A Context-Aware Onboarding Agent for Metaverse Powered by Large Language Models

Jihyeong Hong
z.hyeong@kaist.ac.kr
KAIST
Daejeon, South Korea

Yokyung Lee
ykleeee@kaist.ac.kr
KAIST
Daejeon, South Korea

Dae Hyun Kim
dhkim16@kaist.ac.kr
KAIST
Daejeon, South Korea

DaEun Choi
daeun.choi@kaist.ac.kr
KAIST
Daejeon, South Korea

Yeo-Jin Yoon
yjin.yun@kt.com
KT
Seoul, South Korea

Gyu-cheol Lee
gc.lee@kt.com
KT
Seoul, South Korea

Zucheul Lee
polelee@kt.com
KT
Seoul, South Korea

Juho Kim
juhokim@kaist.ac.kr
KAIST
Daejeon, South Korea

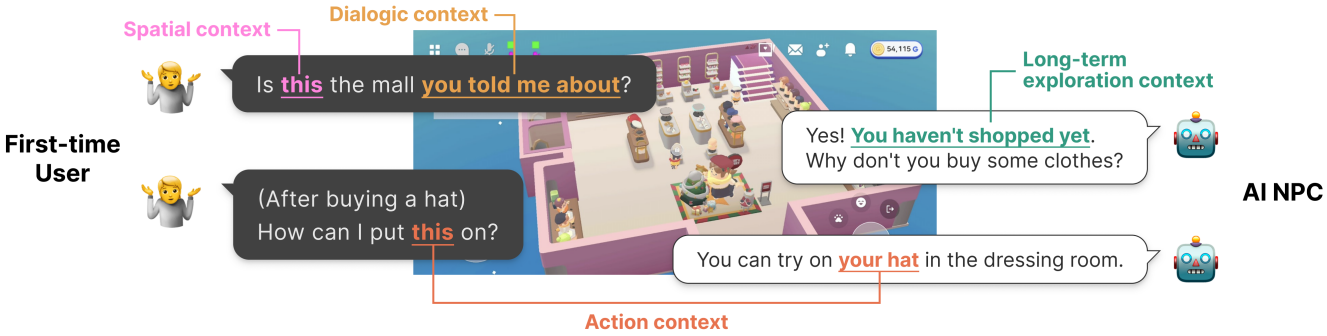


Figure 1: PICAN is an LLM-based pipeline for generating context-aware answers to help users onboard the metaverse. PICAN disambiguates the user question using spatial context, dialogic context, and action context, and goes through relevant user history to customize responses.

ABSTRACT

One common asset of metaverse is that users can freely explore places and actions without linear procedures. Thus, it is hard yet important to understand the divergent challenges each user faces when onboarding metaverse. Our formative study (N = 16) shows that first-time users ask questions about metaverse that concern 1) a short-term spatiotemporal context, regarding the user's current location, recent conversation, and actions, and 2) a long-term exploration context regarding the user's experience history. Based on the findings, we present PICAN, a Large Language Model-based pipeline that generates context-aware answers to users when onboarding metaverse. An ablation study (N = 20) reveals that PICAN's usage of context made responses more useful and immersive than those generated without contexts. Furthermore, a user study (N = 21) shows that the use of long-term exploration context promotes users' learning about the locations and activities within the virtual environment.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; *Natural language interfaces.*

KEYWORDS

metaverse, conversational agent, context-awareness, large-language models

ACM Reference Format:
Jihyeong Hong, Yokyung Lee, Dae Hyun Kim, DaEun Choi, Yeo-Jin Yoon, Gyu-cheol Lee, Zucheul Lee, and Juho Kim. 2024. A Context-Aware Onboarding Agent for Metaverse Powered by Large Language Models. In

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3661579

Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3643834.3661579

1 INTRODUCTION

Metaverse is opening a new chapter of immersive communication and enjoyment, serving a wide range of purposes such as learning, socializing, and gaming. One commonly observed feature of metaverse platforms like Minecraft [34], Roblox [44], and Zepeto [38] is their inherent non-linearity and open-ended nature [23]. They provide virtual spaces where users can freely explore around and try different activities [19, 25, 26, 31], allowing users to have more control over exploration. Such non-linearity of exploration allows creative usage following each user's preference [37], such as decorating their virtual place [39] or laboring for virtual currency [15]. However, users can often feel lost in the abundance of choices. They may not know what to try next, or even achieve milestones without knowing during random explorations [18, 36]. This results in the users of the metaverse often becoming "wanderers" [13], exploring the world with no purpose or destination. They choose what to explore based on their limited understanding of the virtual space, and hesitate to move along because of their unfamiliarity with different strategies and virtual space [48].

As new users of the metaverse must learn a variety of new concepts [28], and as educating them is essential to their engagement and continued usage [3], several types of support have been designed to assist onboarding of such open-world settings. One common method is giving a tutorial with a storyline of predefined order and tasks at the beginning [47]. While this method can make users cover core information, it does not reflect each individual user's interest and cannot respond to subsequent questions the users might have. Another common method is to have an in-world non-player character (NPC) that answers users' questions about the directions toward a destination [12, 22, 50]. While this method can help navigate the metaverse, users cannot ask questions when they do not have any destination in mind.

To help first-time users who are onboarding the open-world metaverse, the system should be designed to observe each exploration and provide corresponding support. In other words, the support should reflect context; the varying information that affects how the intent of user utterance should be interpreted and how responses should be changed to fit into users' current status. However, existing supports lack usage of different contexts of exploration, by restricting user input or using context (e.g., location) for limited purposes. Instead, onboarding supports should be able to adapt to different exploration statuses, or contexts, and provide personalized and on-the-spot help. In this paper, we aim to design an onboarding support pipeline for in-world NPCs, which commonly took roles of instructing users [54], while being aware of user contexts.

We first conducted a Wizard-of-Oz formative study (N=16) to investigate (1) the challenges and opportunities of AI NPC as onboarding support, as well as (2) the types of contextual information required for onboarding support. By asking queries to the NPC who adapted responses to different user contexts, participants could actively explore the metaverse with the freedom of choosing the content and the sequence of onboarding process. Based on the queries that participants asked, we identified the contexts required to understand queries and generate adaptive responses: (1) short-term spatiotemporal context that considers spatial, dialogic, and action contexts to resolve the ambiguity of queries and (2) long-term exploration context to retrieve past relevant actions of the user to create customized responses.

Based on the findings of the formative study, we built PICAN, an LLM-powered pipeline for an interactive NPC that engages in free-form conversation with the first-time users in the metaverse and provides context-aware responses to provide contextual onboarding support. To generate contextual responses, PICAN disambiguates the user question by resolving references utilizing information on the user's location, dialogue, and action, and specifies the descriptions in the response by relating them to the previous experiences of the user.

To validate the correct usage of the context and the positive onboarding effect of our system, we conducted a technical evaluation (N = 20). We compared the full system condition and the conditions without one of the contexts or both. Evaluators found context-aware responses more immersive and useful, especially with short-term spatiotemporal context. We also conducted a user study (N = 21) to observe first-time users' interaction and reaction to the context-aware onboarding AI NPC. We found long-term exploration context could help users familiarize objects and activities in the metaverse by referring to relevant previous experiences.

The contributions of this paper are as follows:

- The requirements and design goals for a metaverse onboarding method that utilizes a free-form conversation with an LLM-powered AI NPC
- PICAN, a pipeline that answers user queries during metaverse onboarding by utilizing short-term spatiotemporal context and long-term exploration context
- The technical evaluation and user study results confirming the importance of short-term spatiotemporal context and long-term exploration context in onboarding assistance

2 RELATED WORK

This work aims to support first-time metaverse user's onboarding process using an LLM-based pipeline to generate AI NPC that can provide customized answers to users' queries based on various user contexts. In this section, we review previous approaches of (1) supporting the user onboarding process in the metaverse, (2) personalized user support in the onboarding process, and (3) applications of language models in virtual worlds.

2.1 Supporting Onboarding Experience in Metaverse

In the context of the expanding metaverse, an increasing number of individuals are engaged to join this new virtual space. As the metaverse possesses characteristics that are different from conventional platforms, newcomers experience distinct user challenges. The key characteristics that encourage first-time users are an open-ended experience with fewer constraints and a wider range of experience. However, at the same time, this can discourage them when their experience is different from their experience, leading to low virtual world self-efficacy and a lack of motivation to enter the metaverse

again [8]. Therefore, guiding user experience based on their expectation is especially important in a metaverse context. Previous work on the metaverse user experience survey highlights technical issues like login troubles, system lag [2], and server setup [43], but Lee and Gu[27] point out a lack of in-depth analysis regarding usability challenges, particularly during onboarding.

Several approaches have been proposed to enhance first-time users' experience in the metaverse platform, often focusing on helping them navigate the virtual environment. Theune et al. [50] developed an agent to help users find out the location by asking clarification queries and making proper gestures. Dijk et al. [12] and Jan et al. [22] also proposed agents that serve as tour guides for non-professional visitors to find their way without previous training. Cosgrove [10] suggested gamification methods to help users get familiar with moving around and navigating the virtual world. However, these approaches focus on navigating to locations, not on the broader spectrum of what users can do in this new environment. Guerra [17] also proposed the need to study multi-purpose agents that can support metaverse users with a wide range of tasks for improving the usefulness of the assistance.

In this context, our research aims to identify what kind of queries first-time users ask frequently and develop a pipeline to properly answer these queries, which can ultimately guide them to a comprehensive onboarding experience.

2.2 Personalized Approaches for User Experience

Previous research has been conducted to assist users in various domains through personalized approaches that take into account their specific context or usage history.

Providing personalized assistance based on the long-term usage history analysis has been a significant focus in this thread. There have been many attempts to make chatbots that generate personalized conversation for user engagement based on the user dialogue history [57, 59]. Bae et al. [6] introduced novel chatbot tasks of managing long-term memory within conversations and providing personal assistance through individual memory storage derived from dialogues. Moreover, Jain et al. [21] investigated the interaction patterns of first-time chatbot users and explored ways to enhance dialogue efficiency by proactively resolving and maintaining context from earlier user messages.

In metaverse, Craftassist [16] was proposed to address ambiguities in connecting user commands to real objects within the Minecraft environment by utilizing dialogue memory and surrounding environments.

Another prevalent approach in personalization is the analysis of the user's current context. Zhang et al. [58] proposed a method to generate an agent's answer personalized to the user based on the user profile. In gameplay situations, personalization is often used to control the difficulties of the game based on the user's current state [41]. For example, Blom et al. [33] proposes a method for personalizing game levels based on facial expression recognition. Similarly, Li et al. [30] suggested a conversational robot that detects user confusion based on their facial expression and adjusts dialogue policies accordingly.

Building upon the findings of prior research, we focused on proposing a personalized agent tailored to the unique demands of the metaverse environment. Given the extensive user freedom within the metaverse, newcomers may struggle with making decisions and understanding how to access required information. These challenges necessitate a substantial degree of personalization. Therefore, we especially focused on providing personalized support for them by addressing both long-term memory on the user's usage log, and short-term user context.

2.3 AI Agents for Virtual Worlds

With the rise of the Large Language Models (LLMs), it has also started to be utilized in virtual worlds, such as role-playing games or metaverse. LLM's ability to generate human-like text has been used to create an agent that generates content based on inputs from developers and users. These agents are used to introduce users to the virtual worlds by generating a narrative [49] or answer to the user's question to introduce the backstories of the MMORPG game [9]. Additionally, AI agents are also utilized to explain the given in-game quests based on object description [51] and the player's request [5]. Ashby et al. [5] utilize knowledge graphs within role-playing games to generate quest descriptions aligned with the game world's current state, leveraging LLM to create quest titles and NPC dialogue. Another approach proposed agents that can do more than present text output, such as executing user commands by generating functions of NPC actions [52].

