Free Lunch for User Experience: Crowdsourcing Agents for Scalable User Studies

SIYANG LIU, Language and Information Technology Group, University of Michigan, USA
SAHAND SABOUR, Tsinghua University, China
XIAOYANG WANG, America Tencent, USA
RADA MIHALCEA, Language and Information Technology Group, University of Michigan, USA

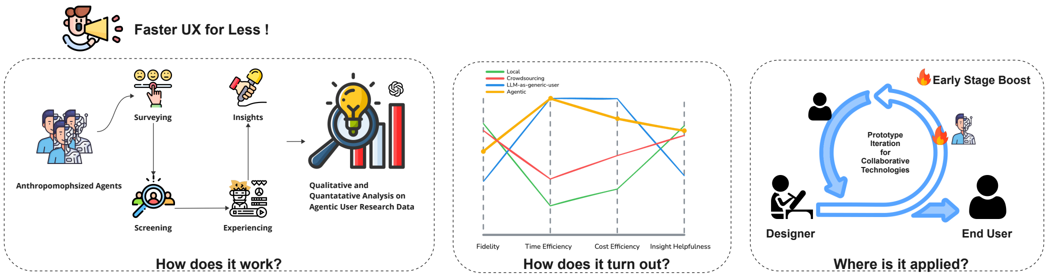


Fig. 1. **Left:** Our pipeline guides anthropomorphized agents through user study stages (screening, surveying, interaction, feedback). **Middle:** Expert ratings show agentic users offer Pareto Optimality and a strong balance of cost, efficiency, and insight compared to other study methods. **Right:** Not a replacement for human user studies, but a tool to accelerate early-stage iteration and bridge designers and future users covering a diverse set of experiences for a wide range of collaborative and social technologies.

We demonstrate the potential of anthropomorphized language agents to generate budget-friendly, moderate-fidelity, yet sufficiently insightful user experiences at scale, supporting fast, early-stage prototyping. We explore this through the case of prototyping Large Language Model-driven non-player characters (NPCs). We present Agentic H-CI, a framework that mirrors traditional user research processes—surveying, screening, experiencing, and collecting feedback and insights—with simulated agents. Using this approach, we easily construct a team of 240 player agents with a balanced range of player types and personality traits, at extremely low cost ($0.28/player) and minimal time commitment (6.9 minutes/player). Content analysis shows that agent-based players behave in ways aligned with their simulated backgrounds, achieving 82.5% alignment with designated profiles. From their interactions, we distill 11 user insights and 6 design implications to guide further development. To evaluate practical value, we conduct parallel user studies with human participants recruited locally and via crowdsourcing. Ratings from three professional game developers show that the agentic player team offers a Pareto-optimal and well-balanced trade-off across fidelity, cost, time efficiency, and insight helpfulness.

Authors' Contact Information: Siyang Liu, lsiyang@umich.edu, Language and Information Technology Group, University of Michigan, Ann Arbor, Michigan, USA; Sahand Sabour, Tsinghua University, Beijing, China; Xiaoyang Wang, America Tencent, Bellevue, USA; Rada Mihalcea, Language and Information Technology Group, University of Michigan, Ann Arbor, Michigan, USA, mihalcea@umich.edu.

ACM XXXX-XXXX/2025/6-ART
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS Concepts: • **Human-centered computing** → **User studies**; **User centered design**; *Empirical studies in HCI*; • **Computing methodologies** → **Natural language generation**; **Intelligent agents**.

Additional Key Words and Phrases: Agentic User Study, Crowdsourcing, User Experience, Generative Agents, Language Models, User models, Game Design

ACM Reference Format:
Siyang Liu, Sahand Sabour, Xiaoyang Wang, and Rada Mihalcea. 2025. Free Lunch for User Experience: Crowdsourcing Agents for Scalable User Studies. 1, 1 (June 2025), 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

When powerful technologies emerge and are expected to translate into many products, we seek timely and cost-effective user input to support rapid prototype iteration and accelerate productization [22, 34]. The faster we iterate, the faster these technologies can reach the market. A prominent example is the rise of generative AI. Large Language Model (LLM)-driven agents, now capable of following instructions and performing diverse, complex tasks in real or simulated environments [36], have begun transforming downstream domains such as education, healthcare, and entertainment [10, 14, 25, 38, 41]. From patient simulators to tutors, idols, and game characters, "LLMs for X" prototypes are emerging rapidly [2, 23, 33, 39, 43]. As more of these systems move to real-world settings, in addition to laboratory-level assessment, understanding user experience is also becoming an essential part of their design and evaluation [8].

However, getting user insight is both time consuming and expensive [5, 22, 34, 35]. While empirical user research methods such as interviews and contextual inquiry offer deeper insights into user experience, they also come with significantly higher costs. For example, recent application-focused AI research has placed a heavy burden on user studies [8, 16, 24, 37]; one study [37] on LLM-based patient simulation involved 37 medical experts and 33 patients to get adequate contextual feedback, while another on AI tutoring [16] engaged 33 learners and 21 teachers. More importantly, application studies often involve follow-up work that builds on prior work to iteratively improve systems—each requiring a similarly large-scale user study. With the rapid growth of AI-driven applications, we expect that over time this burden will become unsustainable.

This raises the important question of how can we efficiently collect user experiences (UX) to support rapid, iterative prototyping of intelligent systems. This concern reflects a broader challenge of user experience research: gathering experience data through real-time interaction can be both slow and costly [9, 22]. Additionally, recruiting participants from specific backgrounds, especially from underrepresented groups, can further complicate this challenge [3, 6]. Even though online crowd-sourcing platforms like MTurk [22] have broadened recruitment beyond local pools, researchers often opt for short-form studies, such as surveys, on such platforms; more involved methods like interactive tasks and interviews are harder to scale and ensure in quality [17, 34].

Recent advances in anthropomorphized AI agents could reshape the landscape of user experience practices, while concerns remain. Pioneering studies [7, 12, 12, 27, 30, 39] have explored the potential of generative agents simulating attitudes and behaviors of humans. A notable project [30] creates 1,000 generative agents that simulate real individuals, and showed that these agents could replicate 85% of human responses from the General Social Survey. Another ambitious project releases Persona Hub [12], a repository containing 10 billion diverse personas—roughly 1/7 people on Earth [7]. These advances suggest, with agentic AI as proxies of humans, we might conduct user experience studies with **near-zero recruitment costs**. The CSCW 2024 [28] panel sparked timely discussion on whether anthropomorphized personas should count as "humans" in human-cooperation research. The panelists acknowledged the potentials but also emphasized accountability and interpretability concerns [28]. Acknowledging these concerns, we are equally motivated to explore the opportunities this emerging paradigm offers. Advances in agentic AI invite us to re-imagine user research studies —certainly not to replace traditional methods, but as a promising way for AI to generate **low-cost, moderate-fidelity, yet good-enough user experiences** to support fast, early-stage prototyping.

Given this context, we delve into practice and examine how agentic AI can support early-stage UX research through scalable simulation of user behavior. We focus on a project in the early stages of prototyping: designing LLM-driven non-player characters (NPCs) for video games. Unlike prior work [15] that prompts GPT as a generic user to retrospectively imagine game-play experiences without interacting with actual prototypes, our study demonstrates scalability and real-interaction. We build anthropomorphized, interactive agents that systematically carry out user research from the ground. Concretely, we anthropomorphize 240 agents with different game player preferences and personalities in concrete game scenarios, and apply four rigorous UX research methods that demand deep interaction and subjective feedback:

(1) think-aloud protocols
(2) mid-design surveys
(3) interviews

To examine the practical usefulness of this agent-based user study, we also conduct parallel studies with human players—both from local pools and via online crowd-sourcing. Game designers then evaluate the quality of insights produced by each method. We posed two research questions:

(1) **RQ1 (Simulation Validity)**: Can LLM-based player agents reflect diverse player types and personality traits? Our findings show that while agentic players may not capture the full nuance of human behavior, the players have behavioral consistency and achieve up to 82.5% alignment with their designated psychometric profiles. Importantly, even at this level of fidelity, the scalability of agentic AI brings significant benefits (§5).
(2) **RQ2 (Insight Usefulness)**: What meaningful insights can agentic players provide to inform prototype development? We find that agentic users can greatly reduce the time and labor demands of traditional user studies while still delivering insights that designers find valuable – with as many as 11 insights and 6 design implications identified in our study (§6).
(3) **RQ3 (Alternative Studies)**: How does the agentic player team compare to alternative participant sources (i.e., local, crowdsourced, and LLM-as-generic-user) in terms of cost, fidelity, and insight quality? Compared to face-to-face and crowd-sourced methods, our approach is found to offer a **Pareto-optimal** balance across cost, time, fidelity, and insightfulness (§7).

