

Design of Generative AI Game NPCs Based on Human-Computer Interaction

Huzhongchenwei *

Computer Science and Technology, Guizhou Normal University, Guiyang, Guizhou, 550025, China

* Corresponding author Email: 1345939442@qq.com

Abstract: The generative artificial intelligence (Generative AI) and the large language models (LLMs) are developing at a very fast pace, which has led to the fundamental change in the way of interaction with the non-player characters (NPCs) in the games. The conventional NPCs have no dynamic reaction mechanisms or long-term memory, and therefore they cannot be used to satisfy the requirements of the authenticity and immersion in the modern open-world games as it is not possible to use them effectively. In the past few years, the topic of designing an NPC based on LLMs has been discussed in detail by many researchers within the academic world including topics like personality consistency, dialogue creation, contextual adjustment, etc. Most current works however deal with the technical aspects and do not provide systematic description of how to evaluate and design using human-computer interaction (HCI). The paper presents and applies the generative AI NPC design paradigm that uses the LLMs through the lens of HCI. This can be done through the introduction of such methods as contextual design, low-fidelity. The design-evaluation-redesign process, which is based on the prototyping, think-aloud and heuristic evaluation of the possibilities of AI NPCs in the dynamic dialogue, contextual adaptation and behavior prediction, will be developed as a series of iterative steps. The system has been prototyped with an open-domain dialogue, with a React front-end and an LLM API that supports contextual memory. The mixed-methods study approach to the experimental research includes the quantitative analysis of the immersion of the players and the efficiency of their task performance and qualitative measurement of the user interaction. There are expected results of having a prototype system of an interactive AI NPC, a collection of guidelines to interact with generative AI NPCs, and a set of experimental and assessment toolkits to conduct research at the intersection of HCI-AI. This paper gives theoretical basis and practical directions to the application of the concept of generative AI to game interaction.

Keywords: Generative Artificial Intelligence; Large Language Models; Non-player Characters; Human-computer Interaction; Game Design; Iterative Evaluation.

1. Introduction

1.1. Research Background

It has been observed that the video game NPCs have been increasingly developed to become more of a background decoration in the last few years and less of an aspect of immersion into the story of the game. The role-playing games (RPGs), open-world games, and other similar types of games use not only the role of the NPCs as tools to develop the plot but also as the main instruments to build up the world of the game. The expectations of players towards the NPCs have moved beyond the delivery of functional tasks, with a focus on the emotional impact, personalization, and narrative involvement.

Nevertheless, the traditional NPC logic is mostly implemented in the form of finite state machines (FSM), behavior trees (BT) or dialogue trees. The FSMs are used to determine the movement of the NPCs by means of a given pattern of states that can be followed, e.g., patrol, detect player, attack and then return. Behavior tree has its own hierarchy of nodes that regulate the priority of the behaviors but still requires all possible combinations of behaviors to be determined by developers. Dialogue trees have very strict limitations on what players can do with them, as it only allows them to choose between pre-established branches. In spite of their merits concerning the logic and controllability of resources, these methods are not more than an engineering of the control of behavior, which does not possess any learning and adaptability. It was discovered that following the

interaction of scripted NPCs 3-5 times, the behavior of the NPCs will be predictable to the players, hence the exploration motivation and immersion would be considerably reduced [1]. According to the technical analysis provided by NVIDIA, when one plays RPGs, the number of dialogues that villager NPCs can give should never exceed three, and enemy NPCs will always have a set of cyclic attacks, which can be learned by the players within a few rounds.

In the meantime, generative AI, in particular large language models (e.g., GPT series, LLaMA) have shown strong natural language understanding and generation abilities that allow generating plausible, varied, and personality-consistent responses to a given context without using predefined scripts. This is a technological breakthrough which offers a different method of game NPC design. In 2025, Southern Methodist University reported that GPT-4 had reached 73.98 percent accuracy on preserving NPC personality consistency over more than 50,500 tests, a qualitative leap over the less than 18 percent performance of previous models [2]. The implication of this finding is that not only can large language models create dialogue, but they can also facilitate the development of characters with consistent personality characteristics, which introduces new opportunities to anthropomorphic NPC design in games.

1.2. Research Questions

This paper answers the following three fundamental research questions:

(1) What is the way to build game NPCs using large language models, which have dynamic dialogue and

contextual adaptation features? This question has several technical dimensions such as model selection, prompt engineering, and context management that attempt to investigate the technical limits of LLM-based NPCs in regard to narrative consistency, response timeliness, and personality stability.

(2) How can low-fidelity prototyping, think-aloud protocols and heuristic evaluation be used to optimize the interaction experience of AI NPCs? HCI has been widely applied in traditional software interface design, but its applicability and design-guiding value in the design of novel AI NPC systems still lack systematic empirical validation.

(3) The capability of the behavior prediction in AI NPCs and its impact on the immersion of players and their ability to complete tasks? This is a question that can be asked by critically thinking about the logic behind the behavior of the NPC - s. In general, people are critical of traditional NPCs because they are too predictable. However, could it be that highly predictable LLM-based NPCs would also bring problems? The "double-edged sword" effect has become the core motivation of this study.

1.3. Research Objectives

The general objective of this research is to develop and put into practice a generative AI NPC prototype system that relies on LLMs and test its usability and user experience by means of systematic HCI evaluation techniques. The specific objectives are:

(1) Develop a prototype of AI NPC with open conversation and contextual response, including the basic LLM integration, context management, and personality setup capabilities.

