International Journal of Human–Computer Interaction

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hihc20

The Effects of LLM-Empowered Chatbots and Avatar Guides on the Engagement, Experience, and Learning in Virtual Museums

Shuhao Zhang, Mingge Ma, Yue Li, Ka Lok Man, Jeremy Smith & Yong Yue

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2494072

The Effects of LLM-Empowered Chatbots and Avatar Guides on the Engagement, Experience, and Learning in Virtual Museums

Shuhao Zhang[a], Mingge Ma[a], Yue Li[a], Ka Lok Man[a], Jeremy Smith[b], and Yong Yue[a]

[a]School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China; [b]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, United Kingdom

ABSTRACT
Virtual museums, as a form of serious game, transcend spatiotemporal constraints of traditional museums through advanced computer technologies, enabling immersive and educational user experiences. With the support of large language models (LLMs) in particular, users can have exploratory interactions with conversational agents. This paper presents a comparative study examining user engagement, experience, and learning across three interaction methods: text labels, chatbots, and avatar guides in a virtual museum. Our study revealed that users were more engaged with the LLM-empowered avatar guide than chatbots and text labels. Additionally, the overall user experience with avatar guides was superior to that of chatbots and text labels. However, no significant differences were observed in users' learning outcomes across the three interaction methods. We discussed the characteristics of these interaction methods and proposed design recommendations for virtual museums and serious games.

KEYWORDS
Serious games; virtual museums; large language model; chatbots; avatar guides

1. Introduction

The traditional museum experience, while rich in historical and cultural significance, often faces limitations in accessibility, interactivity, and personalization. With the advent of virtual technologies and large language models (LLMs), virtual museums have emerged as promising solutions to these challenges. By creating digital replicas of physical museums, virtual museums offer a new way to experience and interact with cultural heritage. For example, Vardhan et al. (2022) designed a system to allow users to visit historic sites through augmented reality (AR). Similar methods were used by researchers, allowing visitors who wear virtual reality head Head-Mounted Displays (VR HMDs) to achieve a natural visiting experience by overlaying a historical virtual environment (Capece et al., 2024; Tennent et al., 2020) and virtual cultural relics (Tsita et al., 2023) onto an actual exhibition room. In contrast, LLM-based agents enhance the virtual museum experience through their language generation capabilities. For example, in the virtual museum designed by Vasic et al. (2024), LLMs were used as guides during a 3D panoramic virtual tour of the Civic Art Gallery of Ascoli, allowing visitors to express their interest and view the requested content. Although conversational agents have been proven effective in the visiting of real museums (Kopp et al., 2005), there remains a lack of research on their impact on user experience and learning in the virtual museum. Thus, investigating the effects of LLM-empowered agents in virtual museums is a crucial area of study. Another important consideration is the format of conversational agents. In the existing systems, chatbots are the most popular method for assisting users during their visits (Kopp et al., 2005; Machidon et al., 2020). However, research indicates that altering the agent's format can influence user experience. For example, Saito (2023) found the voice-only agent showed better reliability, usefulness, understandability, enjoyment, and likability than the text-only agent in art exhibition visiting. Additionally, Bönsch et al. (2021) explored how an agent's active communication style affects user experience, concluding that the level of interaction does not significantly impact user experience. These findings suggest that the design and interaction mechanisms of agents can affect user experiences. However, it is unclear if the findings from traditional conversational agents can be applied to LLM-empowered ones, making this an important topic for exploration in virtual museums.

To fill the research gaps, we present a comparative study that investigates the differences in users' engagement, experience and learning across three interaction methods (text labels, chatbots, and avatar guides) in a virtual museum environment. Our findings indicate that LLM-empowered methods, particularly avatar guides, positively influence user engagement and experience in virtual museums. However, these approaches did not lead to significant enhancements in learning motivation or outcomes. Our main contributions are three-fold. First, we designed a virtual museum that features three types of information presentations, with two of them empowered by a Chinese large language model. Second, we presented an empirical evaluation of the users' engagement, experience, and learning associated with the

CONTACT Yue Li Yue.Li@xjtlu.edu.cn School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou, China

three interaction methods during their visits and interactions with museum artifacts. Finally, we discussed the features of three interaction methods and offered design implications for LLM-empowered virtual museums, providing valuable insights for future virtual museum systems and serious games.

2. Related work

2.1. Virtual museums and serious games

As an important complement to traditional physical museums, virtual museums provide unique opportunities for culture dissemination and education. Compared to the traditional visiting method, virtual museums break the limitation of time and space and allow users to visit and interact with artifacts at any time. For example, by using augmented reality (AR) (Li et al., 2023; Vardhan et al., 2022) and virtual reality (VR) (Ch'ng et al., 2020; Tennent et al., 2020; Wang et al., 2024), users could see and interact with artifacts in virtual forms. Researchers also advocated extending the museum trajectory, encouraging visitors to collect artifacts and take them home (Xu et al., 2024). Such novel strategies not only enhance user interaction but also promote continuous learning and appreciation of cultural artifacts (Xu et al., 2022). Another benefit of being "virtual" is the ability to transcend the constraints of the physical world, including both human-made rules and the laws of physics. For example, touching and moving artifacts are usually not allowed when visiting a real museum. However, allowing interaction with artifacts is a common practice in virtual museum design (Wang et al., 2023). This includes the ability to resize artifacts that may be unusually large or small, providing an unique experience that only virtual environments can offer. About the learning outcomes, Tian et al. (2024) found that virtual museum experiences can be as effective as physical visits in promoting learning and engagement, particularly when combined with interactive and multimedia elements. This highlights the potential of virtual museums to enhance educational experiences while making cultural heritage more accessible. Virtual museums can incorporate multimedia elements such as videos, interactive quizzes, and user-generated content, further enriching the educational experience. Therefore, they can cater to diverse audiences, providing tailored educational content that meets various learning preferences.