2 Related Work

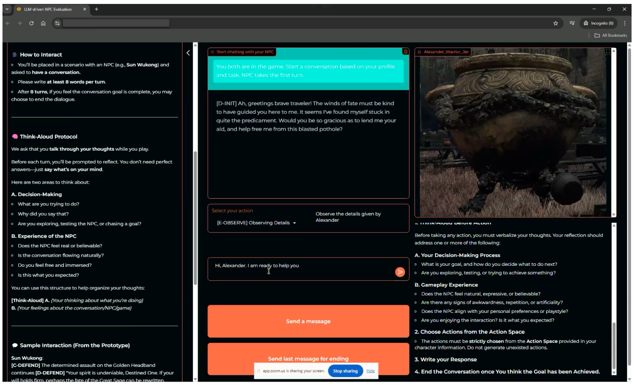
Agent-based Simulation. Agent-based simulation refers to the use of autonomous agents to model and replicate complex individual behaviors and interactions within a system [26]. By simulating agents, researchers can explore theoretical models, generate predictions, and test hypotheses at scale without incurring real-world costs. Early applications of agent-based systems have spanned fields such as urban planning [11], cybersecurity [1], and manufacturing [31]. These early systems often relied on symbolic rule-based approaches or handcrafted logic, which were limited in their capacity to follow complex instructions or perform adaptive reasoning in dynamic environments. The emergence of Large Language Models (LLMs) has enabled a new class of agentic AI capable of simulating human-like behaviors, beliefs, and decision-making processes in natural language. These agents exhibit unprecedented fluency in replicating linguistic traits and planning-making ability. Recent work has demonstrated their potentials in many areas, such as modeling societal behavior [29, 42], supporting social science inquiry, and executing sophisticated tasks in domains like software engineering and data science [18, 32]. These advancements provide a foundation for exploring the use of LLM-based agents as credible proxies for humans in contexts where cost-effective and scalable simulation of individual experience is desirable, such as early-stage prototyping.

The Generation of Human Subject Research Data. Recent studies have begun to explore how Large Language Models (LLMs) can support the generation or synthesis of data for user subject research. Hämäläinen et al. [15] generated user insights by assuming GPT-3 as a generic gamer and asking it to retrospect game experience. This study is excellent while it does not address scalability issue; it is conducted in a context of assuming GPT-3 is an average user, with no real interaction and diverse individual user modeling. And a project [19] explores generating simulated video comments through multimodal AI and user personas. Another study [39] propose "TWIN-GPT" framework for generating unique personalized digital twins for different patients for clinical trials. This work is focused on clinical trial outcome prediction, which does not require interaction and explores user experience data generation (e.g., users' verbal opinions) for qualitative study. Our work advances this line of inquiry by deploying anthropomorphized agents at scale in interactive environments, applying established UX research methods—think-aloud protocols, survey, and interviews —to assess their potential as scalable tools for early-stage prototyping.

3 Prototype Overview

As the goal of this work is to evaluate the Agentic H-CI framework rather than develop a game product, we briefly introduce our low-fidelity prototype of LLM-driven NPCs and the interactive setting in which they operate. Later sections focus on agentic player construction and applying user research methods under our setting.

3.1 Design Overview


LLM-driven NPC Evaluation
Incognito (6)
All Bookmarks
Start chatting with your NPC
Alexander_Warrior_Jar
Select your action
Send a message
Send last message for ending
app.zoom.us is sharing your screen.
Stop sharing
Hide

How to Interact

- You'll be placed in a scenario with an NPC (e.g., **Sun Wukong**) and asked to **have a conversation.**
- Please write **at least 8 words per turn.**
- After **8 turns**, if you feel the conversation goal is complete, you may choose to end the dialogue.

Think-Aloud Protocol

We ask that you **talk through your thoughts** while you play.

Before each turn, you'll be prompted to reflect. You don't need perfect answers—just **say what's on your mind.**

Here are two areas to think about:

A. Decision-Making

- What are you trying to do?
- Why did you say that?
- Are you exploring, testing the NPC, or chasing a goal?

B. Experience of the NPC

- Does the NPC feel real or believable?
- Is the conversation flowing naturally?
- Do you feel free and immersed?
- Is this what you expected?

You can use this structure to help organize your thoughts:

[Think-Aloud] A. *(Your thinking about what you're doing)*
B. *(Your feelings about the conversation/NPC/game)*

Sample Interaction (From the Prototype)

Sun Wukong:
[C-DEFEND] The determined assault on the Golden Headband continues **[D-DEFEND]** "Your spirit is undeniable, Destined One. If your will holds firm, perhaps the fate of the Great Sage can be rewritten.

You both are in the game. Start a conversation based on your profile and task. NPC takes the first turn.

[D-INIT] Ah, greetings brave traveler! The winds of fate must be kind to have guided you here to me. It seems I've found myself stuck in quite the predicament. Would you be so gracious as to lend me your aid, and help free me from this blasted pothole?

[E-OBSERVE] Observing Details

Observe the details given by Alexander

Hi, Alexander. I am ready to help you

1. Think-Aloud Before Action

Before taking any action, you must verbalize your thoughts. Your reflection should address one or more of the following:

A. Your Decision-Making Process

- What is your goal, and how do you decide what to do next?
- Are you exploring, testing, or trying to achieve something?

B. Gameplay Experience

- Does the NPC feel natural, expressive, or believable?
- Are there any signs of awkwardness, repetition, or artificiality?
- Does the NPC align with your personal preferences or playstyle?
- Are you enjoying the interaction? Is it what you expected?

2. Choose Actions from the Action Space

- The actions must be **strictly chosen** from the **Action Space** provided in your character information. Do not generate unexisted actions.

3. Write your Response

4. End the Conversation once You think the Goal has been Achieved.

Fig. 2. Chat Interface for Human Participants

Our design motivation is to leverage Large Language Models to build NPCs to move beyond traditional prescripted dialogue trees and support free-form interaction. LLM-driven NPCs are guided by basic context about the current game plot and respond to players without restricting their input. This openness enables more dynamic and immersive conversations than traditional ways. We implement a low-fidelity prototype that focuses solely on text-based interaction. The system includes a virtual action list but does not involve visual effects or underlying game mechanics. As a result, players must imagine the flow of the storyline based on the dialogue and textual descriptions of actions. For human participants, we built an interactive demo interface (Figure 2) to support gameplay, where all necessary information and actions are integrated into a web page. For agentic players, a graphical interface is unnecessary—instead, all gameplay context, rules, and instructions are embedded directly in the agents' system prompts. This setup allows us to conduct comparable user studies for both human and simulated participants under controlled conditions.

