiorno, S., Klinkert, L., Zhuang, Z., Chawla, T., & Clark, C. (2024). PANGeA: Procedural artificial narrative using generative AI for turn-based, role-playing video games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 20*(1), 156-166. https://doi.org/10.1609/aiide.v20i1.31876