Virtual museums often serve as the foundation for serious games focused on cultural heritage. Serious games are games with a specific purpose other than pure entertainment (Pilote & Chiniara, 2019). Typically, they are designed to improve users' learning interest and effectiveness through gamification and have been used in various fields, such as education (Ding & Yu, 2024), training (Feng et al., 2020), and rehabilitation (Mirza-Babaei et al., 2014). For cultural heritage in particular, many works have proven that serious games improve user experience and learning outcomes. For example, the *Undercroft Game* is a small puzzle game where the player interacts with various non-player characters to uncover clues and solve the puzzles (Petridis et al., 2013). This game aims to increase the audience's interest in cultural knowledge while showing the daily activities of the monks of Coventry's original Benedictine monastery. Although the evaluation of this work was conducted with a small sample size of five users, it showed the potential of using games to learn history knowledge and improve user experience. Embodiment is a commonly referenced concept in the design of games for virtual museums. For example, *Monsters Eat Art* (Rehm & Jensen, 2015) features a "monster" as the narrator, guiding children through a treasure hunt in an art museum to learn about artifacts. The use of a character like the monster serves to capture the attention of young audiences, making the educational content more relatable and engaging. This approach exemplifies how embodiment can transform traditional museum visits into dynamic learning experiences, contributing to a deeper appreciation for cultural heritage. Overall, these examples indicate possibilities for museums and cultural heritage experiences to incorporate virtual technologies and game mechanisms to support playful learning experiences and social activities.

2.2. LLM in games and education

Large language models (LLMs) are a category of foundation models trained on immense amounts of data. Compared to traditional working modes, LLM-supported working systems could understand and generate natural language and other types of content to perform a wide range of tasks in different fields. Over the past few years, LLMs have shown positive effects in business (Cheung, 2024), education (Katz et al., 2023), medicine (Azizi et al., 2023) and entertainment (Hu et al., 2024). The advent of LLMs has transformed game design by enabling more natural interactions between players and games. In traditional games, the interaction between users and non-player characters (NPCs) should follow a set workflow, and the interaction is limited to specific options. Nowadays, the application of LLMs in games provides new opportunities to enhance the interactions. Designers could leverage LLMs to create dynamic, interactive narratives (Harmon & Rutman, 2023; Yong & Mitchell, 2023), facilitate open dialogue with NPCs (Paduraru et al., 2023; Volum et al., 2022) and create game scenes (Kumaran et al., 2023). One of the emerging applications of LLMs is in simulating NPC conversations. LLMs excel at natural language processing, allowing them to produce contextually relevant and engaging responses based on player inputs. This capability not only enriches the interaction with NPCs but also minimizes repetitive dialogue, fostering a more exploratory and immersive gameplay experience (Reed et al., 2022). For example, in *Yandere AI Girlfriend Simulator*[1], players need to escape a locked room by interacting with an LLM-based "girlfriend," engaging in realistic conversations rather than scripted plots based on button clicks. This approach enhances the realism of the interaction, making it more engaging and personal, as players must build rapport with the NPC to progress in the game.

As shown above, LLM-supported games demonstrate advantages in the high freedom and the natural interaction process. These advantages demonstrated in games can also be applied to education contexts. For example, Seßler et al. (2024) found that LLMs could positively affect learners' motivation in math education. Compared to traditional learning methods, students experience less pressure and tension while enjoying greater interest and engagement when using an LLM-supported learning system. In addition, LLMs have shown promising results in improving the quality of engineering essays, demonstrating their practicality in educational settings (Bernabei et al., 2023).

2.3. Chatbots and Avatar Guides

Chatbots are computer programs designed to simulate human conversations with users. They facilitate easy access to information by providing instant responses to inquiries without the need for human intervention or manual research. This capability has made them valuable interactive tools for enhancing user experiences in museums and other cultural venues. For example, *Max*, a conversational agent used at the HNF museum in Germany (Kopp et al., 2005), allows users to communicate via keyboard inputs. The system could generate responses based on predefined rules. Another example is CulturalERICA (Machidon et al., 2020), which integrates with the Europeana database to match specific intents with the context of the conversation. In recent years, ChatGPT has emerged as one of the most well-known and widely used chatbots (OpenAI, 2022). Since its introduction in 2022, its impact has been explored across various domains, including education (Xue et al., 2024), nutriology (Tsiantis et al., 2024) and culture heritage (Ribeiro et al., 2024).