(2) Use an iterative "design-evaluate-redesign" approach to improve interaction experience, and apply HCI formative and summative evaluation methodologies in a systematic way throughout the AI NPC design process.

(3) Investigate the boundaries of applicability and design guidelines for generative AI in game NPC design, offering design knowledge for future research and industrial practice.

2. Theoretical Basis and Literature Review

2.1. Technological Change and the Constraints of Conventional Game Characters NPCs

The technological development of game AI can be seen as the deterministic behavior patterns of early arcade games. RPGs and other types of games have been evolving NPC behaviour modeling over the last four decades, from finite state machines to behavior trees and goal-oriented action planning (GOAP), each generation trying to make NPC behaviour more intelligent and natural [3].

The simplest of all behavior modeling techniques is the finite state machine. The NPC will have a limited number of states to move in and out of and these states are determined by the set of behaviors that it has. As an illustration, a guard NPC may go through a loop of IDLE PATROL CHASE ATTACK RETREAT. The benefits of FSMs can be easily implemented, easy to understand, and not much performance overhead. But their disadvantages are no less obvious: when the complexity of the behaviour increases, the quantity of states also grows exponentially, which causes the phenomenon of the so-called state explosion, and FSMs are unable to cope with those situations that were not specified in advance.

The concept of behavior trees is the use of hierarchical control systems in FSMs, which are capable of being more flexible in terms of scheduling behaviors using composite nodes (Sequence, Selector, Parallel) and decorator nodes. The most popular form of NPC behaviour management is a behavior tree and it has been used in games such as *Horizon Zero Dawn* and *The Last of Us* [3]. However, the nature of the behavior trees is to combine and schedule predefined actions; there is no way that any possible path of the behavior can be learned or adapted by developers, since all of them are designed by developers before hand.

Despite the benefits of these approaches in terms of logical clarity and resource controllability, they are still essentially behavior control engineering, without learning and adaptive features. The reflections on the industry are especially deep: in the course of creating *Watch Dogs: Legion*, Ubisoft specifically pointed out that it is impossible to realize the design objective of each character having an independent personality using traditional NPC systems and therefore procedural generation and AI-assisted behavior modeling have to be introduced. This judgment was further confirmed by subsequent generative AI prototype projects at Ubisoft - their NPC dialogue systems started trying to use LLMs to overcome the limitations of pre-written scripts.

There have also been systematic academic critiques. Lankoski and Bjork (2015) in their research on game design noted that the over predictability of scripted NPCs compromises identification with the virtual character by players, reducing immersion into narrative [4]. In *Artificial Intelligence and Games* book, Yannakakis and Togelius (2018) suggested that once a player has played a couple of rounds with scripted NPCs, they can easily anticipate their behavior patterns and this determinism is exactly what makes games immersion the most hostile to immersion [1].

Interestingly, the first industry and modding communities have already indicated that a new paradigm is emerging. In 2023, it was decided by the *Mount and Blade II: Bannerlord* to use ChatGPT in the game as a means of generating dynamic dialogue with the NPCs. The players were happy with the fact that the NPCs have finally been given a soul, and even though the technical aspect of the implementation of the AI has not yet been stabilized, the response of the market showed how promising the generative AI can be when used in NPC interaction [5].

2.2. Generative AI and Game NPCs Integration: Current Research

In recent years, academic studies have extensively investigated LLM-based NPC design through empirical studies, developing a few key areas of study.

(1) Personality Stability and the anthropomorphic form. The GPT series models were also examined in a systematic manner by Klinkert et al to determine the ability of such models to retain the personality of the NPCs. They used 50-item International Personality Item Pool that was based on the Big Five model of personality, which demonstrated that the accuracy of the personality consistency of GPT-4 is equal to 73.98 percent (more than 50,500 tests) as opposed to less than 18 percent (text-davinci-003) [2]. This step up shows that there is the basic potential to create character building with stable personalities through LLMs. It has been observed in the study that the LLMs are able to translate the background information of fictional characters into uniform dialogue patterns and behavior choices, allowing the NPCs to work

with the logic of the person rather than the script [2].

(2) The Dialogue of Dynamic Players. A group of studies was published by Rimer, Ploug and other researchers in 2025 IEEE Conference on Games which dealt with the LLM-based dialogue systems of NPCs. They developed a specific RPG game *Quest of Aivengarde* and created four different versions of dialogue: one was the control (the traditional dialogue tree), another was rephrase mode (LLM just paraphrase what is written down in the preset content), hybrid mode (LLM partially participates in generating dialogues) and open mode (LLM generates dialogues completely) [6]. The results of the experiments conducted on 64 participants indicated that the full openness of the dialogue resulted in an increased length of the dialogue and the increased interest of players in the overall process, especially in casual players [6-7]. The users also said that the experience of the immersion and naturalness of the open-ended LLM speech could not be compared to the conventional dialogue trees [6-7].

In this regard, Feng et al. (2025) systematically investigated the humanization design of LLM-based NPCs and user experience optimization in terms of HCI. The paper concentrated on three fundamental technologies - retrieval-augmented generation (RAG), knowledge graph-based question answering improvement, and reinforcement learning to optimize dialogue [8]. It was discovered that RAG can alleviate hallucination in LLMs by dynamically accessing external information, but its performance is highly sensitive to the quality and size of the knowledge base [8,9]; the knowledge graph method improves the relationship between NPC speech and game story by means of triple structures (subject-predicate-object), but it requires high construction and maintenance costs [8]; though policy convergence is faster with reinforcement learning, it remains a problem when applied to multi-step tasks [8].