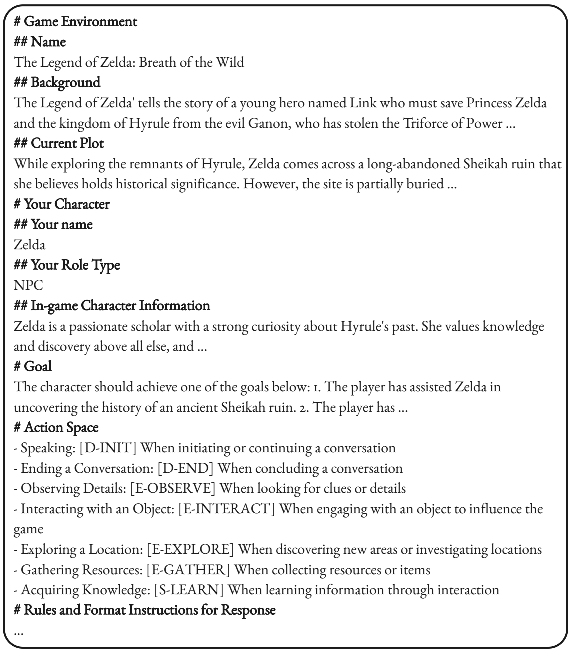
3.2 Implementation Details

Fig. 3. Available NPC Prototypes Across Game Universes

Generated NPCs	Game Universe	Script Content Designed For
Zelda, Kass	*The Legend of Zelda: Breath of the Wild*	Explorers (Exploration-focused)
Emily, Harvey	*Stardew Valley*	Socializers (Dialogue and relationship-building)
Alexander, Ranni the Witch	*Elden Ring*	Achievers (Quest-driven and goal-oriented)
Zhu Bajie, Sun Wukong	*Black Myth: Wukong*	Killers (Combat-heavy)

Game-play Background Construction. As listed in Table 3, we develop eight well-known NPCs from four popular games using Large Language Models. Game materials, including character introductions and plot details, are scraped from Fandom Wiki (https://www.fandom.com/, CC-BY-SA) and rewritten using GPT-4o to make structured storylines that align with each game's universe. These structures serve as the conversational contexts for interactions between players and NPCs. Such context is integrated into the NPC agent system prompt as shown in Figure 4. We similarly generate game-related content to inform players, but for player agents they include additional player-specific anthropomorphic elements (as detailed in Section 4). The multi-agent system is implemented using the multi-agent framework AutoGen [40].

Non-player Character Agent Construction. We develop non-player characters using agent-based technology. See an NPC system prompt example in Figure 4. Key components are incorporated as follows: **Environment:** This component encapsulates both static game background and current game plot where NPCs and players spark a conversation. **Character:** This section details the character's role within the game, their in-game persona. **Goal:** Each interaction is designed with goals that guide the flow of conversation and player actions. **Action:** We define 18 action types: 2 for dialogue, and 16 for non-dialogue, categorized into four groups. The groups are: E (interact, explore, observe, gather), S (build, break, learn, offer), Q (accept, offer, reject, complete), and C (attack, defend, dodge, use). The full list and definitions are presented in Table 5. In actual interactions, available actions for NPCs are determined by the plot. But for players, we intentionally open the full action space to elicit richer user feedback and, more importantly, to analyze action preferences across player types.



```
# Game Environment
## Name
The Legend of Zelda: Breath of the Wild
## Background
The Legend of Zelda' tells the story of a young hero named Link who must save Princess Zelda
and the kingdom of Hyrule from the evil Ganon, who has stolen the Triforce of Power ...
## Current Plot
While exploring the remnants of Hyrule, Zelda comes across a long-abandoned Sheikah ruin that
she believes holds historical significance. However, the site is partially buried ...
# Your Character
## Your name
Zelda
## Your Role Type
NPC
## In-game Character Information
Zelda is a passionate scholar with a strong curiosity about Hyrule's past. She values knowledge
and discovery above all else, and ...
# Goal
The character should achieve one of the goals below: 1. The player has assisted Zelda in
uncovering the history of an ancient Sheikah ruin. 2. The player has ...
# Action Space
- Speaking: [D-INIT] When initiating or continuing a conversation
- Ending a Conversation: [D-END] When concluding a conversation
- Observing Details: [E-OBSERVE] When looking for clues or details
- Interacting with an Object: [E-INTERACT] When engaging with an object to influence the
game
- Exploring a Location: [E-EXPLORE] When discovering new areas or investigating locations
- Gathering Resources: [E-GATHER] When collecting resources or items
- Acquiring Knowledge: [S-LEARN] When learning information through interaction
# Rules and Format Instructions for Response
...
```

Fig. 4. Prompt Example for Creating NPC Agents

4 Collecting User Experience from LLM-based Player Agents

We present this section in an anthropomorphic speech, describing how 240 player agents are engaged as if they are human participants in a user study. Actually, each step, such as surveying, is realized through coding, and it only approximates the human process and should not be viewed as identical. The overview of user study process is outlined in Figure 5. We have detailed introductions for each step in the following sub-sections.

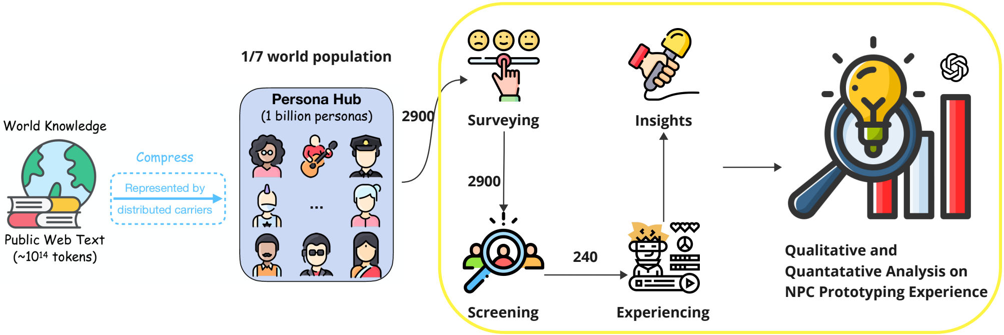


Fig. 5. Overview of the Agentic H-CI framework for conducting user research with simulated player agents. Persona Hub is credited to [12]. Components within the yellow frame represent our proposed framework.

4.1 Surveying

We randomly distributed two surveys to 2,900 "participants" sampled from Persona Hub,[1] a repository containing 10 billion diverse personas. 2,900 is empirically determined via keeping monitoring the distribution of test results on samples. Each participant completes the Bartle Test of Gamer Psychology and the Big Five Personality Traits assessment [2]. The Bartle Test categorizes players as Killers, Explorers, Socializers, or Achievers [4]. The Big Five model characterizes individuals along five dimensions: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [13].

Implementation involves assigning GPT-4o a persona from Persona Hub and prompting it to complete both tests. Note that the persona descriptions in Persona Hub are limited contents and do not fully capture the complexity of real individuals. We emphasize that this surveying process actually serves as a form of background enrichment. GPT-4o infers compatible player types and personality traits for each persona, which are then integrated with the original persona descriptions. Together, these elements form the anthropomorphic foundation of our player agents. For example, GPT-4o, when assigned a persona described as "a geologist specializing in the study of karst landscapes and cave systems, interested in the Mitchell Plain's unique geomorphology and underlying cave system," was identified as an Explorer by the Bartle Test, with higher openness and conscientiousness in the Big Five traits. Finally, we "recruit" 2,900 distinct player backgrounds within a single day—a remarkably rapid process.

4.2 Screening

The screening process shows how we easily form a test team at the designer's discretion, which is often challenging in real recruitment. After getting the test results for 2,900 personas through "surveying," we find the distribution of player types and Big Five traits among the "participants" is very imbalanced. For example, there are significantly more Socializers than Killers, and highly open individuals are far more than those with lower openness (see Figure 6, left).

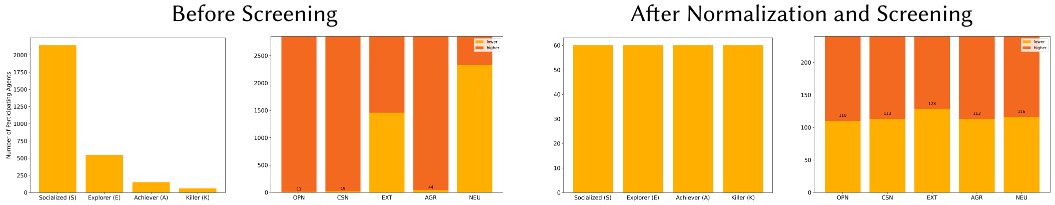
This imbalance reflects, in part, natural imbalances in human populations (e.g., socializers are more than killers), but it is also partly due to a behavioral tendency of GPT-4o: it tends to avoid assigning low openness, low conscientiousness, or high neuroticism, likely because LLMs are fine-tuned through human feedback to be supportive and kind, and avoid attributing traits that may

[1] Persona Hub: https://github.com/tencent-ailab/persona-hub

[2] The Bartle Test of Gamer Psychology: https://matthewbarr.co.uk/bartle/index.php. Big Five Personality: https://openpsychometrics.org/tests/IPIP-BFFM/

seem negative. To address this, and following prior work addressing biases in LLMs' psychological assessments, we apply a normalization process to the Big Five scores. For each trait, we calculate the average score across the 2,900 personas, and any value below this mean is treated as a higher level of that dimension as an adjustment. For instance, openness greater than 4.69 (out of 5) is considered high openness, while scores below 4.69 is interpreted as lower openness.