Researchers have explored ways to improve user experience given the practicability of chatbots. For example, Saito (2023) examined how different modalities of agents affect visitors while viewing paintings in an online art viewing service. They found that voice-only agents were rated the highest in reliability, usefulness, understandability, enjoyment, and likability compared to text agents. Additionally, avatar agents demonstrated greater understandability and enjoyment than text agents. However, their work was not based on LLMs and the content presented was strictly controlled. In another study (Bönsch et al., 2021), the effects of a virtual agent's initiative on user experience were explored, revealing no significant differences between being guided or having exploratory freedom. These findings underscore the necessity of examining both the medium through which users engage with chatbots and the nature and content of those interactions. Specifically, there has been limited research on the novel interactions provided by LLM-empowered chatbots in virtual museums, resulting in a lack of understanding regarding their impact on user engagement, experience, and learning. Exploring these areas is crucial for optimizing the effectiveness of such technologies in enhancing visitor experiences and educational outcomes.

3. System design and implementation

3.1. Apparatus

Our system was developed using a computer with Intel(R) Core(TM) i9-13900K CPU, 32GB RAM, and an NVIDIA GeForce RTX 4090 graphics card. We used Unity (version 2021.3.26f) for the system development and Rhino 7.0 for the virtual museum environment construction. In addition, CUDA11.7 and Python 3.10 were used for the implementation of the LLM chatbot and the avatar guide. We collected text descriptions about museum artifacts from their official website. A `txt` file was created for each museum artifact and used as our local documents.

3.2. Implementation of Chinese LLM agents

The implementation of Chinese LLM agents differ from English ones in its language structure. Chinese language is character-based where each character represents a word or a meaningful part of a word. The training of Chinese LLM requires data that reflects the cultural, historical, and social contexts, which can differ significantly from English and other languages. In addition, the segmentation of Chinese was more complex than languages that uses spaces to delineate words. Despite the challenges, the development of Chinese LLMs is rapidly evolving.

In this work, we adopted ChatGLM (Zhipu, 2024) and LangChain (2024) for the implementation of our LLM agents. ChatGLM is a large language model with 6.2 billion parameters. It uses technology similar to ChatGPT, optimized for Chinese QA and dialogue. LangChain is a framework for developing applications powered by LLMs. It supports the building of context-aware reasoning applications. For our work, we deployed an offline knowledge base including text descriptions about Chinese museum artifacts to support context-aware conversations.

Figure 1 illustrates the workflow of our LLM agent system. We prepared local documents that included text descriptions of the museum artefacts. These were imported and processed through text loading, segmentation, and vectorization (steps 1–7). In addition, the queries were embedded and vectorized, and the top `k` most similar text vectors are mapped with the question vector (steps 8–10). The matched texts are then matched as context along with the question to the `prompt` and submitted to the LLM for generating answers (steps 11–15).

3.3. Interaction methods

We implemented three interaction methods for accessing information about artifacts in a virtual museum: text labels, chatbot and avatar guide.

3.3.1. Text labels (baseline)

The text label placed next to a museum artifact often includes information such as its name, material, age, and historical background. These labels are the standard way that museums adopt to communicate the historical and cultural background of artifacts to visitors. Therefore, we

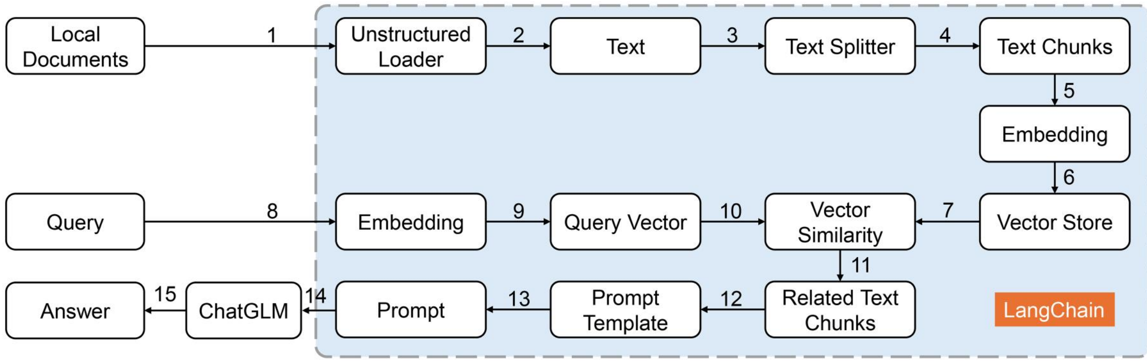


Figure 1. The workflow of our LLM agents. Image adapted from Liu et al. (2024).