(3) Multimodality and Situational Intelligence. The LLMs have been combined with computer vision in the proposal of Radez and Bohak (2026). They are able to get panoramic pictures of the surrounding environment of the NPC and do semantic segmentation and obtain objects as well as the information about their location [10]. Such a data is structured into a JSON structure, which is used as input to the LLM so that the NPCs can dynamically refer to the nearby objects, landmarks, and the other elements of the environment [10]. According to user studies, context-aware NPCs were much more popular among users, proving the success of this method to make the game realistic [10].

(4) Optimization of Efficiency and Deployment on the Edge. Braas et al. (2025) were able to solve the issue of large inference time and hardware needs in LLMs with a modular NPC dialogue system that was developed using small language models (SLMs) [11]. The proposed method is to train small models so as to be used to encode the persona of an NPC and use runtime modules which are memory-swappable, where the model does not need to be retrained or reloaded every time it is used [11]. It has been discovered that at consumer level computers, TinyLlama-1.1B-Chat can be used to achieve acceptable speed of inference, but also keep character dialogue coherence, offering a viable technical solution of running NPCs based on LLMs on commodity personal computers and consoles [11].

(5) Chinese Research. The issue of generative AI in games has also become a research topic in the Chinese academic community. The cultural studies on the production mechanisms of electronic games are one of the areas that have

been studied by Du Liang and Jia Tianxiang (2024), where they discussed the innovation introduced to the game industry with the help of generative AI, and the fact that the game industry is trying to use the power of machine intelligence to make the interaction between human and computer smarter [12]. Li Dianfeng (2024) made an extensive analysis of the future potential and difficulties of the application of intelligent NPCs to video games as well as classified them into three types of artificial intelligence: human-in-the-loop, human-factors-in-the-loop, and non-human-in-the-loop, based on the principles of affective computing and immersive experience [13]. Wu et al. (2025) carried out a systematic literature review at the technical level of the review which included 72 papers out of 6, 120 articles and gave a comprehensive account of the use of generative AI during the process of creating characters in games (Wu et al., 2025).

Although the positive development in the current studies has been achieved, there are still large gaps. First of all, most of the researches concentrate on technical implementation issues (model selection, prompt design, latency optimization, etc.) and relatively low systematic analysis of user experience [1,15]. Second, there is no standardization and empirical justification of the use of classic HCI evaluation techniques (e.g., think-aloud protocols, formative evaluation) to AI NPC design [1,15]. Lastly, the existing literature does not agree on how AI NPC behavior predictability influences immersion - traditional NPCs are criticized as being too predictable but when LLM-based NPCs are overly predictive, will they have the same effect on players' motivation to explore? This question is the entry point of this study [1,7].

2.3. NPC Design Methods from an HCI Perspective

HCI has a human-centered design philosophy, and the main approaches are contextual design, low-fidelity prototyping, formative evaluation, and summative evaluation [1,15]. These techniques can be applied to AI NPC design naturally but have not been developed into a systematic framework of application yet [15].

The contextual design is one of the traditional methods of eliciting requirements in human-computer interaction (HCI) that will help to know the pattern of behavior and needs of the users in real-life situations by observing them and conducting interviews with them [15]. The contextual design in AI NPC can be used in the development of game developers, as it helps them to understand what players would expect when interacting with NPCs in various game contexts.

Low-fidelity prototypes and storyboards are used to validate design ideas quickly and are significant in the initial iterations of AI NPC interaction flows [15]. In comparison with high-fidelity prototypes, low-fidelity prototypes (e.g. hand-drawn sketches, paper models) enable designers to find usability problems easily before investing a lot of engineering resources, which lowers iteration costs.

Think-aloud is a qualitative usability evaluation method in which participants must vocalize their thoughts continuously while performing tasks [15]. It is especially appropriate for evaluating AI NPC systems since the focus of researchers encompasses not just "what the NPC said" but also "how the player takes in and reacts to the NPC utterances" [15].

Formative evaluation focuses on the continuous gathering of feedback during the design process to inform improvements, whereas summative evaluation focuses on the end-point inspection [15]. For AI NPC systems, formative

evaluation can help researchers optimize dialogue quality thresholds, response latency tolerance, narrative coherence and other aspects iteratively [10,15].

The above are mature paradigms of the application of the traditional software interface usability testing. But in case of AI-based interactive systems (in particular, LLM-based dialogue systems), they need to be examined and applied more thoroughly in terms of their applicability, operation rules and performance measures [15]. As an example, the think-aloud of a conventional graphical interface test is based on the users click path and the comprehension of the interface, and the evaluation of the LLM-driven dialogue system is based on the new factors like semantic understanding, character identification and affective reaction. This paper will empirically address this interdisciplinary void [15].

3. Research Methods and System Design

3.1. Overall Research Framework

The current research is a mixed-methods study that integrates qualitative user research with quantitative experimental assessment [15]. The entire process can be broken down into five stages, which are associated with certain methods of HCI and evaluation goals [15].