LLM's ability to generate a comprehensive understanding based on the given text has been also used to build self-exploring agents that can analyze the environments and situations in the metaverse environment described in a text. These agents could learn new skills, make discoveries, and plan their next actions without human intervention by learning from accumulated knowledge found online [14] or using text-based knowledge and memory[60], correcting their behavior by real-time code error feedback [53], and drawing high-level inferences from trivial observations [42]. Language models are also used for generating new narratives in existing video games by understanding environments. Al-Nassar et al. [1] proposed a novel way of generating compelling narratives for tutorial quests within a video game using language models.

However, current applications of LLMs primarily focus on automating tasks that replicate user actions rather than assisting users in getting to know the virtual worlds and make them more engaged. In this paper, we introduce a novel approach to leverage the capabilities of LLMs within virtual worlds. Our approach goes beyond simply comprehending the metaverse environment itself but also incorporates the user's specific context to enhance their overall experience.

3 FORMATIVE STUDY

To understand the value of utilizing NPCs during the onboarding process in the metaverse and to understand the properties of NPCs that users would value during onboarding, we ran a formative study.

3.1 Study Setup

To capture the diverse intent and allow interactions with NPCs without curbing the expectations of the study participants, we ran

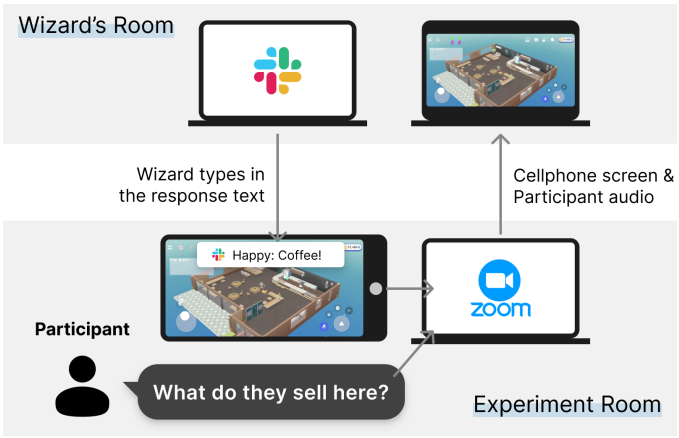


Figure 2: The setup for the formative study. In the experiment room (bottom), the experimenter guides the study with the participant to use a mobile device connected to a monitor. In the wizard's room (top), the wizard (W) receives the mirrored screen and audio through Zoom and types in a response through Slack. This response is sent back to the experiment room as a push notification on the participant's device.

a Wizard-of-Oz study, in which one of the authors acted as a wizard playing the role of an AI NPC. We specifically used *Virtuoville* [1], a 2.5-D mobile metaverse platform, for our formative study. The platform consists of fundamental components of the metaverse such as space to explore, simultaneous interaction with other entities, virtual objects, user avatar, and the ability for the avatar to move around [46]. The platform also allows us to temporarily block other users in the open world for a more controlled setup and includes an NPC that follows the user around which we could use as an imaginary embodiment of the AI NPC, which we named *Happy*.

Prior to the study, the wizard visited every building in *Virtuoville* and explored every interaction with the objects inside to be prepared to generate meaningful responses for any given query. The wizard interpreted the questions along the observation of the user's current screen. When the participant asked for the suggestion of a new place to visit, the wizard suggested one of the places that were not visited yet.

To make the participants think they are interacting with an AI NPC and not a human, we told them that they could start asking queries by saying a wake word *"Hi Happy"*, modeled after those of Apple's Siri [4] or Samsung's Bixby [45]. This wake word also made the queries asked to the NPC distinguishable from the participants' think-aloud comments.

The participants were invited to our experiment room. In the experiment room, an experimenter guided the participants throughout the experiment, providing instructions and leading the interviews; the wizard was present in a separate room (Figure 2). The participants held an Android phone with *Virtuoville* installed and connected to a large monitor for the experimenter to observe. The two rooms were linked via Zoom so that the wizard could observe the participants' mirrored screen and the voice input. The wizard's simulated AI NPC output appeared on the participants' screen as a push notification from Happy.

[1]the pseudonym of the metaverse platform we used

Because the interface of *Virtuoville* is in Korean, we performed the study in Korean to keep the interactions with Happy consistent with the exploration within *Virtuoville*. All quotes in this section are translated from Korean.

3.2 Participants

We recruited 16 participants from the online communities at a university and a company (6 in their 20s, 4 in their 30s, 4 in their 40s, and 2 in their 50s). We only included the participants that have no prior experience in using *Virtuoville* and are familiar with using smartphones. The study lasted around 75 minutes and the participants either chose to receive an equivalent of 11.5 USD via direct deposit or a souvenir. The study was approved by the Institutional Review Board (IRB) at our institution.

3.3 Study Procedure

We first explained the purpose and the procedure of the study, and the participants signed a consent form. We also asked for their permission to record the device screen and voice. While introducing the study, we informed the participants that Happy, the dog shown on the screen, is an AI NPC. We told them that as first-time users, they will learn about *Virtuoville* while exploring by themselves and asking Happy any query when they want to. As a short tutorial, we had the participants ask Happy the two example queries related to the onboarding on the metaverse, *"Did you have lunch?"* and *"How can I move forward?"*.

After the introduction, we asked the participants to share previous experiences with metaverse or similar platforms like video games, especially regarding how they learned to use the platforms. The participants then explored *Virtuoville* for 30 minutes while interacting with Happy. We first asked the participant to find the virtual home and decorate it. For the remaining time the user freely explored *Virtuoville* while asking questions to the NPC. To focus on the queries that appear during the onboarding phase where users explore diverse activities before deeply engaging in a few particular

ones, the experimenter nudged the participants to move on to a different activity by providing tasks (e.g., "Leave home and explore the village", "Earn money in the farm") if the participants remained at the same location for more than 10 minutes.

Once the exploration phase of the study was over, we performed a semi-structured interview. We mainly asked about their overall onboarding experience with Happy and about the situations in which the participants deemed NPC-based onboarding effective or ineffective. To understand the desirable traits of AI NPCs during onboarding, we also asked the participants to recall their interactions with Happy and discuss the suitability and quality of the NPC responses and potential improvements. The first author categorized the questions from participants and discussed samples and definitions of each category with other authors for finalization.

3.4 Results

Based on the analysis of the participants' interactions with Happy and the interview contents, we found that AI NPCs can assist active exploration during onboarding with contextual support. We also defined specific contextual supports required: (1) Understanding short-term spatiotemporal context that considers spatial, dialogic, and action contexts to resolve the ambiguity of queries, (2) Understanding long-term exploration context that considers the user's past activities to retrieve past relevant actions and create customized responses, and (3) Initiating conversation in the context where the user is stuck or lost during the exploration.

Overall, all 16 participants (P1-16) interacted heavily with Happy during the 30-minute study, and a total of 404 queries were collected (M = 25.25, Std = 19.12). The participants asked a wide range of queries. These included the basic description (e.g., *"What does this symbol mean?"* (P5)), the location of certain buildings and activities (e.g., *"Where can I buy furniture?"* (P2), *"Where is a clothing store?"* (P13)) and the instructions for activities (e.g., *"How can I take pictures?"* (P7)). Participants also frequently asked for confirmation (e.g., *"Did I come to the right place? I am bad at finding places."* (P2)) and suggestions (e.g., *"What should I do now?"* (P14)).

Eight participants appreciated the efficiency of the process of getting responses, meaning they could get responses on the spot without additional steps like going through tutorials or searching online. For instance, P14 described that being able to get assistance while staying in the metaverse increased the immersion compared to previously going back and forth between the virtual world app and the YouTube tutorial videos.

Six participants appreciated having more control over which information they got from the NPC based on their own context of exploration. For instance, P3 mentioned that since different users can ask for different help, they can design their own exploration, instead of following steps pre-assigned by the metaverse.

3.4.1 Finding 1. Understanding short-term spatiotemporal context is important for interpreting queries during onboarding. Of the 404 queries collected through the formative study, 98 (24.26% out of all queries, M = 6.125, Std = 7.63) included references to information that is spatially or temporarily proximate to the user.

Out of the queries, 53 (13.12% out of all queries, M = 3.31, Std = 3.57) queries included references to spatial context, concerning the current location or the nearby objects (e.g., *"Is **this door** closed?"* (P9), *"Can't I buy the seed **here**?""* (P15)). As some participants (P1, P9, P12) mentioned, the main usage of the AI NPC was asking about the new places they visited. Since visiting different spaces is a frequent activity during exploration, the queries should be interpreted with the understanding of spatial context to reduce redundant descriptions.

Next, 29 (7.18% out of all queries, M = 1.81, Std = 3.15) queries included references to dialogic context, concerning the previous recent dialogues (e.g., NPC: *"These are apartments and houses."* P9: *"Do **they** become my house if I decorate **them**?"* / NPC: *"You can play games alone, too."* P2: *"**Which one** can I play alone?"*). Dialogic context were required to answer follow-up queries that diverged according to participants' different interests and knowledge.

Finally, 21 (5.20% out of all queries, M = 1.31, Std = 2.24) queries included references to action context, concerning the recent activities such as buying, building, and selecting (e.g., *"Is **this** a lamp?"* (P12) after selecting an object, *"Do **these shoes** look good?"* (P16) after changing clothes). Since users actively interact with the environment in the metaverse, the onboarding-support AI NPCs have to observe the user in real time and refer to each action to describe it further.

Some of the queries referred to multiple types of contexts. For instance, the quest *"Is **this** the garden **you mentioned**?"* (P3) refers to both spatial context and dialogic context.

Spatial context, dialogic context, and action context are all essential in understanding the user query, and we conceptualize them together as short-term spatiotemporal context. We identify them to be "short-term" because they consider the current or recent context.

The onboarding-support AI NPCs should identify the queries that refer to short-term spatiotemporal context, pinpoint the referred objects, and give relevant responses.

3.4.2 Finding 2. Understanding long-term exploration context is important for generating customized responses. Of the 404 queries collected, 26 (6.44% out of all queries) asked for suggestions on the next destination or activity. Examples include direct queries (e.g., *"What should I do next?"* (P15)) and asking for an alternative (e.g., *"What can I do else than placing the furniture?"* (P14)). The latter queries had to be answered considering both the previous experiences and the constraints mentioned in the query. To foster a wide variety of exploration by suggesting new activities or objects, these queries require memory of what the participants have already experienced, which we define as long-term exploration context.