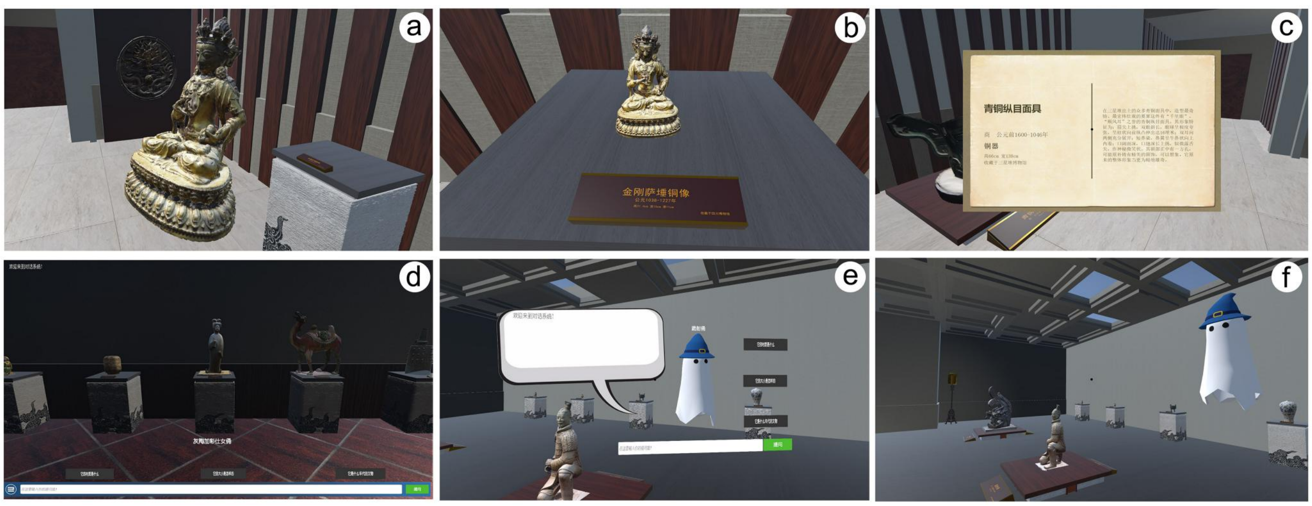


Figure 2. Screenshots showing users' views when performing different actions. (a) Grabbing an artifact in hand; (b) Observing an artifact on the pedestal; (c) Viewing the information label (baseline condition); (d) Using the Chatbot to converse; (e) Conversing with the avatar guide; (f) Having the avatar accompanying them.

implement text labels as the basic interaction method, and use it as the baseline condition for our comparisons. A detailed text label can be accessed when the user clicks on the nameplate (Figure 2(b,c)).

3.3.2. Chatbot

The style of the chatbot references common 2D user interfaces. It consists of an information display area and an input area (Figure 2(d)). We implemented two types of interactions with chatbots. The first is text input: users can type questions into the input area and receive responses in the display area. The other is quick buttons. Given that users are mostly likely to ask questions about the artifact they are viewing, we set up three buttons of frequently asked questions to improve interaction efficiency: (1) What is its material? (2) What era is it from? and (3) What is its size? We identify the item of interest by calculating the distance between users and the artifacts around them, and their camera-facing direction.

3.3.3. Avatar guide

The avatar guide replaces the 2D user interface used in chatbot to a virtual avatar that accompanies users while they move around in the museum environment. For the avatar to fit the environment (Huang et al., 2024), and to avoid the uncanny valley effect (Di Natale et al., 2023), we used a cartoon avatar (Figure 2(f)). This avatar remains with users throughout their visit and can respond to their inquiries. Similar to the chatbot, the avatar guide features an input area and a display area, but the interface is presented in a bubble dialog box. Additionally, three quick buttons were also configured around the dialog box (Figure 2(e)).

3.4. Experimental environment and interactions

The experimental environment includes a tutorial room and three exhibition rooms (Figure 3). The tutorial room includes one artifact that allows users to get familiar with the interaction controls with the three experimental conditions before the formal experiments. Each exhibition room contains six distinct museum artifacts, each accompanied by a brief nameplate indicating its name, era, and the museum from which it was collected. Users can click on the nameplate to activate a more detailed text label (Figure 2(c)).

Users can explore the virtual museum environment by looking around and moving (Li et al., 2019). They could change their viewpoint by moving the mouse and navigating through the museum using the `W A S D` keys. To closely