Stage 1: Needs Analysis. Affinity diagramming is the one of the data organization tools which are combined with the use of contextual design and user interviews [15]. The initial step in the research process is to define the key requirements that should be met when designing AI NPCs based on the literature and industry cases, such as openness, consistency of personality, context awareness, and timeliness of response [2,10]. Then the interview of the players who have RPG experience (5-8) will be done in a semi-structured manner

[7,15] to learn about their current experiences in using NPC systems, their pain points and what they expect to see in an AI-driven NPC system. All these are analyzed by interpretation of all the records of the interviews and grouping of raw statements into core requirement clusters through affinity diagrams [15].

Phase 2: Prototyping. According to the results of the analysis of the requirements, it is necessary to create low-fidelity prototypes (sketches and models) and storyboards in order to demonstrate the interaction between NPCs and players [15]. The reasonableness of the main logic of interaction as well as its usability are ensured by internal checks and basic tests on the useability of the program [15], so that the process of engineering can be corrected at an early stage.

Phase 3: Technical Implementation. High-fidelity prototypes are implemented with the React frontend framework and an LLM API, finishing LLM service integration, context management modules, and frontend interface development [2,11].

Phase 4 – Evaluation & Iteration: After high-fidelity prototype development, a heuristic evaluation for formative purposes is carried out [15], engaging 3-5 experts in HCI and game design to thoroughly examine the system and pinpoint usability problems. Next, a handful of representative users are recruited to take part in think-aloud testing [15] to witness patterns of behaviour and thought whilst interacting naturally. System iterations are undertaken in response to evaluation feedback [15].

Phase 5 - Effectiveness Validation: Summative evaluation via experiment design and questionnaires, quantitatively assessing overall user experience, immersion, and task completion efficacy of the AI NPC system [1,7,15]. The relations between the three phases of research are depicted in Figure 1.

Table 1. Research stage and methodology mapping

Phase	Methods	Matching HCI Content	Outputs
Requirements Analysis	Contextual design, user interviews, affinity diagrams	contextual design, interpretation session	User Requirements Report
Prototype Design	Low-fidelity sketches + storyboards	Sketching exercises, storyboard logic.	Low-fidelity prototype set.
Technical Implementation	LLM API + React frontend	frontend framework, API integration	high-fidelity prototype system
Evaluation and Iteration	Think-aloud + heuristic evaluation	Think-aloud, formative evaluation	Iterative improvement checklist
Effectiveness Validation	Experimental design + questionnaires	summative evaluation, questionnaire design	Experimental data analysis report

3.2. System Design and Technical Implementation

3.2.1. System Architecture

This is the overall structure of this system which is comprised of three major layers:

(1) Frontend Layer of the Interface: The React framework has been constructed and it offers the main interface elements, which are a text input field, NPC response display panel, history of dialogue visualization, and scene status notifications. In order to facilitate data collection over time and its analysis in the future, there is a system of interaction logs that consists of dialogue content, time stamping, as well as delays of the responses of NPCs [15].

(2) LLM Service Layer: Cloud-based LLM services are accessed through RESTful APIs. The first prototype is

planned to use the GPT-4 model, with the architecture reserved for model switching [2,11] so that comparative experiments can be conducted between different LLMs/SLMs in the future. Prompt engineering is one of the main jobs of this layer, which needs to organize context information such as NPC personality settings, current scene information, dialogue history, etc. effectively as input to the LLM [2,8].

(3) Context Management Module: Keeps the conversation of the player and the character personality document of the NPC, as well as tagged information on the current scene of the game [2,11]. The module has a direct influence on the quality of the response of LLMs generated by them [2, 11] - the absence of proper management of context will result in the inability of the NPC to recall what was said before in the conversation and will result in an obvious fragmentation of the persona.

Table 2. System technology stack and functions

Technology Component	Specifications/Framework	Function Description
Frontend Framework	React 18+	UI components and state management
Frontend Language	HTML5/CSS3/JavaScript	Interface and style development
LLM Service	GPT-4 API	Dialogue generation engine
Context Storage	Local database / cloud storage	dialogue history and scene state
Persona Configuration	System prompts	Character consistency control
Auxiliary Evaluation Tools	Logging module	Data collection of experiments.

3.2.2. The Behavior Prediction Ability Design and Comparison System

To systematically investigate the effect of behavior prediction capability on player experience, the system creates two kinds of NPC behavior modes:

Low-prediction mode (baseline condition): The NPC makes responses only on the basis of the current input of the player without taking into account the behavior patterns of the player in the medium- and long-term, as well as progress of the task. This mode is intended to emulate the interaction pattern of conventional scripted NPCs [1], which is used as a reference point.

High-prediction mode (experimental condition): The NPC uses the history of dialogue, tasks completed, unfinished tasks, and information about the current scene context of the player to actively suggest or hint at the current scene and state of the player in responses [7,10]. Indicatively, when a player has not succeeded several times in a puzzle-solving task, the high-prediction mode of the NPC will be proactive in providing clues instead of waiting for the player to request.

The two types of NPCs have the same personality, style and linguistic capability in terms of the tone, but the only difference is that they are not able to use long term context to predict their behaviour [2,11]. It has been designed so that the measured variations will be associated with the independent variable of the predictability itself[15] as opposed to any other confounding variables.