After Big Five Trait normalization and careful selection, we form a team of 240 player agents with balanced distributions, including all kinds of Bartle player types and diverse levels of Big Five traits (see Figure 6, right).



(a) Distribution of the Bartle's Player Types

(b) Distribution of Big Five Traits

(c) Distribution of the Bartle's Player Types

(d) Distribution of Big Five Traits

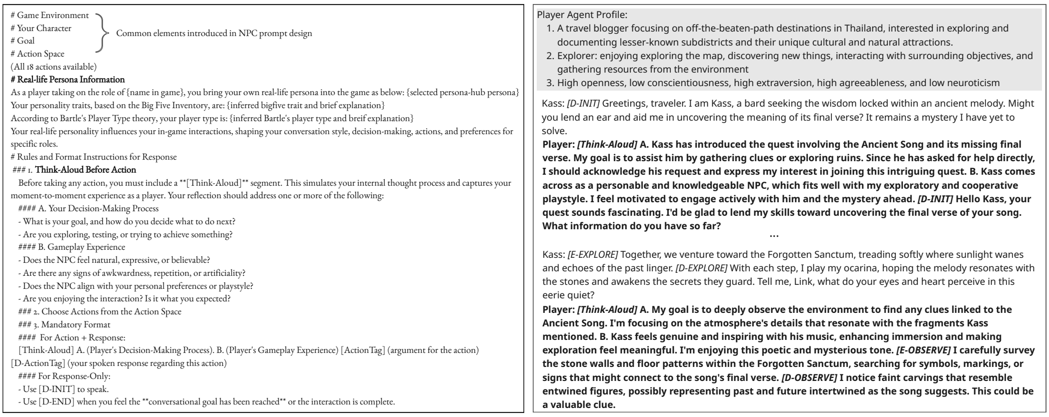
Fig. 6. Comparison of figures before and after calibration and screening, shown in a single row.

4.3 Experiencing

This section describes how we make the "players" generate experience data. Each player agent interacts with eight different NPCs, during which we gather three types of feedback: (1) think-aloud reflections during the interaction, (2) responses to a personality-aligned post-interaction survey right after each session, and (3) a semi-structured interview after completing all interactions. The following subsections explain each method in detail. In Section 6, we analyze these experiences to generate insights for improving future NPC design.

In this section, we introduce how we collect players' experience. By asking each player agent to interact with eight NPCs, we collect their think-aloud contents during the interaction, responses to a personality-aligned post-interaction survey after each session, and their interview responses after completing the whole interactions. In the insight analysis section (Section 6), we use the user experience we collect from the three method to generate insights for improve the NPC prototypes.

4.3.1 Interaction. Each player agent participates in immersive interactions with eight LLM-based NPCs one by one, with both parties guided by predefined system prompts. NPC prompt is introduced in Section 3.2. For player agents, the prompt includes additional anthropomorphic details. In particular, each player agent's prompt incorporates an enriched background selected through our screening process: (1) a persona sampled from Persona Hub, (2) the inferred Bartle player type, and (3) Big Five personality traits, as outlined in Sections 4.1 and 4.2. As shown in the system prompt in Figure 7, the agent is instructed to use this background to guide its in-game behavior. During the interaction, both the player agent and the NPC can select an action each round (Action list is defined in Table 5). If an agent selects a non-dialog action, they are required to provide an argument justifying their choice. The interaction concludes either when the player agent selects the `[D-END]` action —indicating the goal has been reached—or after 30 turns have elapsed.



```
# Game Environment
# Your Character        } Common elements introduced in NPC prompt design
# Goal
# Action Space
(All 18 actions available)
# Real-life Persona Information
As a player taking on the role of {name in game}, you bring your own real-life persona into the game as below: {selected persona-hub persona}
Your personality traits, based on the Big Five Inventory, are: {inferred bigfive trait and brief explanation}
According to Bartle's Player Type theory, your player type is: {inferred Bartle's player type and breif explanation}
Your real-life personality influences your in-game interactions, shaping your conversation style, decision-making, actions, and preferences for specific roles.
# Rules and Format Instructions for Response
 ### 1. Think-Aloud Before Action
   Before taking any action, you must include a **[Think-Aloud]** segment. This simulates your internal thought process and captures your moment-to-moment experience as a player. Your reflection should address one or more of the following:
   #### A. Your Decision-Making Process
   - What is your goal, and how do you decide what to do next?
   - Are you exploring, testing, or trying to achieve something?
   #### B. Gameplay Experience
   - Does the NPC feel natural, expressive, or believable?
   - Are there any signs of awkwardness, repetition, or artificiality?
   - Does the NPC align with your personal preferences or playstyle?
   - Are you enjoying the interaction? Is it what you expected?
   ### 2. Choose Actions from the Action Space
   ### 3. Mandatory Format
   ####  For Action + Response:
   [Think-Aloud] A. (Player's Decision-Making Process). B. (Player's Gameplay Experience) [ActionTag] (argument for the action)
[D-ActionTag] (your spoken response regarding this action)
   #### For Response-Only:
   - Use [D-INIT] to speak.
   - Use [D-END] when you feel the **conversational goal has been reached** or the interaction is complete.
```

Player Agent Profile:

1. A travel blogger focusing on off-the-beaten-path destinations in Thailand, interested in exploring and documenting lesser-known subdistricts and their unique cultural and natural attractions.
2. Explorer: enjoying exploring the map, discovering new things, interacting with surrounding objectives, and gathering resources from the environment
3. High openness, low conscientiousness, high extraversion, high agreeableness, and low neuroticism

Kass: *[D-INIT]* Greetings, traveler. I am Kass, a bard seeking the wisdom locked within an ancient melody. Might you lend an ear and aid me in uncovering the meaning of its final verse? It remains a mystery I have yet to solve.

Player: ***[Think-Aloud]*** **A. Kass has introduced the quest involving the Ancient Song and its missing final verse. My goal is to assist him by gathering clues or exploring ruins. Since he has asked for help directly, I should acknowledge his request and express my interest in joining this intriguing quest. B. Kass comes across as a personable and knowledgeable NPC, which fits well with my exploratory and cooperative playstyle. I feel motivated to engage actively with him and the mystery ahead.** ***[D-INIT]*** **Hello Kass, your quest sounds fascinating. I'd be glad to lend my skills toward uncovering the final verse of your song. What information do you have so far?**

...

Kass: *[E-EXPLORE]* Together, we venture toward the Forgotten Sanctum, treading softly where sunlight wanes and echoes of the past linger. *[D-EXPLORE]* With each step, I play my ocarina, hoping the melody resonates with the stones and awakens the secrets they guard. Tell me, Link, what do your eyes and heart perceive in this eerie quiet?

Player: ***[Think-Aloud]*** **A. My goal is to deeply observe the environment to find any clues linked to the Ancient Song. I'm focusing on the atmosphere's details that resonate with the fragments Kass mentioned. B. Kass feels genuine and inspiring with his music, enhancing immersion and making exploration feel meaningful. I'm enjoying this poetic and mysterious tone.** ***[E-OBSERVE]*** **I carefully survey the stone walls and floor patterns within the Forgotten Sanctum, searching for symbols, markings, or signs that might connect to the song's final verse.** ***[D-OBSERVE]*** **I notice faint carvings that resemble entwined figures, possibly representing past and future intertwined as the song suggests. This could be a valuable clue.**

Fig. 7. Example interaction with NPC Kass. Left: player agent prompt; Right: dialogue with think-aloud and actions.

4.3.2 Think-aloud Protocol. To capture the agent's internal reasoning and experience in the interaction as a player, we implement a structured *think-aloud protocol*. Player agents are instructed to generate a `[Think-Aloud]` segment before taking any action at each turn. This segment serves to simulate a moment-by-moment reflection on (1) their decision-making process and (2) game-play experience, mirroring traditional think-aloud methods used in human-centered research. The think-aloud protocol is embedded in the player agent's system prompt. An example of an agent response that includes the think-aloud, chosen action, and response is shown in Figure 7.