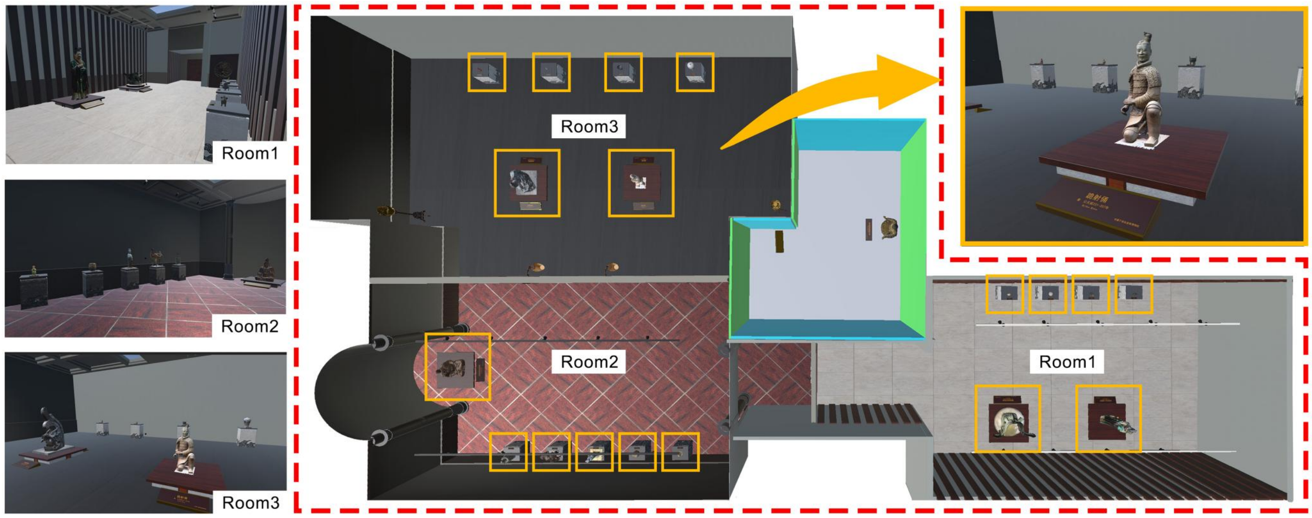


Figure 3. A virtual museum consisting of a tutorial room and three exhibition rooms. Six different artifacts were populated in each room, highlighted in orange.

observe an exhibit, users can grab it by pressing the `E` key (Figure 2(a)). Additionally, the `U O` keys are used to adjust the scale of the exhibit, while the `J-L I-K` keys allow for rotation. Users can access information about the artifacts by clicking on their nameplates, which provide detailed text panels for the baseline condition. For the chatbot and avatar guide, users can activate or deactivate these functions by pressing the + and − keys.

4. User study

4.1. Research questions and hypotheses

In this work, we ask the following research questions:

RQ1: Can LLM-empowered chatbots and avatar guides improve users' engagement in virtual museums?

RQ2: Can LLM-empowered chatbots and avatar guides improve users' experience in virtual museums?

RQ3: Can LLM-empowered chatbots and avatar guides improve users' learning in virtual museums?

For users' engagement, we measured their engagement with MES, time spent in the museum environment and in museum presentations. For user experience, we measured the usability, experience (UEQ-S) and preferences. For learning, we measured users' knowledge (MES), motivation, and test score (AIT). Based on the review of related works (Section 2), we propose the following hypotheses:

H1: LLM-empowered chatbots and avatar guides improve users' engagement in virtual museums, as indicated by the self-reported engagement (H1a), time spent in the museum environment (H1b), and time spent in museum presentations (H1c).

H2: LLM-empowered chatbots and avatar guides improve users' experience in virtual museums, as indicated by user-evaluated usability (H2a), user experience (H2b) and subjective preferences (H2c).

H3: LLM-empowered chatbots and avatar guides improve users' learning in virtual museums, as indicated by the self-reported knowledge (H3a), motivation (H3b) and knowledge test score (H3c).

4.2. Procedure

We conducted a study that took place in a 2×2m space in a university lab. After a brief introduction, we collected participants' consent and asked them to fill out a pre-experiment questionnaire about their demographic information and a pre-experiment artifact information test (AIT), assessing their knowledge about the museum artifacts. Then, participants were familiarized with the devices and adjusted the device to a comfortable physical setting, including the sensitivity of the mouse, keyboard and chair. Before the experiment, participants went through a tutorial to familiarize themselves with the (1) movements in the environment (2) interactions with artifacts, and (3) three interaction methods for accessing information (text labels, chatbot and avatar guide). The formal experiment consists of three sessions, the sequence of which was based on a Latin square design. During each session, participants were encouraged to explore the exhibition room and learn about the artifacts. They were free to spend as much time as they desired. After this, participants completed a questionnaire that measures their engagement, experience, and learning. They also filled in a post-session AIT. Then, they were encouraged to rest fully and inform the researcher when they were ready for the next experimental session. After completing three experimental sessions, participants were invited to participate in a semi-structured interview. The entire experiment lasted approximately 40 min. The experimental procedure is summarized in Figure 4. This study is approved by our University Ethics Committee.

4.3. Measures

The dependent variables of this research include users' engagement, experience and learning. We adopted both subjective and objective measures in the evaluation.

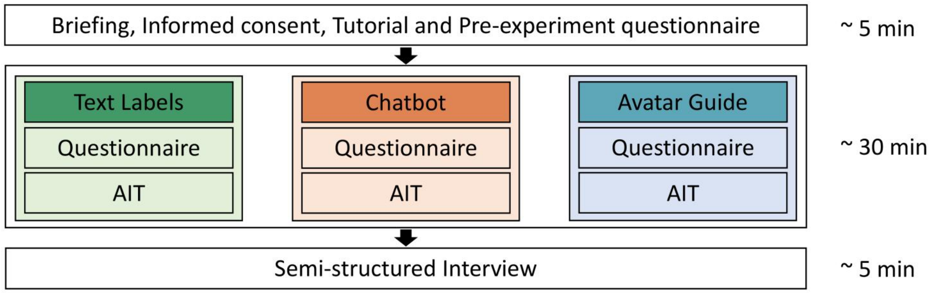


Figure 4. An example experimental procedure.

Table 1. Questions used to measure the engagement, experience, and learning in virtual museums.