3.2.3. Personality Control and Consistency Mechanism

LLM-based NPCs are highly flexible in generating responses, yet this flexibility brings the risk of personality drift: without constraints, the same NPC may exhibit different personality traits when answering different questions, which will seriously affect players' identification with their characters [2]. To solve this problem, this paper proposes a multi-level personality control mechanism:

(1)System prompt level: Each API request has a character definition document which defines the background of the character, personality traits (based on the Big Five personality dimensions), speaking style and emotional tone [2].

(2)Dynamic binding level: The setting of the personality is linked to the identity number of the target NPC and it means that there are different NPCs in the game with independent and stable character traits [2,11].

(3)The boundary level of setting: The safe ways of answering the questions that are put by players can be made in such a way as to prevent failure of alignment (e.g. the NPC does not have enough information on real - world matters).

3.3. Prototype Iteration Process

The principle of progressive refinement in the HCI is used to develop prototypes, which is split into three stages [15]:

(1) Low-fidelity prototype: To begin with, paper-and-pencil sketches of the fundamental layout of the NPC dialogue interface are drawn and interaction storyboards

showing the entire flow of interaction between player input and NPC response are also developed. The internal team reviews are preliminary to confirm the reasonableness of the logic behind the interaction and the smoothness of the user journey [15].

(2) Prototype of medium-fidelity: The prototype is a clickable version of the interface interaction and branching response logic that is implemented in interaction design software, including Figma, which represents the dialogue between NPCs [15]. It is not possible to use the real LLM with the prototype of medium fidelity but it can be used to test the usability of the user.

(3) The prototype is a high fidelity: Lastly, the system is implemented on the basis of the React framework and an API of LLM. The system provides features of context management, personality regulation, and recording dialogues [2,11,15] in real users to conduct interactive experiments.

4. Experimental Setup and Initial Results

4.1. Experimental Objectives and Hypotheses

The paper is supposed to confirm the effects of AI NPCs in various modes of behavior prediction on the immersion of players, their efficiency in performing tasks, and the satisfaction with the game [1,7]. The hypotheses that can be made are as follows: according to the research gaps found in the literature review and the questions proposed by the researcher.

H1: The efficiency of the task performed by players in information gathering and guided tasks are higher in high prediction mode than low prediction mode [7].

H2: The immersion of players in high-prediction mode is much greater than that of low-prediction mode, but this difference is moderated by task type - too much prediction may actually lower autonomy in exploration-driven scenarios [1,7].

H3: The AI NPC response latency is one of the major pain points that can influence player experience, and there is a variance in user patience with the LLM generation speed [10,11].

Testing of the above hypotheses will yield empirical evidence pertaining to the main research question of this study - how the behaviour prediction ability of AI NPCs influences player immersion and task completion efficacy.

4.2. Experimental Design

Independent variable: NPC behavior mode (low-prediction vs. high-prediction), using a between-subjects design [15]. To eliminate learning effects and fatigue effects on the results, participants are randomly allocated to one of two groups.

Dependent variables:

(1) Immersion: It was measured with the validated Game Experience Questionnaire (GEQ), and its sub-dimensions

(e.g., immersion, flow, challenge) were adjusted in accordance with the research requirements. The key outcome variable of this study is immersion.

(2) Task completion time: The duration of three experimental tasks performed by the participants is recorded to determine task efficiency.

(3) System usability: The most popular quantitative usability evaluation tool in HCI is the System Usability Scale (SUS) [1,7,15], which measures system usability and gives a composite usability score of 0 to 100.

Experimental tasks: The participants are required to perform three experimental tasks in a small open-world scenario, namely information gathering, item seeking, and dialogue guidance [6,7]. In the information gathering task, players need to gain background information about the scenario by speaking with a certain NPC; in the item seeking task, players need to collect clues by conversing with various NPCs until they find a target item; in the dialogue guidance task, the quality of responses and interaction experience of NPCs is assessed in natural, open conversation scenarios.

4.3. Participants and Procedure

The study will recruit 32–40 participants with video game experience via online and offline means. The following are the criteria that the participants must meet: (1) they have at least 20 hours of experience in RPG or open-world games within the last six months; (2) they are between 18 and 40 years old; and (3) they can communicate fluently in the language supported by the system. Participants are randomly assigned to two groups [15], each consisting of 16–20 participants.

The experimental procedure is as follows: (1) Welcome and informed consent (5 minutes); (2) Scenario introduction and operation instructions (5 minutes); (3) Game interaction tasks (20 minutes); (4) Questionnaire completion (10 minutes); (5) Semi-structured interview (10–15 minutes). The total time taken is about 50 to 55 minutes [15].

In the experiment, research assistants will observe and document the interaction behaviors of participants with a particular emphasis on the types of reactions that the participants have towards the responses of NPCs (e.g., follow-up questions, topic changes, ignoring suggestions, context switching). In the case of the think-aloud condition, the speech of the participants is audio-recorded and transcribed after the experiment to be analyzed qualitatively.

4.4. Preliminary Findings

According to the initial results of the pilot experiment and heuristic evaluation, there are a number of significant discoveries:

Finding 1: The tension effect between efficiency and autonomy. In the pilot study observations (N=8), the high - prediction mode showed a significant superiority over the low - prediction mode in information gathering and task guidance efficiency, with participants generally perceiving the NPC as more intelligent. However, when the density of predictions went up (e.g., the NPC often and proactively gave hints in exploratory situations), some participants' willingness to explore on their own noticeably dropped [1,7]. Some even said that it felt like the NPC was leading them by the nose and didn't want them to think independently. This finding supports the predictions of H1 and H2 — predictive capability is indeed a two - sided sword, and too much prediction may erode players' sense of self - determination. This

discovery provides an interesting addition to Ploug et al. & s (2025) conclusion that open - ended dialogue enhances player engagement. Although open - ended dialogue increases engagement, the boundary between "help" and "intervention" needs careful definition as interaction depth rises.