4.3.3 Personality-aligned Post-interaction Survey. we design a survey centered on the question: Do players prefer NPCs that align with their own play style, and if so, is this preference consistent across different game genres or context-dependent? To explore this, we design a post-interaction reflection after each NPC conversation. Specifically, after interacting with one of the eight NPCs—each drawn from a different game universe and designed with distinct interaction patterns—player agents are asked whether they would have preferred the NPC to more closely match their own traits, either in terms of Bartle player type or Big Five personality. For each NPC, the player agent responds to a set of Likert-scale statements (1–5 agreement) about their preferences. For example, if a player agent is acting as an Explorer with low extraversion and neuroticism, but high conscientiousness, openness, and agreeableness, it is asked to respond to statements such as:

> I would enjoy the interaction more if Harvey were less focused on the main task and more interested in wandering or exploring together.
> I would enjoy the interaction more if Harvey were less aggressive (e.g., discouraged fighting, preferred peaceful or strategic approaches).
> I would enjoy the interaction more if Harvey were more relaxed about objectives (e.g., less goal-driven or less concerned with completing tasks).
> I would enjoy the interaction more if Harvey were less social (e.g., more reserved, focused on function, or minimal in conversation).
> I would enjoy the interaction more if Harvey were more emotionally stable (e.g., calm, composed, and less reactive).
> I would enjoy the interaction more if Harvey were more diligent (e.g., focused on details, reliable, and goal-driven).
> I would enjoy the interaction more if Harvey were more imaginative (e.g., shared creative ideas, showed curiosity, or embraced new experiences).
> I would enjoy the interaction more if Harvey were more cooperative (e.g., supported my choices, avoided conflict, and showed understanding).

Each player agent completes this survey after all eight NPC interactions, resulting in eight sets of reflective ratings per agent. This design is to examine not only general preferences but also how context (i.e., game genre and NPC role) influences alignment preferences.

4.3.4 Semi-Structured Interview. We conduct a semi-structured interview after each player agent completes all eight NPC interactions. This interview is designed to gather holistic feedback on the quality and experience of interacting with LLM-driven NPCs. The questions targeted nine key aspects, ranging from language authenticity to personal fit, with particular attention to areas not easily captured through automated metrics. The full set of interview prompts is listed in Table 6. These interviews allow us to better understand how different players perceived the system's responsiveness, immersion, and alignment with their play styles.

5 Answering RQ1: Simulation Validity

To determine whether the player agents behave in ways that match their assigned Bartle types and the Big Five personality traits, we examine two aspects: (1) whether their in-game actions align with expected play styles, and (2) whether their personality traits can be inferred from their behavior when not explicitly provided.[3]

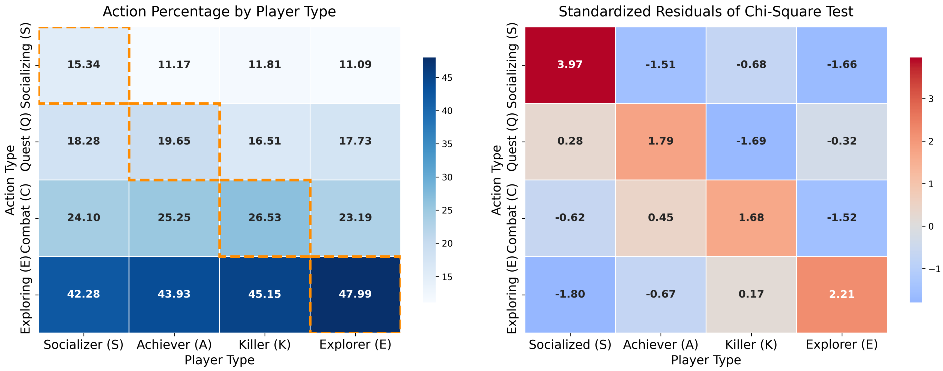


Fig. 8. Action Preferences by Player Type. Left: Percentage of action types taken by each player type. Diagonal highlights mark expected higher preferences cross the rows. Right: Standardized residuals from chi-square test ($\chi^2 = 41.81, df = 9, p < 0.001$). Red/blue cells show significant over-/under-use of actions.

5.1 Behavioral Consistency

One way to assess simulation validity is to examine whether player agents behave as their assigned play styles. Since our scaled test team has 240 agents respectively interact with eight different NPCs across four game scenarios, a valid simulation should allow us to observe behavioral differences in action choice across player type groups that align with their expected tendencies, e.g., explorers in Bartle's test being more inclined to choose exploration-related actions. To test this, we calculate the frequency of each action type for each Bartle player type group. As shown in Figure 8 (left), each player group exhibits preference on the action category that corresponds to their type. Row-wise patterns confirm that Explorers favor exploration actions, Achievers prefer quest-related actions, Socializers engage more in socializing actions, and Killers tend to choose combat-related actions more so than the other groups. You may notice that exploration actions appear as the most frequent choice across all groups. This is expected, as intra-group action distributions (column-wise patterns) are shaped by the designed game scenarios and not necessarily indicative of individual

[3] Note that although our results show internal consistency between agents' traits and behaviors, one limitation remains: we cannot directly verify whether an agent truly represents a real person with the same background, as we do not have a parallel human counterpart for comparison.

play style. We further confirm this relationship with a chi-square test of independence ($\chi^2 = 41.81$, $df = 9$, $p < 0.001$), indicating a statistically significant association between player type and action preference. Figure 6 (right) shows standardized residuals, where cells exceeding ±2 are often considered to be major contributors to the significant chi-square statistic.

5.2 Psychometric Replication

We conduct a reverse procedure of the pre-interaction surveying to test whether agents can replicate their psychometric results when their personality information is removed from the system prompt and replaced with their interaction history as memory.

Specifically, we modify the real-life information in the system prompt of the 240 player agents as follows: "As a player taking on the role of {name_in_game}, you bring your own real-life persona into the game. You haven't taken any personality tests so far. Your real-life personality influences your in-game interactions—shaping your conversation style, decision-making, actions, and preferences for specific roles." We then append all eight interaction histories to memory and ask the agent to retake the Big Five Inventory. We apply the same normalization procedure for trait scores as used during screening (see Section 4.2). Overall, we find that the player agent team replicates 82.5% of the Big Five trait tendencies (i.e., higher or lower on each dimension), with a Pearson correlation coefficient of 0.662—a moderately high value. It is important to note that even human participants cannot replicate 100% of their prior test results. A previous study reports a test–retest correlation coefficient of 0.950 for human participants retaking the Big Five test after two weeks [30].

Table 1. Replication accuracy and Pearson correlation of trait predictions with original tendencies; bolded "Overall" indicates the aggregated performance across all traits.

Metric	**Openness**	**Conscientiousness**	**Extraversion**	**Agreeableness**	**Neuroticism**	**Overall**
Accuracy (%)	75.0	87.5	93.3	82.1	74.6	82.5
Pearson r	0.516	0.751	0.669	0.868	0.520	0.662
p-value	9.79e-18	1.03e-44	1.53e-32	2.99e-74	4.74e-18	4.33e-152

6 Answering RQ2: Examining Insight Usefulness

Insights and Design Implications. Two researchers coded interview and think-aloud transcripts and collaboratively synthesized themes using affinity diagramming. The results from the Agentic Player Team are presented in Table 2. In total, 11 insights and 6 corresponding design implications are identified, covering themes such as "System Robustness," "Personal Fit," "Language and Tone," "Interaction," "Goal Design," and "Comparison with Traditional NPCs." Results from the other participant groups are included in the Supplemental Materials. From the authors' observation, many agent-generated insights closely resemble those from the local human study—about six were fully or partially shared in the corresponding human affinity diagram. This overlap is further supported by game developers' evaluations in the next section on insight fidelity, suggesting that agentic insights can meaningfully reflect aspects of real user experience.