Long-term exploration context could also be used to answer queries about facts, mainly when the response from NPC discusses topics of which the participants might have previous knowledge. For instance, the wizard answered the query *"How can I plant other seeds?"* (P6) by informing the location of a seed market, since the participant had planted a few of the possible seed choices. However, the query may be answered differently based on related previous explorations. If they have not planted any seeds, the response may include information about both buying seeds and planting them. If they have planted seeds before, the response could specify the undiscovered types and where to get them. If they have planted every type of seeds available, the response could inform that there is no other type available, and possibly introduce new activities. Similar consideration of long-term exploration context needed to

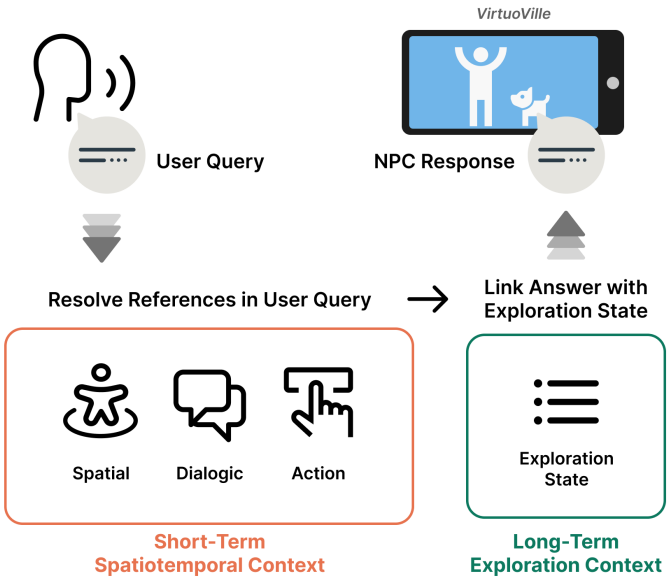


Figure 3: An overview of PICAN. The user utterance or action description is passed as input, short-term spatiotemporal context and long-term exploration context are applied to process the query and generate a response.

be made answering queries like *"Where can I change clothes?"* (P3) and *"What is Event hall for?"* (P16). The answer to the first query depended on places for changing clothes the participant had already visited, and the answer to the second query depended on which functions of the Event hall the participant had already experienced.

3.4.3 Finding 3. NPC needs to initiate conversation while considering the context. The AI NPC in the formative study never initiated a conversation and only responded to the participants' queries. However, participants suggested that the AI NPC sometimes proactively reach out to them and initiate conversation, especially when they feel lost. Some participants (P4, P5, P11, P12, P16) suggested that the NPC proactively introduce the possible activities in a nearby building. Some participants (P2, P8, P13, P14) wanted proactive help from the NPC when they were having hardships during exploration, such as being inactive or repeating same actions. However, as mentioned by P6, P7, P8 and P9, we identified that when the user is engaging in activities, the NPC's proactive approach is unnecessary and can be disrupting.

The onboarding-support AI NPCs have to proactively initiate the conversations to guide first-time users who do not have prior knowledge of the metaverse and the NPC. However, it should carefully observe the context to not disrupt the open exploration in the metaverse where users have high freedom and can constantly change their goals.

4 PICAN: PIPELINE FOR INTERACTIVE CONTEXT-AWARE NPC

Based on the findings from the formative study, we propose PICAN (Figure 3), a two-stage pipeline powered by LLMs that generates a context-aware and exploration-promoting response to a user's query during the onboarding phase. PICAN processes the input speech-to-text converted user query that either directly asks for suggestions purely dependent on the NPC's biases or for factual information of *Virtuoville*. In the first stage, PICAN resolves context-dependent references to objects, locations, and interactions using spatial, dialogic, and action contexts from the input query. PICAN reformulates the query with the resolved references to support a proper understanding of the query during subsequent processing. Then, in the second stage, PICAN utilizes the user's exploration state available in the long-term exploration context to generate suggestions that promote further exploration and reword the response in terms of prior exploration to reinforce knowledge gained from previous explorations.

We built our pipeline and optimized our prompts based on GPT-4-1106 model with 0 temperature[40]. To provide information about *Virtuoville* to the LLM, we created a *Virtuoville* knowledge document based on official documentation provided by the creators of *Virtuoville* and our test uses of *Virtuoville*. The document contains the names and characteristics of the locations, objects, features, and functions in *Virtuoville*. Because incorrect answers can negatively impact the onboarding process and lower the user's trust in the NPC, we applied the retrieval-based augmentation generation technique based on this *Virtuoville* knowledge document to keep the generated responses faithful to the provided document while avoiding hallucinations [29]. We include the *Virtuoville* knowledge document and prompts we used within our pipeline in the supplemental materials.

4.1 Data Logging and Input Processing

PICAN continuously collects and logs information that it can use as short-term spatiotemporal context and long-term exploration context. In particular, it keeps logs of (1) user location (i.e., building, coordinates; Figure 6 (b)), (2) user actions (e.g., *"plant seed"* , *"move furniture"* ; Figure 6 (d)), (3) exploration history (i.e., whether the user has visited certain locations and whether the user has performed certain action; Figure 6 (e)), and (4) conversation history (Figure 6 (c)).

Users interact with PICAN through speech input, whose start and end points are specified via pressing start (Figure 9 (a)) and end (Figure 9 (b)) buttons. Because user input is not necessarily in the form of a query (e.g., *"I don't want to go to the Virtumall."*), our pipeline uses an LLM to formulate it into a query with clear information request intent (e.g., *"Could you recommend me somewhere to go to that is not Virtumall?"*) so that our pipeline can generate responses that are informative in nature instead of sympathetic responses (e.g., *"I feel sorry that you don't want to go to the Virtumall. I am sure you will still find interesting things there."*)

Even when the user does not provide direct input, PICAN monitors the user logs to detect predefined patterns in which it proactively intervenes. Our pipeline specifically detects periods of inactivity or lack of interactions (empirically set to 2 minutes) and situations when the user is performing repetitive actions without changes to user attributes such as changes in the user's inventory status or coin count. For example, if the user moves back and forth

around a wall looking for an entrance, PICAN feeds a query for suggestion as input (*"What should I do now?"*). We exclude repetitive actions with changes to user attributes since users often purposefully engage in repetitive actions to *farm* items or coins.

4.2 Stage 1. Reference Resolution Based on Short-Term Spatiotemporal Context

PICAN first utilizes the short-term spatiotemporal context to resolve references to various objects and interactions so that LLM is able to produce an accurate response based on the query and the *Virtuoville* knowledge document. This process consists of three steps: (1) extending recent dialogue-dependent queries (2) detecting references that need resolving and (3) resolving detected references.

4.2.1 Step 1. Extend query based on recent dialogue. First, the dialogic context is used to check if there is no crucial information missed without hints that can be detected in reference detection (Step 2). PICAN uses the LLM module to infer the omitted detail by referring to the most recent user query and NPC response.

4.2.2 Step 2. Detect references to objects and interactions. PICAN detects all words in the input query that make references to objects and interactions such as demonstratives (e.g., this, that), possessives (e.g., its, their), and locative adverbs (e.g., here, there). In our example in Figure 4, it detects the word *'this'*. In addition, we detect comparative words (e.g., other, else) if they do not contain the comparison target. For example, in the query *"Are there any other places I could go to?"*, the specific place for comparison is not specified, which makes it challenging for our pipeline to determine which place should be excluded in the response.

4.2.3 Step 3. Resolve Detected References. To resolve the reference words detected in Step 2, PICAN uses the short-term spatiotemporal context to replace or specify each of the words with its actual referent. For spatial context, our pipeline retrieves the user's current location from the log and lists all objects within the visibility radius, the longest distance visible on the device screen, as candidates of the correct referent of the reference word (*Spatial* in Figure 4). For dialogic context, our pipeline assumes that the most recent query-response pair includes information about the correct referent (*Dialogic* in Figure 4). For action context, our pipeline retrieves the 10 most recent user actions from the log as candidates of the correct referent (*Action* in Figure 4). Based on the candidates, our pipeline prompts an LLM to select the candidate that can replace the reference word in the query only if it can provide logical reasoning in making the decision, leveraging the commonsense knowledge possessed by LLMs (e.g., the knowledge that trees are green when resolving *"this green object"*). In our example in Figure 4, "this" is a demonstrative that refers to an object nearby when it is placed at the end of a query, the nearest object is most likely to be the referent. If the resolved reference is the direct answer to the query, the model appends details to the query to preserve the user's original intent (e.g., User: *"What is **this**?"*, Referent: "apple tree", Query Processor: *"What is **this tree**?"*). For example, if the user asks *"What is this?"* but the model replaces the referent with *"What is apple tree?"*, the subsequent answer generation models may return a description of the referent rather than answering what the referent is. When the reference is ambiguous and PICAN cannot resolve it, the pipeline requests the user to provide additional details in their query by responding, *"I'm sorry I don't understand what action you are referring to. Please approach me again when you have the details!"* The query resulting from this step is fed directly into an LLM with the *Virtuoville* knowledge document to generate a preliminary response (e.g., User: *""What is a Coin?""*, Preliminary Response: *""Coin is the currency used in Virtuoville for purchasing items""*).

4.3 Stage 2. Suggestion Generation & Response Rewording Based on Long-Term Exploration Context

PICAN responds to queries asking for suggestions by recommending objects or interactions that have not been explored to promote exploration and responds to queries asking for information by linking the information to prior explorations to reinforce the information. PICAN first classifies the query as either a suggestion query or information query using chain-of-thought prompting to prompt an LLM to perform logical reasoning on whether a query is dependent on the NPC's personal opinions, preferences, or judgments (suggestion query) or not (information query). When the query is classified as an information query, the preliminary response from Stage 1 Step 3 is utilized.

4.3.1 Generating Recommendations for Suggestion Queries. For a suggestion query, PICAN takes the reference-resolved query from Stage 1 Step 3 and generates a recommendation that provides a high exploratory value and is easily accessible. We note that users often impose constraints when asking for suggestions. For instance, the suggestion query *"What should I do next that is **not in the Virtumall**?"* imposes two constraints: 'action' and 'not in the *Virtumall*'. Hence, PICAN prompts an LLM to filter the list of explorable actions, retaining only those that meet the constraints specified in the query.

Next, it scores the possible explorable actions by summing the ranks of (1) the number of times the action has been performed and (2) the proximity of the action from the user's current location. If the action is explored more times and can be performed closer to the user than other actions, it has a higher score. Based on the scores, PICAN generates a response that recommends the exploration with the highest score. In our example (Figure 5 (b)), the highest scoring action 'harvest Tree' is returned.