Finding 2: Memory can be a positive emotion. The result of the think-aloud tests was that NPCs were good in long-term memory (i.e., when they remember what the player said or asked during the previous dialogue), and it had been found to have an overall positive emotional impact on participants, with many comments like very thoughtful and quite realistic. This discovery has demonstrated how important it is to build up credibility of NPC by using the concept of memory, as Braas et al. (2025) argue about the importance of long-term memory in the quality of interaction between the NPC and the player. It is interesting to note that in case NPCs remembered too much or even recalled memories at times, other people also felt watched [11] and this shows that there is a need to moderate the use of memory.

Finding 3: Latency as the key factor. It was determined that latency in LLM is a bottleneck to fluency of interaction [10-11,16]. During the pilot study, it was observed that the longer time response (more than three seconds) was, the more varied the behavior of the participants was, the eyes were wandering and the face was impatient, and the players would often type the same command again or they would believe that the system had crashed. The result is a direct reflection of the real-time problem in the user research by Radez and Bohak (2026) [10] and Braas et al. (2026) recommending the use of SLMs as an optimization tool [11]. These trade-offs on efficiency have been found to be the most critical in terms of optimizing Japanese LLMs (average inference time 3.36 seconds) where the similar results are obtained [16]. Thus, during the next experiments, which will be conducted on a large scale, the variable of response latency will be confounding, which should be tightly controlled.

Finding 4: The problem of the consistency in personality. Although, in the low-prediction control group, there was a prompt system which had provided the character persona settings and the participants were still unable to identify the differences between the characters in NPCs after they had interacted with them briefly, this indicates that even though the design of the personality is prompted-based, it may not be enough to satisfy the industry standards of the distinct character personalities (the difference in personality between NPCs) should be more intense than a simple LLM output management. This result gives an insight into the future implementation of technologies based on further binding of personas like fine-tuning or RAG augmentation (as found in the studies conducted by Feng et al. [2,8]).

All of the above results are exploratory results based on the pilot experiment and heuristic evaluation; statistical significance will be tested across the full sample data from the main experiment [2,8].

5. Discussion

5.1. Theoretical and Practical Importance of Research Contributions

The main theoretical results of the present study are as follows:

First, this paper systematically incorporates the traditional HCI evaluation techniques (contextual design, low-fidelity

prototyping, think-aloud, heuristic evaluation) into the LLM-based AI NPC design process, developing an iterative framework of design-evaluation-redesign [1,15]. This is a particular answer to the Yannakakis and Togelius (2018) appeal to make AI game design more player-centered [1]. In contrast to most studies that focus on technology, the present study highlights that AI NPC design should not only be concerned with the question of whether it can be done but also whether it contributes to improving the immersion and autonomy of players. The cross-application of the methods of HCI in this paper confirms the changes required to formative evaluation methods like think-aloud when applied to evaluate LLM dialogue systems (e.g., extending evaluation dimensions beyond click paths to semantic understanding and character identification), which offers empirical support to the evolution of HCI methodology in the context of new AI interaction paradigms [15].

Second, the research has been conducted on the basis of the low-predictive and high-predictive behavior mode of the NPC as it was demonstrated by the comparative experiment. It is shown that there are no absolute values in predictability; the greater the predictability, the better it will be (the prediction process is helpful to information-gathering tasks) or the less the predictability, the worse it will be (the prediction process is harmful to exploration-based tasks). This conclusion can provide a base of empirical analysis of further studies on the adaptive correspondence of the task type and intensity of predictions made by AI NPCs, and this can be viewed as an extrapolation of the conventional approach to the problem of the trap of predictability of NPCs: in the context of LLMs, it may not always be more beneficial to have smart than stupid NPCs; instead, the role of positioning and objectives of the scenario [1,7].

Third, the paper synthesizes an initial list of interaction design principles to generative AI NPCs [2,6-8,11]:

The principle of just right should be applied to memory, that is, contextual coherence in NPC dialogue without feeling watched or narrative interference due to excessive memory, with proper forgetting processes.

It is necessary to provide prediction mechanisms to be dynamic according to the type of task, that is, in goal-oriented situations it should be enough to predict to improve performance and in exploration-related activities it must be less proactive to keep the players free.

Usability (e.g. the delay in response) must be regarded as a primary design consideration, it is not an issue of technical implementation but that of usability, which can change dramatically after 3 seconds of delay.

The personality settings that are unique are the soul of a character - system-level prompts might not be enough to make sharp distinctions between different characters, and more in-depth persona-binding methods like fine-tuning or RAG may be needed.

Although these guidelines are the product of initial investigation, they offer a reference starting point for further research and industrial practice.

5.2. Reflections on the Adaptability of Classic HCI Methods to New AI Interaction Paradigms

In using these classic HCI methods in this study, we gained some reflections on a number of deeper issues:

Modifications of the think-aloud approach in LLM dialogue systems. The conventional think-aloud in usability

testing involves participants performing standardized task sequences on graphical interfaces with verbalizing their interface-related paths. Nevertheless, when it comes to open-ended dialogue interactions based on LLMs, the thought process of the participants includes not only what the NPC said but also several aspects like my perception of the emotions of the NPC, the difference between the intention of the NPC and my own purpose, and my role-playing in the interaction with the NPC. Conventional think-aloud instructions tend to be too generic; assessing AI NPCs can demand more specific guidance strategies.