	Item	Question
Engagement	E1	I enjoyed visiting this exhibition.
	E2	I felt engaged with the exhibition.
	E3	My visit to this exhibition was very interesting.
	E4	I felt I was experiencing the exhibition, rather than just visiting it.
	E5	My visit to the exhibition was inspiring.
Usability	U1	This system's capabilities meet my requirements.
	U2	This system is easy to use.
User experience	UX1	Obstructive (1) – supportive (7)
	UX2	Complicated (1) – easy (7)
	UX3	Inefficient (1) – efficient (7)
	UX4	Confusing (1) – clear (7)
	UX5	Boring (1) – exciting (7)
	UX6	Not interesting (1) – interesting (7)
	UX7	Conventional (1) – inventive (7)
	UX8	Usual (1) – leading edge (7)
Knowledge	K1	The information provided about the exhibits was clear.
	K2	I could make sense of most of the things I saw and did at the exhibition.
	K3	I liked the graphics associated with the exhibition.
	K4	My visit enriched my knowledge and understanding about specific exhibits.
	K5	I discovered new information from the exhibits.
Motivation	M1	How much have you enjoyed using this interaction method to access information?
	M2	How much are you willing to access information with this method in the future?
	M3	Compared to other learning methods (like books and video), how much are you willing to learn through this method?

4.3.1. Engagement

The Museum Experience Scale (MES) (Othman et al., 2011) includes a subscale that measures the engagement with five questions rated on a 5-point Likert scale (Table 1). Each question had the same weight and the mean of the five questions was reported as the *self-reported engagement*. In addition, we recorded system logs to obtain objective engagement measures. Specifically, we tracked the amount of *time users stayed in the museum environment* by calculating the difference between the times they entered and exited the exhibition room. In addition, a timer begins when users open the user interface for any interactive method and stops when they close it. The total duration of *time users spent accessing artifact information* in each room was compiled and used for analysis.

4.3.2. Experience

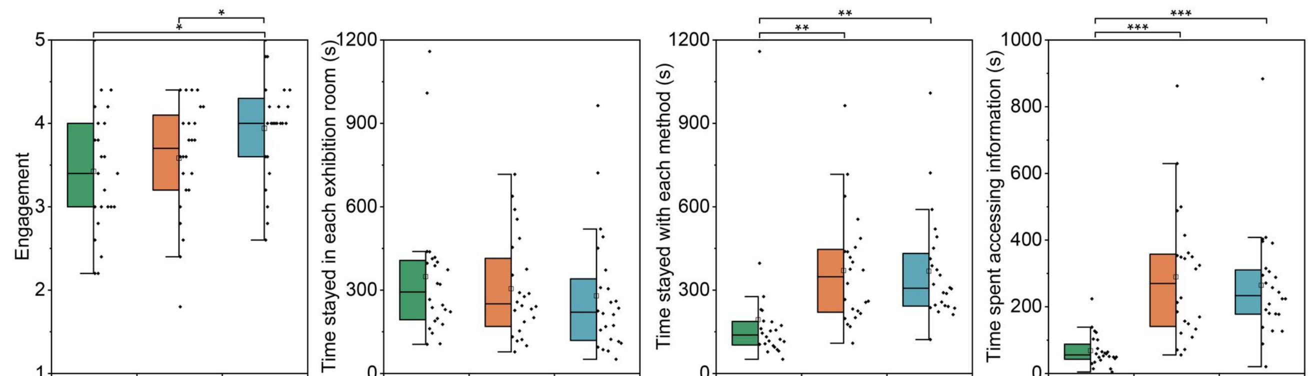
We invited participants to evaluate the *usability* of each interaction method using the two questions of UMUX-LITE (Lewis et al., 2013). Users evaluate it using a 7-point Likert scale. In addition, we adopted the short form of the user experience questionnaire (UEQ-S) (Schrepp et al., 2017). It consists of eight questions: four of these questions represent pragmatic quality and four hedonic quality aspects (Table 1). Items were rated on a 7-point Likert scale. The *user experience* was calculated using the provided toolkit (Schrepp et al., 2017).

4.3.3. Learning

The "Knowledge" subscale of the Museum Experience Scale (Othman et al., 2011) was adopted to understand the learning aspect (Table 1). It includes five questions that measure users' knowledge gained from the exhibition and exhibits. Users evaluate it using a 5-point Likert scale. Each question had the same weight and the mean of the five questions was reported as the *knowledge*. In addition, participants answered three questions about their *motivation* (Cordova & Lepper, 1996), rated on a 7-point Likert scale (Table 1). We followed a similar approach as Xu et al. (2024) to measure *learning outcome* using a self-designed Artifact Information Test (AIT). It consists of 36 multiple-choice questions, with 12 questions for the artifacts in each room. The questions cover aspects such as the material, era, size, and the collected museum of an artifact. Each question has five options, including an "I don't know" option placed at the end.

4.4. Participants

We had 24 participants (13 female, 11 male) who voluntarily signed up for the study, with an average age of 22.33 ($SD = 2.34$). Participants evaluated their knowledge of



	Label	Chatbot	Avatar	Room1	Room2	Room3	Label	Chatbot	Avatar	Label	Chatbot	Avatar
Mean	3.43	3.58	3.94	348.46	304.79	279.04	193.46	370.42	368.42	68.10	289.66	264.40
SD	0.74	0.69	0.61	251.82	180.56	219.07	218.92	199.81	192.83	48.41	194.71	166.85
		a			b			c			d	

Figure 5. Box-plots and tables of means and standard deviations showing the (a) engagement, (b) time stayed in each exhibition room, (c) time stayed with each interaction method and (d) time spent accessing artifact information when using text labels, chatbot, and avatar guide.

traditional Chinese culture (2.21±0.72), the frequency of museums visiting (2.46±0.93), and their knowledge of 3D graphics (3.29±1.20) on a 5-point Likert scale.