4.3.2 Response Rewording for Information Queries. When the user's query is an information query, PICAN modifies the preliminary response from Stage 1 Step 3 to incorporate the user's exploration state thus far. Our pipeline first uses an LLM to extract the (1) location, (2) action, and (3) (action target) object described in the reference-resolved query and the preliminary response. In our example (Figure 5 (a)), the query *"What is this tree?"* and *"That is an apple tree"* would extract *"Virtufarm"* and *"apple tree"* for location and object (there is no action involved). Then, PICAN filters the exploration state log for those with the same location, action, and object as relevant to answering the query. For the previous example, all explorable actions that take place in Virtufarm (e.g., Plant, Harvest, Water) and involve an *"apple tree"* (e.g., Plant, Harvest, Water) are selected. Based on the relevant exploration states, PICAN

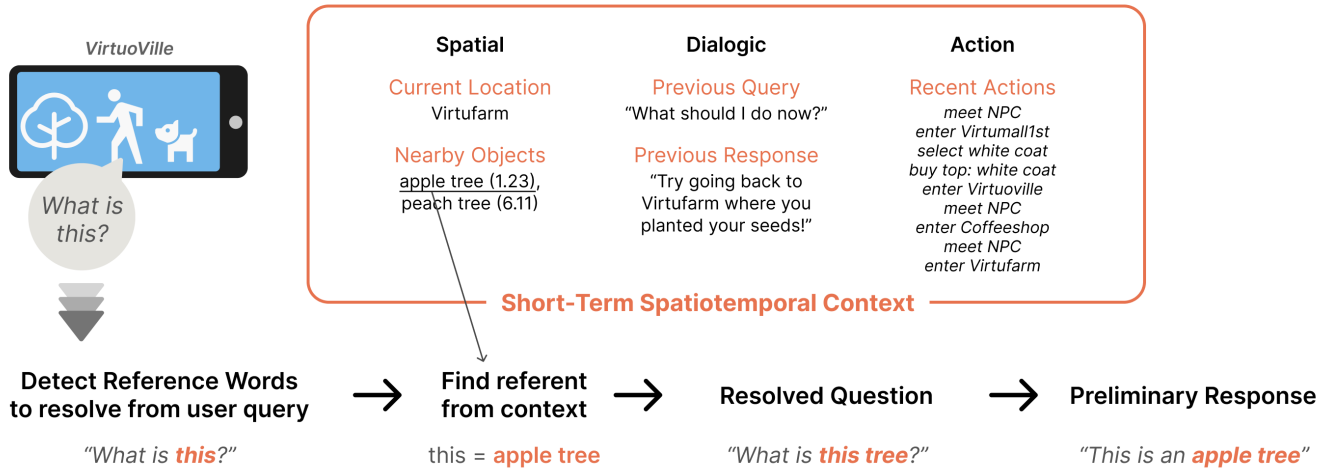


Figure 4: Example of the user query *"What is this?"* passed to the reference detector and resolver models to resolve "this" using short-term spatiotemporal context. The resolved query *"What is this tree?"* is passed to an answer-generating model to generate a direct answer with only short-term spatiotemporal context

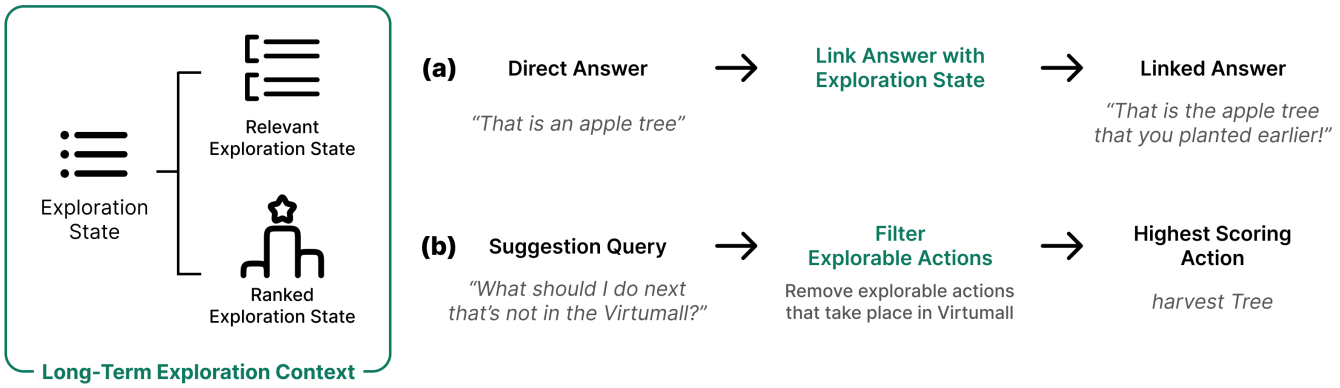


Figure 5: Long-term exploration context is applied differently for an information query and suggestion query. Our pipeline either extracts the exploration state by relevance or ranks them by their relative distance and explore count. (a) For information query, the relevant exploration state is linked to the answers generated for information queries in the previous stage. (b) For suggestion query, the explorable action is retrieved based on the ranked exploration state, and filtered based on the query.

rewords the response to reflect the relevant contexts whenever possible by using an LLM. For our example, the original response "That is an apple tree." is rephrased to "That is an apple tree that you planted earlier!" using the exploration state: *"Plant"* - *"apple tree"*. Our pipeline considers the number of times a user performs an explorable action to reflect the user's knowledge of it.

4.4 Output Presentation

To ensure that the presented output is consistent with the guide NPC's persona while avoiding unnecessary repetitions or jargon in the output, PICAN prompts an LLM to reword all outgoing output such that they (1) sound like a natural, friendly, and personal response from a cheerful guide, (2) do not sound similar to a response that has been generated before, (3) answer the user's original query directly and clearly with no unnecessary information and (4) do not use any terms or concepts that may be unfamiliar to the user. Through this process, the response "That is the apple tree that you planted earlier!" in Figure 5 (a) would transform to a response like "That is the apple tree that you brilliantly planted earlier! Keep going!"

5 PIPELINE EVALUATION

To evaluate the performance of our PICAN, we ran a human-based pipeline evaluation by comparing responses generated from five different conditions.

- **PICAN condition**: Responses generated with PICAN, including the short-term spatiotemporal context and the long-term exploration context,

- **No long-term context condition**: Responses generated with PICAN without long-term exploration context (Stage 2),
- **No short-term context condition**: Responses generated with PICAN without short-term spatiotemporal context (Stage 1)
- **No context condition**: Responses generated without short-term spatiotemporal context (Stage 1) and long-term exploration context (Stage 2), and
- **Non-LLM baseline condition**: Responses generated with a BERT-based question-answering pipeline [11]. This condition was included to observe the benefit of utilizing LLM for question-answering tasks in the metaverse. based on the *Virtuoville* documentations provided to the LLM models. We fed the documentation about the knowledge of *Virtuoville* to the LLM modules in PICAN as a context for answering questions.

The pipeline evaluation was conducted to test the following two hypotheses:

H1. The two context modules of PICAN produce accurate responses and utilize relevant contexts.

H2. The two context modules of PICAN generate responses that convey useful information and promote immersion.

Regarding H2, We evaluated immersion to see whether PICAN motivates users during the onboarding process and assessed usefulness to measure the effectiveness of PICAN in the onboarding process.

5.1 Study Design

We randomly sampled four user queries per session for a total of 40 queries from ten pilot study sessions. The pilot study was conducted for the purpose of data collection and system improvement and followed the exact same procedure as the user study outlined in Section 6. For each of the 40 queries, we generated five versions of responses according to the five conditions with short-term spatiotemporal context and long-term exploration context gathered during the study. To reduce the task load per evaluator, we split the 40 queries into four bins of 10 queries and presented one bin of 10 queries in a randomized order to each evaluator. For each query, we presented the responses in a randomized order without any indicators of the conditions to avoid potential biases. To ensure the response integrity, each evaluator answered two attention check questions related to the context material (What is the current location of the user character? How many times has the user character visited the farm so far?).

We first asked evaluators to go through two documents; 1) a document about buildings and NPCs of *Virtuoville*, and 2) the format of the contexts used by our system shown in Figure 6, such as how to interpret the given long-term exploration context. Evaluators were required to answer three comprehension questions while going through the document. Then, for each of the 10 queries the evaluator is assigned to, we showed each evaluator the query, spatial context, dialogic context, action context, long-term exploration context, and the five responses generated for the five conditions (Figure 6). Based on the provided information, the evaluator gave each response usefulness rank and immersion rank (rank 1 to rank 5, comparing five responses for the corresponding query). They also provided a short justification with respect to the top two and the bottom two responses.

After looking at the query, the evaluator looked at the answer and checked whether each of the responses utilizes short-term spatiotemporal context and long-term exploration context respectively; choosing one between "correct usage", "incorrect usage", and "no usage". The evaluators performed the evaluation asynchronously online through Google Forms. We presented the query-response pairs in English to remove the evaluation's dependency on translation accuracy and focus more on the evaluations of the core contributions of our work.

The context usage of each response was decided based on the majority vote between 5 evaluators designated to the response. Among 200 responses for each context, 18 responses for short-term spatiotemporal context and 20 responses for long-term exploration context did not have a majority vote (e.g., two evaluators saying "incorrect usage", two evaluators saying "no usage"). Therefore, we recruited one more evaluator for each of the contexts. For 4 responses for short-term spatiotemporal context and 1 response for long-term exploration context that still did not have a majority vote, the evaluator who previously covered the different context gave additional decisions.

5.2 Evaluators

We recruited evaluators through online communities of universities. We only included evaluators that have sufficient English fluency to analyze the English query-response pairs. Out of 27 applicants, 20 finished the tasks, resulting in having 5 evaluators for each query. Each session lasted roughly 75 minutes, and the evaluators received an equivalent of 19 USD via direct deposit. This evaluation process was approved by Institutional Review Board (IRB) at our institution.

5.3 Results

H1. The modules of PICAN make accurate and relevant usage of context. The PICAN condition made the most correct usage of both contexts. Figure 7 shows how many responses generated from each condition used short-term spatiotemporal context or long-term exploration context correctly, incorrectly, or did not use the context.

Short-term spatiotemporal context-aware conditions (The PICAN condition (12 responses) and the no long-term context condition (10 responses)) made more correct usage of short-term spatiotemporal context than those unaware of the context (The no short-term context condition (5 responses) and the no context condition (6 responses)). Short-term spatiotemporal context-aware conditions were able to respond correctly to the question that says "here", or asks about the nearby object (e.g., *"Is this friend Sam (NPC)?"*) by referring to the user location. Short-term spatiotemporal context-unaware conditions were often considered to have used the context correctly when the subject matter of the question was not in the scope of short-term spatiotemporal context. For instance, when a user asked *"Is Sara (NPC) near me right now?"* when Sara is far away, all conditions except the non-LLM condition created a correct response that said: *"Sara cannot be found in the vicinity"*. We ran significance test between correct usage and combined count