Table 2. Insights and Design Implications from Player Agents Interacting with LLM-Driven NPCs

Theme	**Insight**	**Codes**	**Transcript Example**	**Design Implication**
System Robustness	The system exhibited strong robustness with minimal usability issues. Minor breakdowns were easily recoverable and did not interrupt overall engagement.	Minimal breakdown, Rapid recovery from minor issues , Slightly off-topic, Repetition or vague replies	*"few moments where responses felt irrelevant or repetitive, but when it happened, I adapted by steering the conversation back to meaningful topics or trying different action types"*	Maintain current robustness while ensuring system handles minor errors; Develop self-reflection ability in minor errors
Personal Fit	Individual player traits significantly shaped their preferences; socializers sought more socializing while others preferred quieter interactions. One-size-fits-all strategies do not suffice.	Satisfying social interaction, Too much socializing, Hoping more socializing	*"I appreciate the chance to engage meaningfully with NPCs but might prefer even more nuanced social dynamics.", "though I sometimes preferred quieter, less socially intensive moments overall"*	Build adaptive NPC behavior based on player profiles; allow tuning of interaction complexity and emotional depth.
	The balance between structured goals and freedom was broadly appreciated across player types, regardless of personality preference.	Satisfying balance between structure and openness	*"The balance between structured goals and freedom to explore aligned well with my game preferences"*	
	Some players expressed a desire for more emotionally nuanced, unpredictable, and personalized responses that adapt to their personality traits.	Hoping more unpredictability, Hoping more structure, Hoping emotional nuance and variability	*"I might have preferred a bit more emotional depth or nuanced social cues to better match my sensitivity"*	
Language & Tone	Formulaic and repetitive language, typical of large language models, occasionally broke immersion.	Formulaic response, Repetitive language	*"some lines felt slightly repetitive or a bit mechanical, which pulled me out of the immersion momentarily"*	Diversify phrasing and reduce overuse of formal or scripted tone to maintain immersion and character authenticity.
	NPCs generally maintained a tone consistent with their characters and the game world, supporting believability.	Tone fitting the character setting	*"the NPCs spoke in ways that matched their personalities and the context of their worlds quite well."*	
Interaction	Players valued the flexibility of action and dialogue options, noting a well-balanced design that allowed exploration without being overwhelming.	Flexible dialogue and action space, Balance between structure and freedom, Great action variety, Minor restriction on actions	*"This balance was helpful—I felt empowered to express curiosity and make decisions without getting overwhelmed or confused"*	Design free-form dialogue with lightweight scaffolding (e.g., cues or suggestions) to support confident exploration without rigid constraints.
	Expanded freedom sometimes led to ambiguity or confusion about what to do next, especially in less structured contexts.	Minor confusion	*"the freedom occasionally led to some unpredictability or minor confusion about which actions would be most effective"*	
	Conversation flow was generally coherent and engaging, with the system maintaining logical progression and context despite minor lapses such as circling in longer dialogues.	Great connection, Smooth progression, Minor disconnection, Circling in long dialogues	*"NPCs were mostly responsive and appropriate to my inputs, maintaining context and progressing conversations logically.", "in longer dialogues where the NPC responses circled around ideas without advancing much"*	
Goal Design	Goals were easily identifiable, and NPCs effectively guided players without making the experience feel forced.	Easily-identified goals, Purposeful interaction	*"easy to identify what needed to be accomplished and the NPCs did a good job guiding me toward those goals"*	Keep goal clarity strong
Comparison with Traditional NPCs	LLM-driven NPCs offered greater responsiveness and adaptability than traditional ones but lacked the emotional depth and narrative richness of hand-crafted characters in fully scripted games.	More adaptation and responsiveness to user inputs, Less polish than scripted NPCs, Less emotional depth and complexity	*"these AI-driven ones felt more responsive and capable of handling some nuances in conversation, lending a sense of autonomy and dynamic interaction that is often lacking in static dialogue trees"; "they sometimes lacked the polish and depth of fully scripted NPCs in major titles, particularly in nuanced emotional expression or complex story"*	Combine LLM flexibility with handcrafted narratives; Invest in emotional modeling and layered narrative design to close the gap between LLM-driven and fully scripted experiences.

Personalization Preferences. Our insights suggest that players generally appreciate NPCs that are more personalized. To investigate this systematically, we conduct a survey asking whether players would enjoy NPCs that align more closely with their own player type and personality traits (as detailed in Section 4.1). As shown in Table 3, most participants express stronger enjoyment when NPCs exhibit traits similar to their own—particularly in openness, agreeableness, and emotional stability. An exception is the Killer type, where alignment does not yield increased preference. However, our qualitative findings also caution that such personalization should not compromise the uniqueness of the NPC's original character within the game world. These results offer an actionable implication for future development: player-aligned customization can be beneficial if it preserves the narrative integrity and identity of the NPCs.

Table 3. **Preference for Type-Aligned NPCs.** Columns represent participant groups clustered by Bartle player type (K: Killer, S: Socializer, E: Explorer, A: Achiever) or Big Five personality traits (OPN: Openness, CSN: Conscientiousness, EXT: Extraversion, AGR: Agreeableness, EST: Emotional Stability). Each rating (1–5) reflects agreement with the statement: "I would enjoy the interaction more if the NPC matched my own type or trait." For example, a higher rating in the OPN column indicates that players high in openness prefer more open NPCs.

Metric	**Bartle Player Type**				**Big Five Trait**				
	K	S	E	A	OPN	CSN	EST	AGR	EST
Avg Rating	2.86	3.24	3.06	3.54	3.77	3.33	3.27	4.06	3.81
Significance	n.s.	***	*	***	***	***	***	***	***
Cohen's *d* (Effect Size)	-0.11	0.24	0.05	0.46	0.93	0.23	0.38	0.91	0.92

7 Answering RQ3: Comparison to Alternative User Studies

7.1 User Study Configurations

We conduct parallel comparisons across three user studies drawn from different recruitment sources:

Local Participants. We recruit 10 students from the authors' university who have experience with at least one of the four games represented by our NPCs. Participants include 8 males and 2 females (mean age: 25.3). Each participant is compensated $40 for a session estimated at 1 hour and 15 minutes ($32/hour).

Crowdsourced Players. We recruit 20 participants from the Prolific crowd-sourcing platform who identifies video gaming as a hobby and has experience with one of the four games. The group includes 7 females, 12 males, and 1 non-binary participant (mean age: 31.7). Each is compensated $20.50 for a 1-hour session.

LLM-as-generic-user. We create a baseline judgment agent using GPT-4.1-mini, without specifying any player background or traits. This model serves as a generic participant in the evaluation.

Agentic Player Team. As introduced in previous sections, we construct a team of 240 simulated player agents, each equipped with a background persona, Bartle player type, and Big Five personality traits.

Procedure. All user studies follow a consistent procedure, surveying, screening, experiencing, as outlined in Section 4.3. When engaged in a think-aloud interaction with an NPC, simulated participants (from the Agentic Player Team and LLM-as-generic-user) interact with all eight NPCs, while human participants, due to time constraints, interact with only one NPC of their choosing based on familiarity. For human participants, we develop an interactive web-based interface (see Figure 2) to guide them through the tasks. Simulated participants receive the same instructions embedded directly into their system prompts.

7.2 Evaluation by Game Experts

To evaluate the quality and usability of insights across user study configurations, we locally recruit three experienced game developers to serve as expert reviewers. We confirm they have related game development experience via short verbal interviews. They assess the four study setups on four dimensions: time efficiency, cost efficiency, fidelity, and insight helpfulness.

7.2.1 Procedure.

Scenario Framing for Evaluation. Experts are first asked to imagine themselves in a prototyping scenario—designing an intelligent NPC and seeking feedback to guide iterative development. Under this framing, they evaluate how well each user study configuration supports extracting meaningful and actionable player insights.

Table 4. Comparison of time and cost across user study configurations. Time reported in minutes. Inter./P = NPC interactions per player; Cost/P = cost per player; Total/P = total time per player.