5. Results

In total, we collected 72 questionnaire samples (24 participants × 3 interaction methods). No significant outliers were found in the data, and we did not exclude any data from the experiment. Statistical analysis was performed using IBM SPSS Statistics 26. The Shapiro–Wilk tests of normality and the Q–Q plots showed that the distributions of engagement, usability, knowledge, motivation, and post-experiment AIT scores were approximately normally distributed. We performed repeated measures ANOVA to analyze the effects of chatbots and avatar agents on these measures. The distributions of other measures (i.e., pre-experiment AIT scores, time spent in the museum, and time accessing information) did not meet the normality test assumption. Therefore, we used the Friedman tests for the analysis. Greenhouse-Geisser correction was applied when the collected data did not satisfy the sphericity test assumption. Bonferroni adjustment was applied for *post-hoc* tests to avoid inflated Type I error. We report Bonferroni-adjusted *p*-values, which is compared against the threshold value of 0.05.

5.1. Engagement

Figure 5 shows the results about user engagement. A repeated measures ANOVA showed that engagement differed significantly among the three interaction methods, $F(1.58, 36.25) = 5.89, p = 0.010, \eta^2 = 0.204$. *Post hoc* analysis revealed that participants reported significantly higher engagement for the avatar guide than text labels ($p = 0.029$) and the chatbot ($p = 0.011$). H1a is partly supported.

A Friedman test showed no difference in participants' time stayed in the three exhibition rooms, $\chi^2(2) = 2.58, p = 0.275, W = 0.054$. However, there was a significant difference in the time users stayed with each interaction method, $\chi^2(2) = 16.08, p < 0.001, W = 0.335$. Users stayed significantly less time in an exhibition room with text labels, compared to the chatbot ($p = 0.001$) and the avatar guide ($p = 0.003$). H1b is supported. In addition, a Friedman test showed a significant difference in the time spent accessing information with the three methods, $\chi^2(2) = 36.33, p < 0.001, W = 0.757$. Users spent a significantly shorter time accessing information for the text label condition than the chatbot ($p < 0.001$) and avatar guide conditions ($p < 0.001$). H1c is supported.

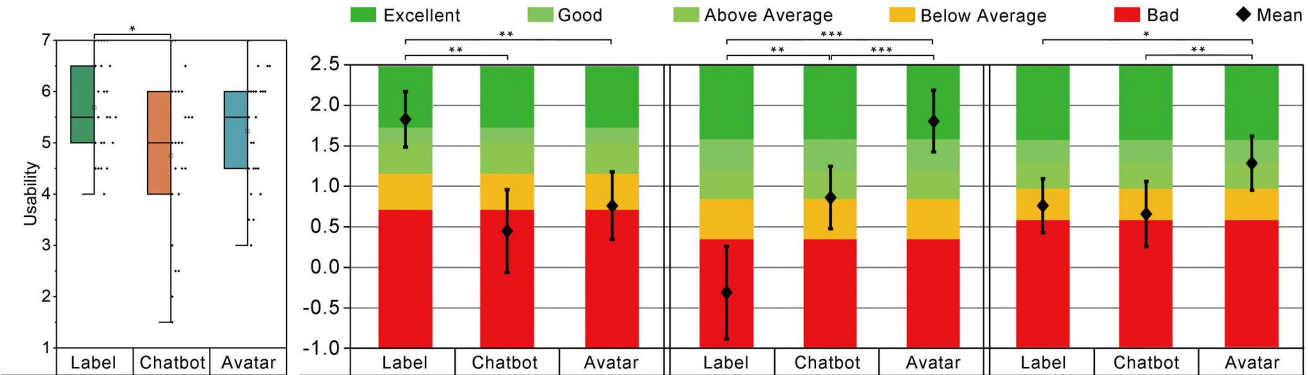
5.2. Experience

Figure 6 shows the results about user experience. A repeated measures ANOVA determined that usability of the three interaction methods differed significantly, $F(2, 46) = 4.35, p = 0.019, \eta^2 = 0.159$. The text labels showed significantly higher usability than the chatbot ($p = 0.043$). H2a is not supported.

The overall user experience differed significantly for the three methods, $F(2, 46) = 7.54, p = 0.001, \eta^2 = 0.247$. The avatar guide showed significantly greater user experience than the text labels ($p = 0.019$) and the chatbot ($p = 0.001$). Significant differences were shown for the pragmatic quality ($F(1.46, 33.53) = 14.69, p < 0.001, \eta^2 = 0.390$) and hedonic quality ($F(1.42, 32.60) = 32.00, p < 0.001, \eta^2 = 0.582$), but the pairwise comparison showed different patterns. The text labels showed significantly higher pragmatic quality than the chatbot ($p = 0.001$) and avatar guide ($p = 0.003$). On the other hand, the avatar guide showed significantly higher hedonic quality than the text labels ($p < 0.001$) and the chatbot ($p < 0.001$). The chatbot also showed significantly higher hedonic quality than text labels ($p = 0.001$).