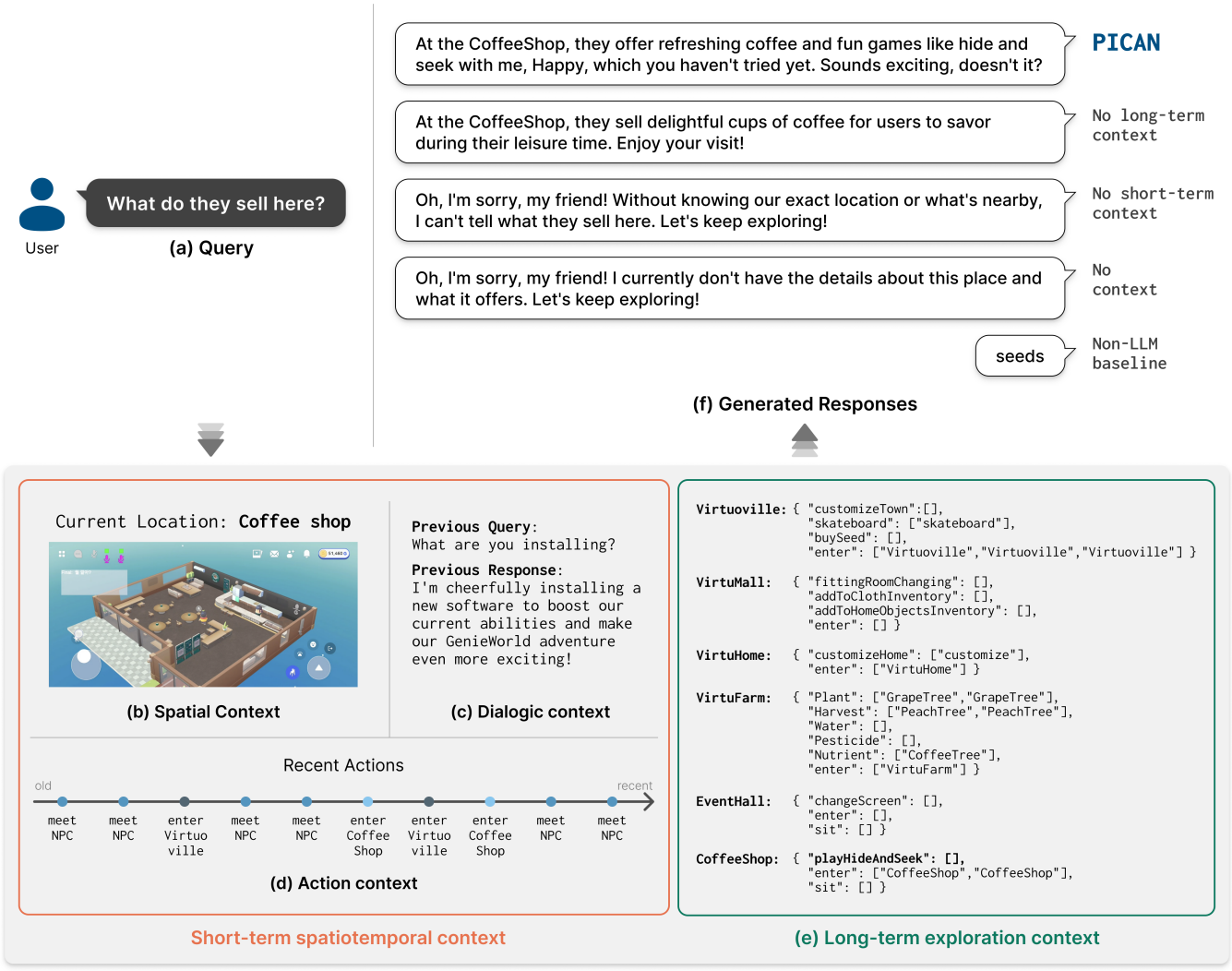
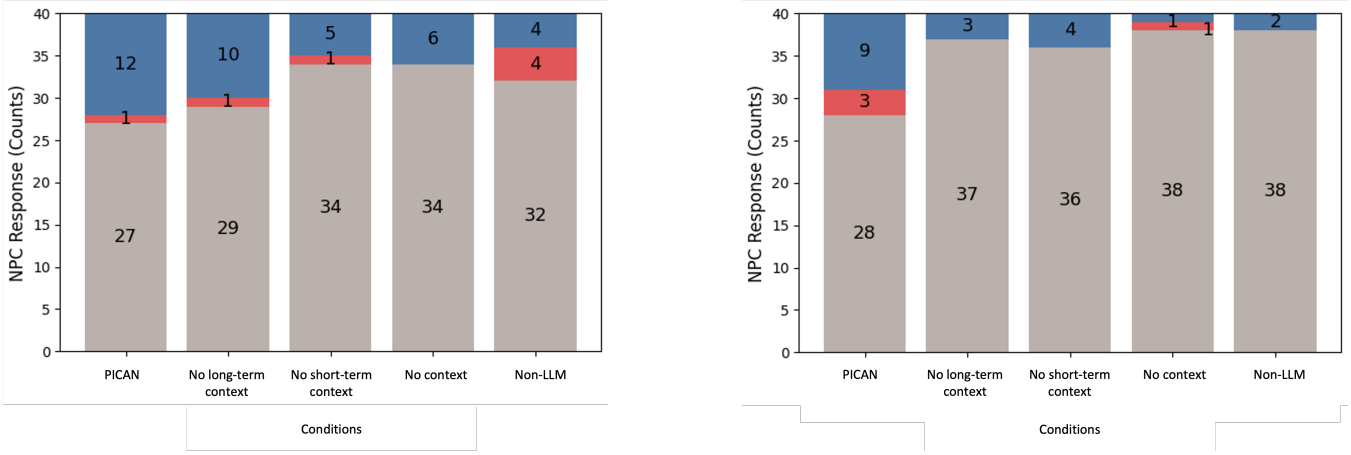


Figure 6: An example of information provided to the evaluator for each query. We provide (a) the query provided by the user to initiate a conversation, (b) short-term spatiotemporal context, displayed as (b-1) the location and (b-2) screenshot, (c) dialogic context, or the immediate conversations preceding the query, (d) action context, or the most recent actions, (e) long-term exploration context, and (f) the responses generated for the five conditions. In this example, the spatial context ((b)) saying that the user was in the Coffee shop shows that the PICAN condition used short-term spatiotemporal context correctly ("At the Coffee shop,"). Also, the fact that the player has not tried to hide and seek before ("playHideAndSeeks":[] in (e)) shows that the PICAN condition utilized the long-term exploration context correctly ("hide and seek, which you have not tried yet").

of incorrect usage and non usage. Cochran's Q test determined that there was a statistically significant difference between conditions (Q=11.8, p=0.019). The posthoc pairwise Mcnemar test with Bonferroni correction showed only showed significant difference between pair of PICAN condition and Non-LLM condition (p = 0.0468), potentially due to low sample count.

Regarding the usage of long-term exploration context, the PICAN condition made the most correct usage among all conditions (9 responses). It generated responses that referred to the related past activities of the user to describe the places (e.g., *"Indeed, Sara is at Virtumall, the place where you've been busy expanding your home objects collection!"*), or suggested activities that the user has not tried. The no short-term context condition showed relatively less usage of the context (4 responses) than the PICAN condition (9 responses) because some queries necessitate short-term spatiotemporal context. For instance, when a user asked *"What do they sell here?"* in the coffee shop, the PICAN condition could utilize both contexts and responded *"At the CoffeeShop, they offer refreshing coffee and fun games like hide and seek with me, which you haven't tried yet."*, whereas the no short-term context condition responded that it does not know where the user is referring to. However, the PICAN condition also made most incorrect usages (3 responses) such as



(a) Short-term spatiotemporal context usage

(b) Long-term exploration context usage

Figure 7: How the responses generated for each condition utilized (a) short-term spatiotemporal context or (b) long-term exploration context based on a simple majority vote of the evaluation evaluators. Blue ■ bars indicate that the system used the context *correctly*, red ■ bars indicate that the system used the context *incorrectly*, introducing incorrect information, gray ■ bars indicate that the system did not use the context.

retrieving irrelevant context. For instance, when a user asked about the name of a blue-haired NPC in front of *Virtumall*, the PICAN condition mentioned the previous farming experience of the user. We ran significance test between correct usage and combined count of incorrect usage and non usage. Cochran's Q test determined that there was a statistically significant difference between conditions (Q=12.125, p=0.016). The posthoc pairwise Mcnemar test with Bonferroni correction showed only showed significant difference between pair of PICAN condition and No context condition (p = 0.0468), potentially due to low sample count.

Overall, the portion of responses that did not have any context was high even for the PICAN condition (27 for short-term spatiotemporal context and 28 for long-term exploration context) because we randomly selected queries from the user study where participants of the study frequently asked about basic facts in the metaverse that do not consider the contexts.

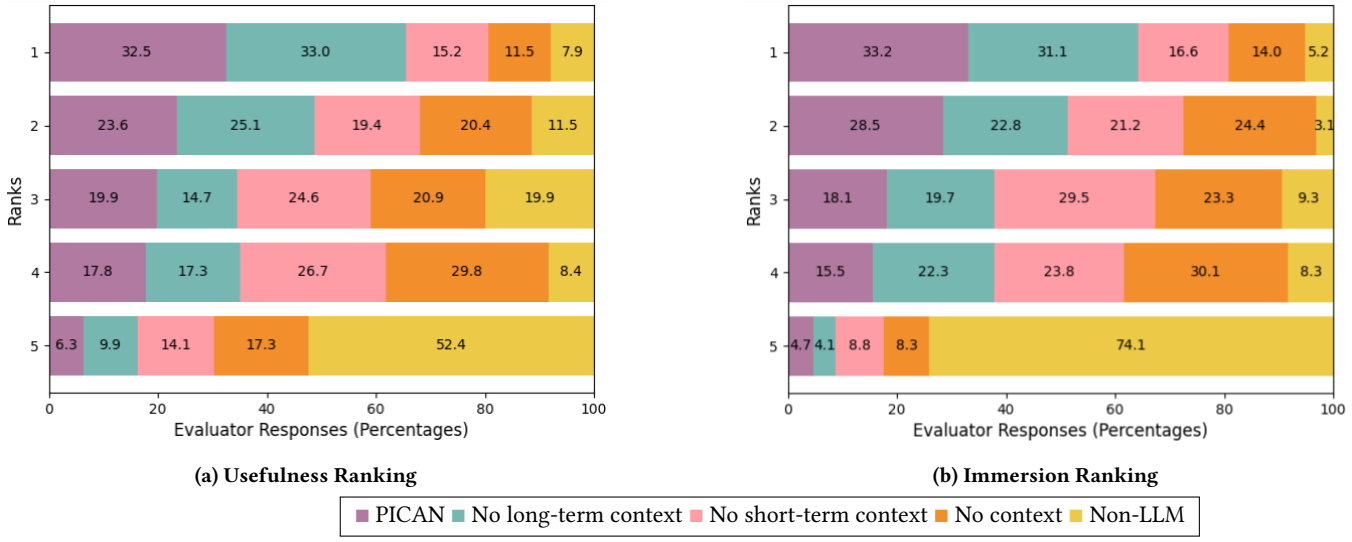
H2. The modules of PICAN make responses that are useful and immersive. Figure 8 shows the composition of the usefulness and immersion rankings by each condition. Nine responses for usefulness and 7 responses for immersion ranking were removed due to reporting multiple ranks for a condition. Regarding usefulness, Friedman's test indicated a statistical difference (χ^2 = 107.91, p <.001) among the usefulness rankings of the five conditions. A posthoc analysis using Nemenyi's test found significant differences in all the pairs of the conditions with short-term spatiotemporal context (the PICAN condition and the no long-term context condition) and the conditions without it (the no short-term context condition and the no context condition) (Table 1a). The average rank (the smaller the number is, the higher the ranking is) for each condition was; PICAN: 2.42, No long-term context: 2.46, No short-term context: 3.05, No context: 3.21, Non-LLM: 3.86. The conditions with short-term spatiotemporal context scored the highest average ranks, with the PICAN evaluated as the most useful. Evaluators commented that the responses of PICAN condition were more useful because they were accurate and more detailed.

Regarding immersion, Friedman's test indicated a statistical difference (χ^2 = 220.05, p <.001) among immersion rankings of the five different conditions. A posthoc analysis using Nemenyi's test found significant differences in pairs "the PICAN condition and the no short-term context condition", "the PICAN condition and the no context condition" and "the no short-term context condition and no context condition" (Table 1b). The average rank for each condition was; PICAN: 2.30, No long-term context: 2.46, No short-term context: 2.87, No context: 2.94, Non-LLM: 4.43. The conditions with short-term spatiotemporal context scored the highest average ranks, with the PICAN evaluated as the most immersive. Evaluators commented that the responses of PICAN condition were more immersive because they "feel like they are actually talking to the user". The accuracy of the information in the responses was also considered important for immersion.

For both usefulness and immersion, short-term spatiotemporal context played a more significant role than long-term exploration context. The non-LLM condition ranked significantly lower compared to the other conditions. Based on the evaluators' comments, a major reason was due to giving short responses often not in full sentences. Although the ability to generate sentences that sound natural and utilize given information is a strength of LLMs as seen in our pipeline evaluation result, a more thorough comparison with non-LLM techniques might better demonstrate the effect of LLMs.