The heuristic measures of evaluation. The classical method of heuristic analysis, which was suggested by Nielsen, aims at the overall usability of graphical interfaces. Evaluation of semantic quality of content of dialogue is also needed in LLM-based NPC systems as well as interface. As an example, does the NPC have to deal with referential expressions (coreference resolution failures) properly? Is it logical that the words spoken by the NPC are appropriate in the situation? Does the character and tone of the NPC suit the status of the player character? These aspects of the evaluation are not yet addressed in the available checklists of HCI heuristics. This implies that there may be a new area of research: developing heuristics of AI dialogue system [15].

Role positioning problems in formative assessment. The conventional formative evaluation is usually performed by professional user experience researchers or target users. In LLM dialogue systems, however, the capability to evaluate the difference of NPC personalities and semantic coherence demands high language sensitivity of evaluators. Non-expert users can have difficulties in recognizing subtle differences in personality or detecting minor deviations in semantics. This scenario shows that there is an opportunity to integrate a dual coding architecture of HCI experts and domain experts in the formative analysis of AI interactive systems [15].

5.3. Research Limitations and Future Directions

The research has its shortcomings which can be considered in the future:

From a technical perspective, the current system only supports text-based interaction, without non-verbal feedback channels such as facial expression, body gesture, and voice. Multimodal expressivity of NPCs has a strong effect on immersion in real games. The lack of non-verbal channels could have influenced the overall assessment of NPC interaction quality in the current results.

As to latency optimization: The response time of the LLM (currently averaging 24 seconds) is a factor in the real-time performance of interactions, especially when there is a need to respond quickly to feedback - this delay may be an impediment. It was also found that latency leads to discomfort and disruption of flow among players as observed by pilots. Future research will take into account the following strategies to enhance latency: using smaller SLM models to deploy at the edge to minimize the number of network round trips; employing prompt caching and request batching techniques to minimize overheads of individual API calls.

In terms of the size and applicability of the experiment: The proposed number of participants in the full scale research is 32-40 (the pilot study on the directionality has been conducted, N=8). Despite the fact that the sample size is enough to conduct an initial statistical analysis, the findings are not generalizable. Also, the experience of gaming of the

players will be mainly based on RPG and open-world games, therefore, it is hard to extrapolate these results to other types of games (e.g., strategy games, MMOs). Moreover, there may be a great difference in how players would accept the behavior of NPC as being predictive, i.e. the player in collectivist cultures would prefer more aggressive assistant-like characters and the player in individualistic cultures would prefer the same level of conversation with NPCs. It remains to be seen in future comparative studies.

In terms of the maturity of design guidelines: The design guidelines suggested in this study (e.g., "just right" memory use, matching prediction intensity to task type, prioritizing latency constraints for usability) are still at an initial refinement stage. These guidelines need systematic validation by means of quantitative testing in a large - scale main experiment and the addition of more scenarios (e.g. , simultaneous multi - player interaction, voice - based natural language input) to enhance ecological validity further.

As to evaluation systems: The current LLM-driven NPC system evaluations are still based on questionnaires and semi-structured interviews, without adequate data collection of real player natural behavior. The next step is to develop multimodal behavior tracking technologies (e.g., eye tracking, facial expression coding) or implicit indicator modeling according to interaction sequences (e.g., frequency of user correction behaviors in dialogue, frequency of topic switching) to measure the seamless integration of the human-collaborator mode between players and LLMs [10-13,15].

6. Conclusion and Outlook

6.1. Research Conclusion

In the HCI perspective, this paper suggested and developed a generative AI NPC design framework that is founded on large language models. With the addition of traditional HCI techniques like contextual design, low-fidelity prototyping, think-aloud, and heuristic evaluation, a full iterative cycle of requirements analysis - prototype design - technical implementation - evaluation and iteration - effectiveness validation was built. The research pre-approves that:

(1) The NPCs created by LLM are much more interactive and flexible than the NPCs created by conventional scripts in open dialogue and contextual interaction [1,2] with respect to personality consistency of interaction (the level of control that is shallow based on prompt) and dynamic response.

(2) The behavior prediction feature is a two-sided sword - the moderate level of prediction has a positive impact on the efficiency of completing tasks, but excessive prediction can reduce the autonomy and exploratory motivation of players. This result gives an initial empirical answer to the academic and industrial discussion about how smart should AI NPCs be [1,7].

(3) HCI methodologies are highly applicable and valuable for guidance throughout the AI NPC design process, and the HCI focus on iterative, participatory design processes is well-suited to development of LLM-driven NPC systems [15].

(4) Latency is a basic usability constraint; after the threshold of 3 seconds, the response time, user experience drops significantly and other interaction qualities cannot offset this experiential violation [10,11].

6.2. Main Contributions

This study is one of the first empirical works to systematically integrate HCI formative and summative

evaluation throughout the whole generative AI NPC design process, giving a methodological example for the intersection of HCI and AI games [1,15].