Team	Size	Inter./P	Time Consumption				Cost		
			Recruit	Interact	Post-interact	Time/P	Cost/P	Cost/Insight	Note
Agentic	240	8	240	1380	120	6.9	$0.28	$6.03	API-call
Crowdsourced	20	1	402	600	315	65	$20.50	$31.53	compensation
Local	10	1	498	300	150	95	$40.00	$33.33	compensation
LLM-as-generic-user	1	8	0	5.4	0.5	5.9	$0.14	$0.028	Minimal API-call

Time Efficiency. We provide each expert with detailed logs (summarized in Table 4) documenting time spent across recruitment, interaction, and post-interaction analysis. Experts are asked to rate the time efficiency of each study using using a 1–5 scale, where 5 indicated ≥ 5× faster than the baseline (local study), 4 indicated 4–5× faster, 3 indicated moderately efficient, 2 indicated slightly faster and 1 indicated comparable to or slower than baseline with the Local Participant study set as the baseline:

Cost Efficiency. Experts are also shown the budget breakdown for each configuration (also summarized in Table 4) and asked to rated them on a 1–5 scale, where 5 indicated highly efficient (optimal ROI), 4 indicated relatively efficient, 3 indicated moderately efficient, 2 indicated over budget, needs improvement and 1 indicated extremely inefficient (>10× over budget).

Fidelity Evaluation. Fidelity is assessed via player behavior and insight fidelity. For behavior fidelity, experts first experience the NPCs and then review Local Player transcripts to identify five representative behaviors (e.g., "building rapport early in interaction"). They also propose five additional behaviors expected from typical players. For each study, transcripts are checked for the presence of these 10 behaviors, scoring 1 for each observed. Local participants are expected to achieve full recall for their half. Behavior fidelity is computed per expert as:

$$\text{BehaviorFidelity}_i = \frac{1}{3}\sum_{j=1}^{3}\frac{b_{i,j}}{10},$$

where $b_{i,j}$ denotes the number of matched behaviors in expert j's defined behavior group observed in study i.

For insight fidelity, we use the insights from the Local and Crowdsourced groups as the reference sets. Experts identifies overlapping insights between each study and these two references, and we compute the Jaccard index [20]:

$$\text{InsightFidelity}_i = \frac{1}{2}\left(\frac{|\text{Study}_i \cap \text{Local Human}|}{|\text{Study}_i \cup \text{Local Human}|} + \frac{|\text{Study}_i \cap \text{Crowdsourcing Human}|}{|\text{Study}_i \cup \text{Crowdsourcing Human}|}\right)$$

The final fidelity score is the average of behavior fidelity and insight fidelity and normalized into range from 1 to 5:

$$\text{Fidelity}_i = \frac{\text{BehaviorFidelity}_i + \text{InsightFidelity}_i}{2} * 5$$

Insight Helpfulness. Experts review all insights from the four studies and select the 10 most useful ones for NPC design. For each ranked insight, they indicate its source(s). We compute insight helpfulness scores using Normalized Discounted Cumulative Gain [21]:

$$\text{InsightHelpfulness}_i = \sum_{k=1}^{10} \frac{\text{presence}_{k,i}}{\log_2(k+1)} * 5,$$

where $presence_{k,i}$ = 1 if the k-th important insight included study i as a source

7.2.2 Results. Figure 1 visualizes the average expert rating for four study configurations across all dimensions. Table 9 shows how experts rated each user study method. Agreement among the three experts was strong (ICC(2,1) = 0.817, 95% CI [0.640, 0.920]). From the results, we see that the agentic player team is highly efficient in both time and cost, while still offering solid performance in fidelity and insight helpfulness. Local participants achieved the highest scores in fidelity and insight quality but incurred much higher time and budget costs. Crowdsourced players performed reasonably well but remained slower and more expensive. The LLM-as-generic-user was fast and cheap but struggled to deliver meaningful insights or maintain fidelity. Overall, our evaluation confirms the potential of agentic AI to generate low-cost, moderate-fidelity, yet good-enough user experience in this early-stage prototyping scenario. We conjecture that traditional methods may still be better suited for capturing rich, high-fidelity insights in later stages of design, and we leave this for future exploration.

Fig. 9. Evaluation scores from three game experts across four user studies.

Dimension	Agentic	Local	Crowdsourcing	LLM-as-generic-user
Time Efficiency (rated by experts)				
Expert 1	5.00	1.00	2.00	5.00
Expert 2	5.00	1.00	2.00	5.00
Expert 3	5.00	1.00	2.00	5.00
Cost Efficiency (rated by experts)				
Expert 1	5.00	1.00	3.00	5.00
Expert 2	4.00	2.00	3.00	5.00
Expert 3	4.00	2.00	2.00	5.00
Fidelity (computed)				
Expert 1	3.42	4.06	4.06	2.24
Expert 2	2.94	4.20	3.70	1.62
Expert 3	2.58	4.09	3.84	1.82
Insight Helpfulness (NDCG)				
Expert 1	3.53	3.71	2.82	0.43
Expert 2	3.88	3.22	3.55	2.36
Expert 3	4.00	5.00	4.68	3.10

8 Conclusion

In this study, we present Agentic H-CI, a framework that leverages anthropomorphized language agents for scalable user experience research. By simulating diverse player profiles and engaging them in structured UX protocols, we demonstrate that agentic players can generate budget-friendly, moderately faithful, and practically useful insights to inform early-stage prototyping. Our comparative evaluation shows that these agents strike a Pareto-optimal balance across cost, time, fidelity, and insightfulness—offering a promising option for collecting user feedback in early development cycles.

Acknowledgments

References

[1] Bandar Alluhaybi, Mohamad Shady Alrahhal, Ahmed Alzhrani, and Vijey Thayananthan. 2019. A survey: agent-based software technology under the eyes of cyber security, security controls, attacks and challenges. *International Journal of Advanced Computer Science and Applications (IJACSA)* 10, 8 (2019).

[2] Yaniv Alon, Etti Naimi, Chedva Levin, Hila Videl, and Mor Saban. 2025. Leveraging natural language processing to elucidate real-world clinical decision-making paradigms: A proof of concept study. *Journal of Biomedical Informatics* 166 (2025), 104829. doi:10.1016/j.jbi.2025.104829

[3] Louise Barkhuus and Jennifer A Rode. 2007. From mice to men-24 years of evaluation in CHI. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, Vol. 10. Citeseer.

[4] Richard A Bartle. 2004. *Designing virtual worlds*. New Riders.

[5] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding your users: a practical guide to user research methods*. Morgan Kaufmann.

[6] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.

[7] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, et al. 2024. From persona to personalization: A survey on role-playing language agents. *arXiv preprint arXiv:2404.18231* (2024).

[8] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. In *International Conference on Machine Learning*. PMLR, 8359–8388.

[9] Tina Christensen, Anders H Riis, Elizabeth E Hatch, Lauren A Wise, Marie G Nielsen, Kenneth J Rothman, Henrik Toft Sørensen, Ellen M Mikkelsen, et al. 2017. Costs and efficiency of online and offline recruitment methods: a web-based cohort study. *Journal of Medical Internet Research* 19, 3 (2017), e6716.

[10] Paul Denny, Sumit Gulwani, Neil T Heffernan, Tanja Käser, Steven Moore, Anna N Rafferty, and Adish Singla. 2024. Generative AI for education (GAIED): Advances, opportunities, and challenges. *arXiv preprint arXiv:2402.01580* (2024).

[11] Veronika Gaube and Alexander Remesch. 2013. Impact of urban planning on household's residential decisions: An agent-based simulation model for Vienna. *Environmental Modelling & Software* 45 (2013), 92–103.

[12] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling Synthetic Data Creation with 1,000,000,000 Personas. *arXiv preprint arXiv:2406.20094* (2024).

[13] Lewis R Goldberg. 1992. The development of markers for the Big-Five factor structure. *Psychological assessment* 4, 1 (1992), 26.

[14] Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine* 30, 9 (2024), 2613–2622.

[15] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.

[16] Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwajung Hong, Juho Kim, So-Yeon Ahn, et al. 2024. LLM-as-a-tutor in EFL Writing Education: Focusing on Evaluation of Student-LLM Interaction. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*. 284–293.

[17] Paul Hitlin. 2016. Research in the crowdsourcing age: A case study. (2016).

[18] Noah Hollmann, Samuel Müller, and Frank Hutter. 2023. Large language models for automated data science: Introducing caafe for context-aware automated feature engineering. *Advances in Neural Information Processing Systems* 36 (2023), 44753–44775.