6 USER STUDY

The pipeline evaluation focused on the individual responses generated by the PICAN and their context usage, where evaluators rated usefulness and immersion. To observe how users make use of context-awareness of PICAN and which aspects of context-aware responses affect user's satisfaction, we designed a user study to



(a) Usefulness Ranking **(b) Immersion Ranking**

Figure 8: The rankings of the response broken by each condition for (a) usefulness and (b) immersion. For example, the top left purple ■ bar on (a) Usefulness Ranking means that 62 responses that ranked best for usefulness are from the PICAN condition.

Table 1: The pairwise p-value between (a) usefulness ranking and (b) immersion ranking of different conditions calculated by the Nemenyi test. The pairs with significance (p <.05) are marked with an asterisk(*) and bolded

	1.	2.	3.	4.	5.
1. PICAN	-	.900	**.001***	**.001***	**.001***
2. No long-term context		-	**.002***	**.001***	**.001***
3. No short-term context			-	.854	**.001***
4. No context				-	**.001***
5. Non-LLM					-

(a) Usefulenss Ranking Pairwise p-value

	1.	2.	3.	4.	5.
1. PICAN	-	.857	**.004***	**.001***	**.001***
2. No long-term context		-	.075	**.021***	**.001***
3. No short-term context			-	.900	**.001***
4. No context				-	**.001***
5. Non-LLM					-

(b) Immersion Ranking Pairwise p-value

observe how users interact with the AI NPC managed by PICAN as they explore the metaverse.

As with the formative study, the interactions with the AI NPC were in Korean, the language of the interface in *Virtuoville*. To pipe the voice input into the PICAN, we first used the Google Cloud text-to-speech tool to transcribe user input and GPT-4 API call to translate this input into English, the language of the pipeline, and then to translate the responses back into Korean. Because the tools for transcription were often inaccurate (e.g., transcribing "coin" as "go-in" (meaning "the deceased" in Korean)), we displayed the transcript to the user for transparency. Since the transcription and translation errors are irrelevant to the core contributions of our work, we cautioned the participants about these errors and asked them to focus on the contents of their queries and the NPC's responses.

6.1 Participants

We recruited 21 participants (U1–U21) using an online community at a university. As with the formative study, we only included the participants who have not used *Virtuoville* before and are comfortable with using smartphones. In addition, we added the requirement that the participants have prior experience with onboarding in any virtual world (e.g., tutorial of video games, beginner's quests in social metaverse systems, etc.) so that they have a point of reference when experiencing onboarding with AI NPC. The study lasted 90 minutes and we compensated the participants an equivalent of 15 USD via direct deposit. The study was approved by Institutional Review Board (IRB) at our institution.

6.2 Procedure

We first described the purpose and the procedure of the study to the participant and asked for permission to record the device screen, the participant's voice, and any logs generated within the system during the study. Next, we performed a pre-interview about prior experiences with metaverse or other virtual worlds as well as their mental model of the NPCs within the platforms. After the pre-interview, we introduced the mobile interface (Figure 9) and guided the participant to try buttons that start (🎙) and end (🎙) the recording of their queries.

We first gave Task 1, where the participant interacted with our AI NPC through three example queries aimed at giving them a sense of its capabilities: a query about in-world coins, a query about a specific building in *Virtuoville*, and a query about recommended next steps. We then had the participants freely explore *Virtuoville* with interactions with our intelligent NPC for the first 20 minutes. After the 20 minutes, to induce experiencing outputs that incorporate longer-term experiential context from the first 20 minutes,

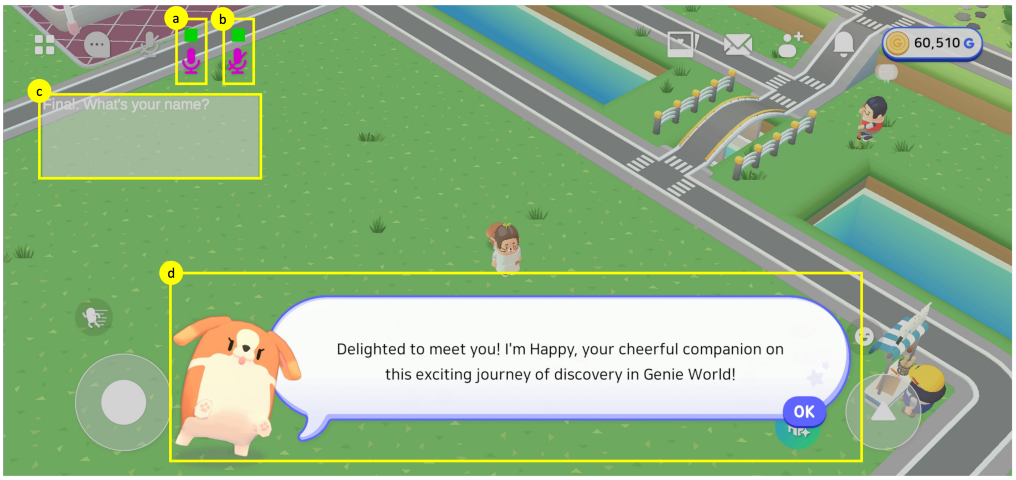


Figure 9: The screenshot of the user study interface. (a) A button that starts listening to the query; (b) A button that ends listening to the query; (c) the transcript of the user query; and (d) the response from the NPC.

we gave Task 2 about searching for a specific NPC somewhere in the *Virtuoville* world. If the participant did not ask any queries that involve short-term spatiotemporal context until the end of the first search of an NPC, the experimenter hinted by saying "The AI NPC knows about your surroundings and status", or "Could you *check* if that is the NPC you were indeed looking for?" This was to give the user experience of getting context-aware answers so that we could get related feedback in the post-interview.

Finally, we conducted a semi-structured interview, focusing on their overall experience, especially concerning their prior onboarding experience. We also asked for qualitative feedback on our AI NPC and how it uses context, as well as possible improvements and potential future usages.

6.3 Interaction Analysis

We analyzed the logs of user questions, NPC responses, and transcript of post-interview. We analyzed the user queries that followed NPC response and post-interview data to identify the detailed cases of satisfaction and dissatisfaction in context usage. Regarding the usage of short-term spatiotemporal context, we considered a response a missed case if the user repeated a query about the same object. We considered a response successful if the user extended the dialogue about the same object by adding on to the response, but not asking for the same thing. This was to see whether the usage of short-term spatiotemporal context made PICAN accurately interpret queries and give corresponding responses.

Regarding the usage of long-term exploration context, we decided whether a response is satisfactory based on whether users found the response useful in the post-interview. This was to see how long-term exploration context could be used to make responses more useful for onboarding.

6.4 Results

Participants were engaged in having a free-form dialogue with the NPC while exploring the metaverse, asking 556 questions in total (M= 26.48, Std=9.57). Participants chose to ask different types of queries for the same goal, highlighting the importance of high user freedom in the types of queries. For instance, when dealing with searching for NPCs (Task 2), some participants preferred asking direct factual queries (*"Where is [NPC name]?"*), some actively walked around and asked about the identity of nearby NPCs (*"Who is this?"*), while some asked both types of queries. This section discusses some satisfactory and unsatisfactory cases of usage of each context, focusing on how participants interacted with the system during tasks and evaluated them in the post-interview.

6.4.1 *Spatial context.*

Participants used different question formats that require spatial context: asking about their own location (e.g., *"Where am I?"*); referring to an object as *this*; referring to an object as *this* + object type (e.g., this building, this person, this friend); referring to vicinity (e.g., here, nearby, near me); and referring to vicinity with direction (e.g, in front of me, next to me). These questions all described the location of the participant's character, or nearby area and object. There was only one question that used demonstrative "that" to describe a far-away object (*"How can I ride that cable car?"*).

Satisfactory cases Participants (1) asked about unknown spaces using references and (2) continued conversation with the NPC regarding the referred object.

First, participants commonly used the words "here" and "this building" to describe a new place they visited. U10 described the importance of the NPC's awareness of spatial context, saying that there are various components of surroundings in the metaverse, such as the current place and nearby objects, and being able to skip the verbal explanation of spatial context increases the usability of NPC. Participants (U7,U8) especially appreciated being able to refer to new places without knowing their names.

Second, after the NPC identified the referred object, they continued discussing the object. For instance, a participant first asked about the current location by referring to it as "here". After receiving a response saying that it was an EventHall, they proceeded to ask what they could do at the place. Similarly, when U2 asked how

long they had to wait for "this" in front of a coffee tree, PICAN identified that a coffee tree was nearby and that the query was about the time required for the coffee tree to grow. After receiving a response about the time estimate for full growth, the participant asked whether they could leave the place and come back until the tree is fully grown.

Unsatisfactory cases While many participants learned about nearby objects and buildings by asking simple queries using references, they often had to (1) specify their reference and (2) repeat the rephrased question when the response only considered vicinity. They also commented on (3) discrepancy between spatial context of query and response, and (4) limited usage of short-term spatiotemporal context when describing directions.

First, some participants often had to specify their query when spatial context was identified as a larger area than they hoped. For instance, when U7 asked if Amy is here while standing in front of EventHall, PICAN resolved "here" as the whole town and gave a generalized response about Amy's location. Therefore, U7 revised the question to ask if Amy is in front of the EventHall.

Second, some participants had to keep asking the same or rephrased question when PICAN often did not clarify where the target was, but simply said it could not be found nearby. Such a case was often observed while the participants were searching for NPCs in Task 2. For instance, when U11 asked where Tom was, PICAN did not specify the whereabouts but said they could not be found nearby. They had to repeat the question about Tom four more times moving around. Such responses could also affect how users perceived the AI NPC's intent and knowledge. For instance, U9 and U15 thought PICAN was intentionally hiding the information to make it harder to find. U15 thought that even though they kept searching for NPCs while interacting with PICAN, less trained users would think PICAN is incapable of answering about NPCs and stop asking related questions.

Third, some participants noticed the discrepancy between spatial context considered in question and spatial context they perceived. U1 pointed out that since they are constantly moving around the metaverse for exploration, their location at the moment of query input and the NPC response is different. One participant was perplexed when they thought they were in front of *Virtumall* but NPC said *"If you are in front of Virtumall"*, as if they were not. This was the result of the difference between the range of vicinity that the user and PICAN perceived.

Lastly, some participants wished that the spatial context awareness was extended to not just describe nearby objects, but also give customized directions toward far-away objects. When asked whether they think the NPC considered their location well when answering questions, some participants (U2,U3,U5,U17) wished the NPC described the detailed path from the current location to the destination.

6.4.2 *Dialogic context.*

Satisfactory cases Participants asked questions or described themselves with sentences that needed previous dialogue for the NPC to fully interpret. For instance, U21 told PICAN to pick "one" after getting a response about types of seeds, and U16 asked "Is he inside?" after getting a response about Sam (an NPC).

Unsatisfactory cases Participants were unsatisfied with the usage of dialogic context when experiencing a cascade of misinformation. Once made a mistake such as hallucination or misinterpreting the context, PICAN often continued giving wrong information by referring to the previous dialogue. For instance, the system first gave misinformation saying Sara (NPC) was on a farm, when in fact they were in front of a mall. Then, when the participant asked what Sara was wearing, PICAN interpreted the query as *"What is Sara doing while working on a farm?"* using dialogic context. This resulted in creating the wrong answer once again. U10 commented that while they think keeping dialogic context would be helpful, remembering everything including misinformation is problematic. To deal with such cases, U9 even wished they could remove some of the dialogue histories once they knew some responses were wrong so that they could stop the cascading misinformation.