In practice, the study introduced a prototype system of LLM-based generative AI NPCs and defined initial guidelines and evaluation frameworks for interaction design of AI NPCs, including fundamental dimensions like memory management, prediction adjustment, real-time constraints, and differentiated personas [2,6-8,11].

In terms of openness and reproducibility, the technical roadmap, experimental paradigm and evaluation toolchain of this project offer a reference point and experimental architecture to researchers in the HCI-AI intersection [14,15].

6.3. Outlook

The further development of the technical approach will include the multimodal generation, which is a speech synthesis and binding of facial expressions, to increase the number of channels of non-verbal communication of NPCs, as well as to examine the possibilities of human-computer interaction in a more natural way. The gradual transition towards cloud based system to edge optimization will be implemented: the compression techniques (model quantization, Int8/Int4) will allow the reduction of latency levels on consumer-grade average hardware and generative NPC can be deployed on such devices in real-life game development settings. It will be discussed about multi-NPC cooperation dialogue and modeling group intelligence behavior, not one-to-one machine but co-evolution and cross-talk between the NPCs, which would make the ecological reality of the NPCs more realistic [10-11,14].

In the path of evaluation, a more detailed AI NPC behavior database will be created that includes not only an AI interaction behavior but also the data on dialogue logs and task efficiency measures, emotional ratings, correction behaviors of users in order to conduct further analysis across different disciplines. The introduction of AI NPCs will be based on laboratory prototypes as well as their real-life application within the development pipeline of games to test the ecological validity and practicality of the design principles in practice [14, 15].

In terms of the theoretical aspect, on the basis of the further analysis of the concept of predictability, the range of application of the capabilities of intelligent NPC in the form of assistance to the accomplishment of different types of tasks will be discussed. The prediction dimension is going to be divided into smaller parts - the simple binary design of experiment and a more complex indicator system that will be able to provide the soft design of personal game intelligent agents [1,7].

References

- [1] Yannakakis, G. N., & Togelius, J. (2018). Artificial intelligence and games. Springer. <https://doi.org/10.1007/978-3-319-63519-4>.
- [2] Klinkert, J., Clark, C., & Buongiorno, S. (2025). Personality-consistent NPCs using large language models. arXiv preprint, arXiv:2402.14879. <https://doi.org/10.48550/arXiv.2402.14879>.
- [3] Rabin, S. (Ed.). (2019). Game AI Pro 360: Gather the wisdom of the world's best AI programmers. CRC Press.
- [4] Lankoski, P., & Björk, S. (Eds.). (2015). Game design research: An introduction to theory & practice. ETC Press, Carnegie Mellon University.

- [5] Yang, D., et al. (2024). GPT for games: An updated scoping review (2020–2024). arXiv preprint, arXiv:2411.00308. <https://doi.org/10.48550/arXiv.2411.00308>.
- [6] Rimer, E., Ploug, R., Petersen, A. K. S., & Scirea, M. (2025). Talking to NPCs: Three LLM-driven approaches to dynamic RPG dialogue. In 2025 IEEE Conference on Games (CoG). IEEE Press. <https://doi.org/10.1109/CoG64752.2025.11114413>.
- [7] Ploug, R., Rimer, E., Skov Petersen, A. K., & Scirea, M. (2025). Open-ended NPC dialogue favors casual players: A pilot comparison of three LLM-driven dialogue systems. In Proceedings of the IEEE 2025 Conference on Games (CoG). IEEE Press. <https://doi.org/10.1109/CoG64752.2025.11114150>.
- [8] Feng, Y., Miao, Y., & Zhou, Y. (2025). Research on the humanization design of game NPCs and user experience optimization based on large language models. In Proceedings of the 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence (ICIAAI 2025) (pp. 789–796). Atlantis Press. https://doi.org/10.2991/978-94-6463-823-3_94.
- [9] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv preprint, arXiv:2312.10997. <https://doi.org/10.48550/arXiv.2312.10997>.
- [10] Radež, G., & Bohak, C. (2026). Empowering NPC dialogue with environmental context using LLMs and panoramic images. arXiv preprint, arXiv:2604.19192. <https://doi.org/10.48550/arXiv.2604.19192>.
- [11] Braas, M., & Esterle, L. (2025). Fixed-persona SLMs with modular memory: Scalable NPC dialogue on consumer hardware. arXiv preprint, arXiv:2511.10277. <https://doi.org/10.48550/arXiv.2511.10277>.
- [12] Du, L., & Jia, T. X. (2024). Generative artificial intelligence and the innovation of video game production mechanisms. *Art Research*, (2), 58–66.
- [13] Li, D. F. (2024). Digital maze and wax wings — Prospects and challenges of generative intelligent NPCs in video games. *Journal of Beijing Film Academy*, (12), 39–48.
- [14] Wu, Z., Chen, Z., Zhu, D., Mousas, C., & Kao, D. (2025). A systematic review of generative AI on game character creation: Applications, challenges, and future trends. *IEEE Transactions on Games*. <https://doi.org/10.1109/TG.2025.3564869>.
- [15] Ternar, D. A., Denisova, A., Cunha, J. M., Kultima, A., & Guckelsberger, C. (2025). Generative AI in game development: A qualitative research synthesis. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3613905.3642142>.
- [16] Shido, T. (2025). Examination of model efficiency and its application in the game field for Japanese large-scale language models. In Proceedings of the 23rd International Conference of Asia Digital Art and Design Association. Tokyo Metropolitan University.