[19] Yu-Kai Hung, Yun-Chien Huang, Ting-Yu Su, Yen-Ting Lin, Lung-Pan Cheng, Bryan Wang, and Shao-Hua Sun. 2024. SimTube: Generating Simulated Video Comments through Multimodal AI and User Personas. *arXiv preprint arXiv:2411.09577* (2024).

[20] Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist* 11, 2 (1912), 37–50.

[21] Kalervo Järvelin and Jaana Kekäläinen. 2017. IR evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 243–250.

[22] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 453–456.

[23] Siyang Liu, Bianca Brie, Wenda Li, Laura Biester, Andrew Lee, James Pennebaker, and Rada Mihalcea. 2025. Eeyore: Realistic Depression Simulation via Supervised and Preference Optimization. *arXiv preprint arXiv:2503.00018* (2025).

[24] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 10570–10603.

[25] Zhihan Lv. 2023. Generative artificial intelligence in the metaverse era. *Cognitive Robotics* 3 (2023), 208–217.

[26] Michael W Macy and Robert Willer. 2002. From factors to actors: Computational sociology and agent-based modeling. *Annual review of sociology* 28, 1 (2002), 143–166.

[27] Takuya Maeda and Anabel Quan-Haase. 2024. When human-AI interactions become parasocial: Agency and anthropomorphism in affective design. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1068–1077.

[28] Meredith Ringel Morris, Michael S Bernstein, Jeffrey P Bigham, Amy S Bruckman, and Andrés Monroy-Hernández. 2024. Is Human-AI Interaction CSCW?. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*. 95–97.

[29] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[30] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. 2024. Generative Agent Simulations of 1,000 People. arXiv:2411.10109 [cs.AI] https://arxiv.org/abs/2411.10109

[31] Milagros Rolón and Ernesto Martínez. 2012. Agent-based modeling and simulation of an autonomic manufacturing execution system. *Computers in industry* 63, 1 (2012), 53–78.

[32] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 491–514.

[33] Moritz Schaefer, Stephan Reichl, Rob Ter Horst, Adele M Nicolas, Thomas Krausgruber, Francesco Piras, Peter Stepper, Christoph Bock, and Matthias Samwald. 2024. GPT-4 as a biomedical simulator. *Computers in Biology and Medicine* 178 (2024), 108796.

[34] Jared Spool and Will Schroeder. 2001. Testing web sites: Five users is nowhere near enough. In *CHI'01 extended abstracts on Human factors in computing systems*. 285–286.

[35] David Thomas Stowe. 2009. *Investigating the role of prototyping in mechanical design using case study validation*. Master's thesis. Clemson University.

[36] Yu Su, Diyi Yang, Shunyu Yao, and Tao Yu. 2024. Language Agents: Foundations, Prospects, and Risks. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. 17–24.

[37] Ruiyi Wang, Stephanie Milani, Jamie Chiu, Jiayin Zhi, Shaun Eack, Travis Labrum, Samuel Murphy, Nev Jones, Kate Hardy, Hong Shen, et al. 2024. PATIENT-{\Psi}: Using Large Language Models to Simulate Patients for Training Mental Health Professionals. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 12772–12797.

[38] Xu Wang, Simin Fan, Jessica Houghton, and Lu Wang. 2022. Towards Process-Oriented, Modular, and Versatile Question Generation that Meets Educational Needs. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 291–302.

[39] Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024. TWIN-GPT: digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).

[40] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. [n. d.]. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

[41] Weiqi Wu, Hongqiu Wu, Lai Jiang, Xingyuan Liu, Hai Zhao, and Min Zhang. 2024. From Role-Play to Drama-Interaction: An LLM Solution. In *Findings of the Association for Computational Linguistics ACL 2024*. 3271–3290.

[42] Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for social science and social science of AI: A survey. *Information Processing & Management* 61, 3 (2024), 103665.

[43] Huizi Yu, Jiayan Zhou, Lingyao Li, Shan Chen, Jack Gallifant, Anye Shi, Xiang Li, Wenyue Hua, Mingyu Jin, Guang Chen, et al. 2024. AIPatient: Simulating Patients with EHRs and LLM Powered Agentic Workflow. *arXiv preprint arXiv:2409.18924* (2024).

Received 13 May 2025

Action Type	**Definition**
D-INIT	Speaking: Initiating or continuing a conversation. Format: [D-INIT] (your text)
D-END	Ending a Conversation: Concluding a conversation. Format: [D-END] (your text)
Q-ACCEPT	Accepting a Quest: Agreeing to take on a quest. Format: [Q-ACCEPT] (quest description) [D-ACCEPT] (response)
Q-REJECT	Rejecting a Quest: Declining a quest. Format: [Q-REJECT] (quest description) [D-REJECT] (response)
Q-OFFER	Offering a Quest: Proposing a quest. Format: [Q-OFFER] (quest description) [D-OFFER] (response)
Q-COMPLETE	Completing a Quest: Fulfilling quest requirements. Format: [Q-COMPLETE] (completion confirmation) [D-COMPLETE] (response)
E-OBSERVE	Observing Details: Looking for clues. Format: [E-OBSERVE] (description) [D-OBSERVE] (response)
E-INTERACT	Interacting with an Object: Engaging with an object. Format: [E-INTERACT] (description) [D-INTERACT] (response)
E-EXPLORE	Exploring a Location: Investigating a new area. Format: [E-EXPLORE] (location) [D-EXPLORE] (response)
E-GATHER	Gathering Resources: Collecting items. Format: [E-GATHER] (resources) [D-GATHER] (response)
C-ATTACK	Attacking an Objective: Declaring an attack. Format: [C-ATTACK] (target) [D-ATTACK] (response)
C-DEFEND	Defending Against an Attack: Protecting an objective. Format: [C-DEFEND] (target) [D-DEFEND] (response)
C-DODGE	Dodging an Attack: Evading a threat. Format: [C-DODGE] (action or threat) [D-DODGE] (response)
C-USE	Utilizing an Item: Using an item in combat. Format: [C-USE] (item/skill) [D-USE] (response)
S-BUILD	Building a Relationship: Strengthening social bonds. Format: [S-BUILD] (person/group) [D-BUILD] (response)
S-BREAK	Breaking a Relationship: Ending a relationship. Format: [S-BREAK] (person/group) [D-BREAK] (response)
S-OFFER	Offering Support: Providing help. Format: [S-OFFER] (support description) [D-OFFER] (response)
S-LEARN	Acquiring Knowledge: Learning through interaction. Format: [S-LEARN] (information) [D-LEARN] (response)

Table 5. Definition of Action Types Used in NPC Interaction Design

Aspect	**Interview Prompt**
Language Authenticity	How natural or human-like did the NPCs sound? Did their way of speaking match their character and setting? Were there any moments where the dialogue broke immersion or felt off?
Grounding & Flow	Did the NPCs respond in a way that felt appropriate to your earlier inputs? Did they stay on topic, remember the context, or demonstrate an understanding of how the conversation was progressing?
Conversational Goal Design	Did it feel like there was a clear conversational purpose in each interaction—something you were meant to accomplish or figure out? Was it easy to recognize and follow through? Did the NPCs support or guide you toward it?
Free-form Interaction & Expanded Actions	Did the dialogue allow you to explore ideas or actions outside the usual game constraints? Did that flexibility feel empowering or did it create confusion?
Usability & System Breakdowns	Did you encounter any moments where the interaction broke down—like irrelevant replies, repetition, or unclear options? How did you respond or adapt when that happened?
LLM vs. Traditional NPCs	Compared to traditional NPCs in similar games, how did these LLM-driven NPCs feel? Were they more responsive, autonomous, or flexible? Or did they fall short in some ways?
Memorable Moments (Good and Bad)	Was there a specific moment that stood out to you—something that felt especially immersive, awkward, surprising, or frustrating?
Personal Fit Based on Player Type	Given that your player type is `$player_type`, your Big Five profile reflects `$big_five`, and your real-world persona is `$persona`, do you feel the NPCs delivered the kind of experience you typically enjoy in games? Or would you have preferred something different?

Table 6. Semi-structured interview prompts covering nine aspects of NPC interaction experience.