User queries that required action context to answer were not observed.

6.4.3 *Long-term exploration context.*

Satisfactory cases Participants appreciated when PICAN (1) recalled a previous action at a place, (2) object related to a subsequent action, and (3) recalled an object to review.

First, some participants (U7, U9, U11, U20) mentioned that they could easily understand where they have to go by referring to their past actions at the destination. For instance, when discussing a response from NPC (*"You can indeed enter Virtufarm; it's right where you've been tending to your coffee trees!"*), U9 said it helps because he often remembered place not by its name, but what he did there. U8 and U10 specified the hardships of learning and remembering new names of places in the metaverse as first-timer users, which makes place descriptions with previous experience especially helpful.

Second, some participants thought that when PICAN recalled an object related to the subsequent action, it made the response align better with user intention. For instance, when U10 asked how he could try on clothes, the NPC said *"To change your clothes, go to your inventory, pick the lovely white knit, and choose "wear" to put it on!"*. Since the NPC Mentioned the "white knit" that U10 bought before, he appreciated that the NPC understood that he was trying to put on the new clothes he bought.

Third, some participants thought they could review the information they knew when PICAN reminded them. For instance, U20 mentioned that when PICAN explained where *Virtumall* is in relation to the places that they already visited, they could remind themselves where such places were.

Unsatisfactory cases Participants often (1) failed to realize the recalled past action when they could not understand the textual description of their past experience. They were also unsatisfied with the responses that recall (2) past actions in too much detail or (3) recall actions irrelevant to the intent of the query.

First, some participants did not know what the recalled action was because they failed to match the text description of the action to what they did. For instance, U10 described that they did not know that what they clicked was the "change screen" button, and were perplexed when the NPC said that they had already done it (Table 2).

Second, some participants commented that the level of detail when describing the past experience is important. U13 commented that the past action was described in too much detail in one of the responses. When PICAN described the location of EventHall as somewhere "beyond where you picked up your green two-tone camping item.", U13 thought sharing the exact detail of a related object (e.g., mentioning the exact name of the product they bought) disrupted the flow of conversation and they would prefer a more general explanation.

Third, some participants preferred a short answer without the addition of long-term exploration context considering the intent of the query. When U12 asked about the identity of the nearby person, they wished PICAN simply responded with their name. They thought recalling the past actions in the building nearby was unnecessary.

7 DISCUSSION

In this section, we discuss the implications and considerations of LLM-based context-aware NPCs when onboarding metaverse.

7.1 Using Context to Interpret Questions and Customize Responses

In the user study, PICAN's understanding of short-term spatiotemporal context allowed participants to seamlessly refer to spatial context and dialogic context with the NPC using short natural language expressions. Specifying precise inputs in lengthy utterances to a conversational agent is a cumbersome task [20], and being aware of short-term spatiotemporal context and resolving references made NPC's response more useful (Section 5.2). PICAN's understanding of long-term exploration context made the responses more customized to the user's knowledge level of the metaverse. It mainly helped users learn about new objects and locations by connecting them to what they already know (Section 6.4). By doing so, users could review their knowledge and learn about new concepts by connecting to past experiences.

PICAN only considered location, action, and object when extracting relevant logs from the exploration state, and attached them to a simple response. User study participants also commented on the expansion of the concept of the *relevancy* of logs and how to *describe* them. For instance, participants with experience playing Minecraft suggested that long-term exploration context could be utilized to explain crafting, where users fill in 3x3 squares with certain materials in certain positions. Instead of describing the positions one by one, the agent may simply say "pickaxe pattern (U11)" or "box pattern (U9)" if the user has a similar crafting experience. In such case, the new *relevancy* measure would be "shape of pattern", and *description* would be "using jargon". Considering the unsatisfactory cases of long-term exploration context in user study also dealt with *relevancy* (with the intent of question) and *description* (using unfamiliar language or in too much detail), improving these concepts may improve the context-awareness.

Future works may also explore how short-term spatiotemporal context and long-term exploration context may benefit onboarding in different domains that involve the multimodal environment and procedural knowledge, such as web design or physical world exploration. For instance, in web design tutorials, the short-term spatiotemporal context may include visual components (spatial context), previous questions (dialogic context), and web edit activities (action context). Long-term exploration context may refer to logs from the exploration state that share edit methods or menu locations with the current response.

Table 2: The interaction between a user and the NPC where the user failed to match the verbal description (change screen option) with the action the user was doing

Entity	Comment or Action
User	[Stands in front of the screen of the Event Hall]
	How can I change the view?
NPC	To change the view in the Event Hall, just select the "changeScreen" option!
User	[clicks "change screen" button with "post PDF on screen" option]
	[clicks "change screen" button with "post Youtube video on screen" option]
	Where is the change screen option?
NPC	The screen change option is in the Event Hall, where you accessed YouTube and PDF before!

7.2 Using LLM for Onboarding Agent

Our system design, evaluation, and user studies mainly focused on developing the context awareness of the onboarding agent. However, we could also identify some benefits and challenges of using LLM for metaverse onboarding agents. Using LLM, PICAN managed to understand free-form user questions and make logical reasonings such as figuring out relevant user history. When compared with their past experiences of learning about the new metaverse mainly through internet search, the formative and user study participants appreciated being able to ask questions whenever they wanted in natural language. Not having to leave the metaverse to search for answers online or in tutorials also made the onboarding experience feel more immersive. However, some participants reported sometimes having a hard time forming the questions that the NPC could answer well and hoped they were provided a list of example questions in the beginning. Moreover, the hallucination of LLM often tricked users into looking for non-existing objects and cascaded along the dialogue, inducing more confusion. Hallucination in LLM yet remains even with tactics such as RAG and fine-tuning [24, 55], and future works may explore not only how to reduce hallucination but also how to recover from it. For instance, future works may add a fact-checking pipeline to stop the cascade of hallucination along dialogic context, or design a response scenario for when a user finds out that the informed building does not exist.

7.3 Enhancing the context-usage to initiate conversation

In our user study, contexts that activated the NPC to suggest help or respond without the user's question were rarely observed. We hypothesize this is because the participants were actively engaging in different activities without stopping or repeating problematic actions since they were solving multiple tasks in a short span of time. In studies that are designed to better resemble real user situations without any time constraints and tasks, we expect our proactive approach to be observed more than our user study. Furthermore, to avoid being disruptive, PICAN provided very limited proactive approaches, meaning it rarely initiated dialogues. If the system has a better understanding of users' goals that are constantly changing, it could provide more proactive approaches. Existing approaches of multimodal goal recognition in an open world [35] may help provide

better context-aware assistance. The high variance in the preferred level of proactiveness also exposed the need for the personalized design of timing and the contents of the proactive approach. To combine goal recognition and personalization with our question-answering module, identifying the exact goal or the problem of the user and translating it to the corresponding query would be important.

7.4 Generalizing context-aware onboarding agents over time

PICAN was designed to support the onboarding phase of metaverse users. However, we could also observe how interaction with PICAN may change over time. Two participants said they would ask more questions about the metaverse once they were comfortable with the fundamentals. Five participants said they would switch to other subjects, such as their everyday lives, primarily for entertainment. It is anticipated that the various conversational agent (CA) usage settings may influence how information- or emotion-focused users' utterances are [56]. Unlike interaction with other CAs like smart speakers where users rarely change the subject of conversations [7], the increased familiarity with the metaverse over time may change the main role of the NPC from informative to more emotional or relationship-focused. In the transition, the immersive responses generated by context awareness (Section 5.3) may positively affect the relationship, since having memory of the users helps them feel closer to agents [32]. Future research may examine how users interpret the NPC's evolving role — which was initially thought to be informative — especially in situations when interactions are personalized with responses that are sensitive to the user context. Such question-and-answer personalization could have an impact on users' perception of the NPC and potentially affect the frequency and topic of future interactions.

8 LIMITATIONS

Although our study revealed the usefulness of PICAN in metaverse onboarding, we acknowledge some limitations.

First, the deployment and the evaluations were conducted on a single metaverse platform (*Virtuoville*). We believe that the PICAN could be generalized to other platforms since the pipeline structure does not rely on the particular design of the metaverse, and can be adapted with platform-specific information such as object locations and possible activities. Future research may evaluate the performance of PICAN in different metaverse platforms, varying in complexity.

Second, the participants' pool was skewed towards people having the same nationality in their 20s–30s. Future research may expand the target user to other languages and age groups, who may have different usages of and reactions towards the AI NPC.

Third, although PICAN could answer a wide range of user queries, queries about the information that is out of range of PICAN's knowledge scope received incorrect or ambiguous responses (e.g., queries about a small object in *Virtuoville* that was not in knowledge scope). Although these queries were rarely observed, we believe extending the knowledge scope can widen the range of queries that can be answered.

Lastly, the response appeared around 15–20 seconds after the user asked a query, and the latency may have prevented the users from asking queries related to the current context. Future work may explore the methods that increase the efficiency of the pipeline.

9 CONCLUSION

In this work, we propose PICAN, a pipeline designed to provide context-aware guidance to help users during the metaverse onboarding process. Our pipeline evaluation shows the importance of utilizing context, especially short-term spatiotemporal context for useful and immersive responses. Furthermore, our user study reveals that PICAN was capable of supporting fundamental needs of spatial context and dialogic context awareness and generating responses that utilize long-term exploration context for in-context explanations.

ACKNOWLEDGMENTS

The authors thank the members of KAIST Interaction Lab for their constructive feedback throughout the work. This work was supported by a grant of the KAIST-KT joint research project through AI2XL Laboratory, Institute of Convergence Technology, funded by KT [G01220645, Designing the UI/UX to support senior users on metaverse].

REFERENCES

[1] Suzan Al-Nassar, Anthonie Schaap, Michael Van Der Zwart, Mike Preuss, and Marcello A. Gómez-Maureira. 2023. QuestVille: Procedural Quest Generation Using NLP Models. In *Proceedings of the 18th International Conference on the Foundations of Digital Games* (Lisbon, Portugal) *(FDG '23)*. Association for Computing Machinery, New York, NY, USA, Article 50, 4 pages. https://doi.org/10.1145/3582437.3587188

[2] Büşra Alma Çallı and Çağla Ediz. 2023. Top concerns of user experiences in Metaverse games: A text-mining based approach. *Entertainment Computing* 46 (2023), 100576. https://doi.org/10.1016/j.entcom.2023.100576

[3] Erik Andersen, Eleanor O'rourke, Yun-En Liu, Rich Snider, Jeff Lowdermilk, David Truong, Seth Cooper, and Zoran Popovic. 2012. The impact of tutorials on games of varying complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 59–68. https://doi.org/10.1145/2207676.2207687

[4] Apple. 2011. Siri. https://www.apple.com/siri/ Accessed: 2024-02-09.

[5] Trevor Ashby, Braden K Webb, Gregory Knapp, Jackson Searle, and Nancy Fulda. 2023. Personalized Quest and Dialogue Generation in Role-Playing Games: A Knowledge Graph- and Language Model-Based Approach. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 290, 20 pages. https://doi.org/10.1145/3544548.3581441

[6] Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022. Keep Me Updated! Memory Management in Long-term Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3769–3787. https://doi.org/10.18653/v1/2022.findings-emnlp.276

[7] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24. https://doi.org/10.1145/3264901

[8] Katherine Bessière, Jason B. Ellis, and Wendy A. Kellogg. 2009. Acquiring a Professional "Second Life": Problems and Prospects for the Use of Virtual Worlds in Business. In *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (Boston, MA, USA) *(CHI EA '09)*. Association for Computing Machinery, New York, NY, USA, 2883–2898. https://doi.org/10.1145/1520340.1520416

[9] Corbin Brown. 2024. Ubisoft NEO NPCs Gameplay: AI-Powered Characters in Video Games. https://www.youtube.com/watch?v=1od2pIs9220

[10] Peter J Cosgrove. 2016. *The effects of gamification on self-efficacy and persistence in virtual world familiarization*. Ph. D. Dissertation. University of Missouri–Columbia. https://doi.org/10.32469/10355/56469

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. https://doi.org/10.18653/v1/N19-1423 arXiv:1810.04805 [cs.CL]

[12] Betsy Dijk, Rieks Akker, Anton Nijholt, and Job Zwiers. 2003. Navigation Assistance in Virtual Worlds. *Informing Science The International Journal of an Emerging Transdiscipline* 6 (01 2003). https://doi.org/10.28945/519

[13] Maricel A Esclamado and Maria Mercedes T Rodrigo. 2022. Are all who wander lost? An exploratory analysis of learner traversals of Minecraft worlds. In *International Conference on Artificial Intelligence in Education*. Springer, 263–266. https://doi.org/10.1007/978-3-031-11647-6_48

[14] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. 2022. MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge. 35 (2022), 18343–18362. https://doi.org/10.48550/arXiv.2206.08853

[15] Maxwell Foxman. 2022. Gaming the system: Playbour, production, promotion, and the metaverse. *Baltic Screen Media Review* 10, 2 (2022), 224–233. https://doi.org/10.2478/bsmr-2022-0017

[16] Jonathan Gray, Kavya Srinet, Yacine Jernite, Haonan Yu, Zhuoyuan Chen, Demi Guo, Siddharth Goyal, C. Lawrence Zitnick, and Arthur Szlam. 2019. CraftAssist: A Framework for Dialogue-enabled Interactive Agents. https://doi.org/10.48550/arXiv.1907.08584 arXiv:1907.08584 [cs.AI]

[17] Abraham Guerra. 2011. *A framework for building intelligent software assistants for virtual worlds.* Technical Report. Pace University.

[18] Anisha Gupta, Dan Carpenter, Wookhee Min, Jonathan Rowe, Roger Azevedo, and James Lester. 2022. Enhancing multimodal goal recognition in open-world games with natural language player reflections. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 18. 37–44. https://doi.org/10.1609/aiide.v18i1.21945

[19] Susan Hazan. 2008. Musing the Metaverse. *2008 Annual Conference of CIDOC* (2008).

[20] Mohit Jain, Ramachandra Kota, Pratyush Kumar, and Shwetak N Patel. 2018. Convey: Exploring the use of a context view for chatbots. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–6. https://doi.org/10.1145/3173574.3174042

[21] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N. Patel. 2018. Evaluating and Informing the Design of Chatbots. In *Proceedings of the 2018 Designing Interactive Systems Conference* (Hong Kong, China) *(DIS '18)*. Association for Computing Machinery, New York, NY, USA, 895–906. https://doi.org/10.1145/3196709.3196735

[22] Dusan Jan, Antonio Roque, Anton Leuski, Jacki Morie, and David Traum. 2009. A Virtual Tour Guide for Virtual Worlds. In *Intelligent Virtual Agents*, Zsófia Ruttkay, Michael Kipp, Anton Nijholt, and Hannes Högni Vilhjálmsson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 372–378. https://doi.org/10.1007/978-3-642-04380-2_40

[23] Anssi Kanervisto, Janne Karttunen, and Ville Hautamäki. 2020. Playing minecraft with behavioural cloning. In *NeurIPS 2019 Competition and Demonstration Track*. PMLR, 56–66. https://doi.org/10.48550/arXiv.2005.03374

[24] Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. 2024. Unfamiliar Finetuning Examples Control How Language Models Hallucinate. *arXiv preprint arXiv:2403.05612* (2024). https://doi.org/10.48550/arXiv.2403.05612

[25] Jeremy Kemp and Daniel Livingstone. 2006. Putting a Second Life "metaverse" skin on learning management systems. In *Proceedings of the Second Life education workshop at the Second Life community convention*, Vol. 20. The University of Paisley San Francisco.

[26] Bokyung Kye, Nara Han, Eunji Kim, Yeonjeong Park, and Soyoung Jo. 2021. Educational applications of metaverse: possibilities and limitations. *Journal of Educational Evaluation for Health Professions* 18 (2021). https://doi.org/10.3352/jeehp.2021.18.32

[27] Han Jin Lee and Hyun Hee Gu. 2022. Empirical Research on the Metaverse User Experience of Digital Natives. *Sustainability* 14, 22 (Nov 2022), 14747. https://doi.org/10.3390/su142214747

[28] Sang-Gun Lee, Silvana Trimi, Won Ki Byun, and Mincheol Kang. 2011. Innovation and imitation effects in Metaverse service adoption. *Service business* 5 (2011), 155–172. https://doi.org/10.1007/s11628-011-0108-8

[29] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 33 (2020), 9459–9474. https://doi.org/10.48550/arXiv.2005.11401

[30] Na Li and Robert Ross. 2023. Hmm, You Seem Confused! Tracking Interlocutor Confusion for Situated Task-Oriented HRI. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction* (Stockholm, Sweden) *(HRI '23)*. Association for Computing Machinery, New York, NY, USA, 142–151. https://doi.org/10.1145/3568162.3576999

[31] Andy Luse, Brian Mennecke, and Janea Triplett. 2013. The changing nature of user attitudes toward virtual world technology: A longitudinal study. *Computers in Human Behavior* 29, 3 (2013), 1122–1132. https://doi.org/10.1016/j.chb.2012.10.004

[32] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105. https://doi.org/10.48550/arXiv.2307.15810

[33] Paris Mavromoustakos Blom, Sander Bakkes, Chek Tan, Shimon Whiteson, Diederik Roijers, Roberto Valenti, and Theo Gevers. 2014. Towards personalised gaming via facial expression recognition, In Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 10, 1, 30–36. https://doi.org/10.1609/aiide.v10i1.12707

[34] Microsoft. 2024. *Minecraft.* https://www.minecraft.net/ Accessed: 2024-02-09.

[35] Wookhee Min, Bradford Mott, Jonathan Rowe, Robert Taylor, Eric Wiebe, Kristy Boyer, and James Lester. 2021. Multimodal Goal Recognition in Open-World Digital Games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* 13, 1, 80–86. https://doi.org/10.1609/aiide.v13i1.12939

[36] Mu Mu and Murtada Dohan. 2023. Unstuck in Metaverse: Persuasive User Navigation using Automated Avatars. *IEEE Communications Magazine* 61, 9 (2023), 56–62. https://doi.org/10.1109/MCOM.002.2200608

[37] Stephan Müller, Mubbasir Kapadia, Seth Frey, Severin Klinger, Richard P Mann, Barbara Solenthaler, Robert W Sumner, and Markus Gross. 2015. Statistical analysis of player behavior in Minecraft. In *Proceedings of the 10th International Conference on the Foundations of Digital Games*. Society for the Advancement of the Science of Digital Games.

[38] Naver Z Corporation. 2020. *Zepeto.* https://web.zepeto.me/ Accessed: 2024-02-09.

[39] Josef Nguyen. 2016. Minecraft and the building blocks of creative individuality. *Configurations* 24, 4 (2016), 471–500. https://doi.org/10.1353/con.2016.0030

[40] OpenAI. 2023. GPT-4 Technical Report. https://doi.org/10.48550/arXiv.2303.08774 arXiv:2303.08774 [cs.CL]

[41] Panagiotis D. Paraschos and Dimitrios E. Koulouriotis. 2023. Game Difficulty Adaptation and Experience Personalization: A Literature Review. *International Journal of Human–Computer Interaction* 39, 1 (2023), 1–22. https://doi.org/10.1080/10447318.2021.2020008 arXiv:https://doi.org/10.1080/10447318.2021.2020008

[42] 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* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. https://doi.org/10.1145/3586183.3606763

[43] Anton Petrov. 2014. Using Minecraft in Education: A Qualitative Study on Benefits and Challenges of Game-Based Education. https://api.semanticscholar.org/CorpusID:63435870

[44] Roblox Corporation. 2024. *Roblox.* https://www.roblox.com/ Accessed: 2024-02-09.

[45] Samsung. 2017. Bixby. https://www.samsung.com/us/apps/bixby/ Accessed: 2024-02-09.

[46] Stefan Seidel, Nicholas Berente, Jeffrey Nickerson, and Gregory Yepes. 2022. Designing the metaverse. (2022).

[47] Vaishnavi Shah. 2018. *Examining the Effect of Different Types of Tutorials on New Players of a Compute Science Teaching Game.* Ph. D. Dissertation. Northeastern University. https://doi.org/10.17760/D20289866

[48] Chen Si, Yusuf Pisan, and Chek Tien Tan. 2016. Understanding players' map exploration styles. In *Proceedings of the Australasian Computer Science Week Multiconference*. 1–6. https://doi.org/10.1145/2843043.2843480

[49] Yuqian Sun, Ying Xu, Chenhang Cheng, Yihua Li, Chang Hee Lee, and Ali Asadipour. 2022. Travel with Wander in the Metaverse: An AI chatbot to Visit the Future Earth. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP55362.2022.9950031

[50] Mariët Theune, Dennis Hofs, and Marco van Kessel. 2007. The virtual guide: a direction giving embodied conversational agent. In *Proc. Interspeech 2007*. 2197–2200. https://doi.org/10.21437/Interspeech.2007-598

[51] Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-Tuning GPT-2 on Annotated RPG Quests for NPC Dialogue Generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games* (Montreal, QC, Canada) *(FDG '21)*. Association for Computing Machinery, New York, NY, USA, Article 2, 8 pages. https://doi.org/10.1145/3472538.3472595

[52] Ryan Volum, Sudha Rao, Michael Xu, Gabriel A DesGarennes, Chris Brockett, Benjamin Van Durme, Olivia Deng, Akanksha Malhotra, and Bill Dolan. 2022. Craft an Iron Sword: Dynamically Generating Interactive Game Characters by Prompting Large Language Models Tuned on Code. In *The Third Wordplay: When Language Meets Games Workshop.* https://doi.org/10.18653/v1/2022.wordplay-1.3

[53] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. https://doi.org/10.48550/arXiv.2305.16291 arXiv:2305.16291 [cs.AI]

[54] Henrik Warpefelt and Harko Verhagen. 2015. Towards an updated typology of non-player character roles. In *Proceedings of the international conference on game and entertainment technologies*. 1–9.

[55] Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Cheng Niu, Randy Zhong, Juntong Song, and Tong Zhang. 2023. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *arXiv preprint arXiv:2401.00396* (2023). https://doi.org/10.48550/arXiv.2401.00396

[56] Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. 3506–3510. https://doi.org/10.1145/3025453.3025496

[57] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long Time No See! Open-Domain Conversation with Long-Term Persona Memory. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 2639–2650. https://doi.org/10.18653/v1/2022.findings-acl.207

[58] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2204–2213. https://doi.org/10.18653/v1/P18-1205

[59] Hanxun Zhong, Zhicheng Dou, Yutao Zhu, Hongjin Qian, and Ji-Rong Wen. 2022. Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, 5808–5820. https://doi.org/10.18653/v1/2022.naacl-main.426

[60] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. https://doi.org/10.48550/arXiv.2305.17144 arXiv:2305.17144 [cs.AI]