5.3. Learning

Figure 7 shows the analysis results about learning. A repeated measures ANOVA showed no significant difference in knowledge when using the three methods, $F(2, 46) = 2.85, p = 0.068, \eta^2 = 0.110$. H3a is not supported. However,



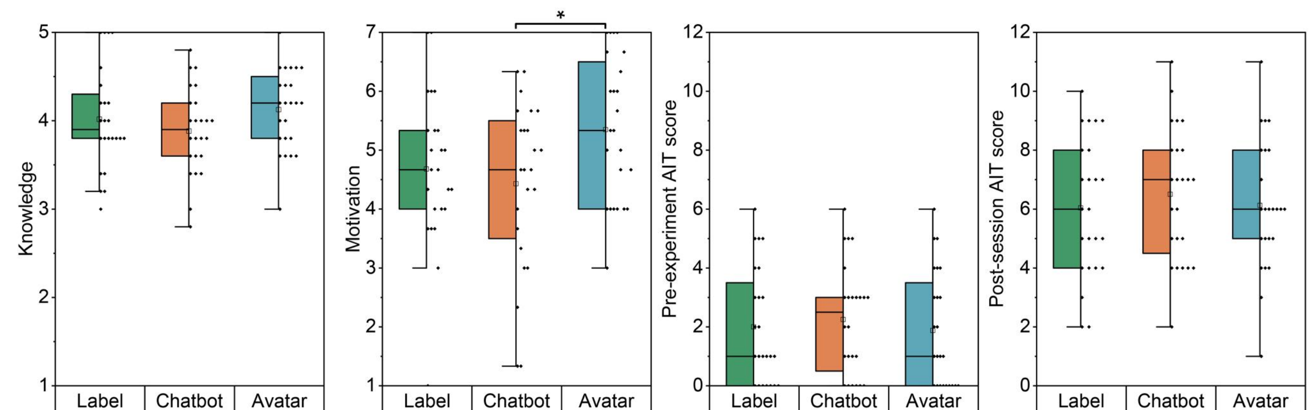
	Label	Chatbot	Avatar
Mean	5.69	4.75	5.23
SD	0.96	1.53	1.11

a

	Label	Chatbot	Avatar	Label	Chatbot	Avatar	Label	Chatbot	Avatar
Mean	1.85	0.46	0.77	-0.31	0.87	1.81	0.77	0.66	1.29
Scale	Pragmatic Quality			Hedonic Quality			Overall		

b

Figure 6. Box-plots and tables showing the (a) means and standard deviations of usability and (b) the pragmatic quality, hedonic quality, and overall user experience when using text labels, chatbot, and avatar guide.



	Label	Chatbot	Avatar		Label	Chatbot	Avatar		Label	Chatbot	Avatar		Label	Chatbot	Avatar
Mean	4.02	3.88	4.13		4.68	4.43	5.34		2.00	2.25	1.88		6.04	6.50	6.13
SD	0.59	0.49	0.46		1.90	1.43	1.22		1.96	1.87	1.96		2.33	2.27	2.23
	a				b				c				d		

Figure 7. Box-plots and tables of means and standard deviations showing the (a) knowledge, (b) motivation, (c) pre-experiment AIT scores and (d) post-session AIT scores when using text labels, chatbot, and avatar guide.

participants' motivation differed significantly, $F(2,46) = 4.52, p = 0.016, \eta^2 = 0.164$. *Post hoc* analysis revealed that users showed significantly higher motivation to access information when using the avatar guide than using the chatbot ($p = 0.003$). H3b is not supported.

The pre-experiment AIT scores showed no significant difference, $\chi^2(2) = 2.58, p = 0.276, W = 0.054$. Similarly, no significant difference was found for the post-session AIT scores, $F(2,46) = 0.291, p = 0.749, \eta^2 = 0.013$. H3c is not supported. Comparing participants' pre- and post-test AIT scores, paired-samples t tests showed significant improvements in their learning outcome when using text labels ($t(23) = -6.73, p < 0.001$), the chatbot ($t(23) = -9.41, p < 0.001$), and the avatar guide ($t(23) = -7.56, p < 0.001$).

5.4. Interview

One participant skipped the interview so we had 23 interview responses. Figure 8 shows the interview questions and results. Eleven participants preferred using the avatar guide for accessing information, while twelve favoured the text labels. This preference was primarily influenced by the actions required to access information, with some participants noting that the two LLM-based interaction methods were more complex than the traditional text labels ($N = 7$). However, more than half of the participants reported a better overall experience with the avatar guide ($N = 14$). They found the question-based approach engaging ($N = 3$) and appreciated the presence of an avatar for communication ($N = 7$). In contrast, three participants preferred the text labels due to their directness and effectiveness. The remaining six participants indicated no strong preference.

Participants reported some issues during the experiment. Two users mentioned that the response time of the LLMs was sometimes longer than expected. P10 raised concerns about the accuracy of LLM responses. We examined the system logs related to this issue and found that, while most answers were correct, the italicized sentence did not originate from the provided source. In addition, one participant noted the difficulty for the LLMs in providing appropriate answers.

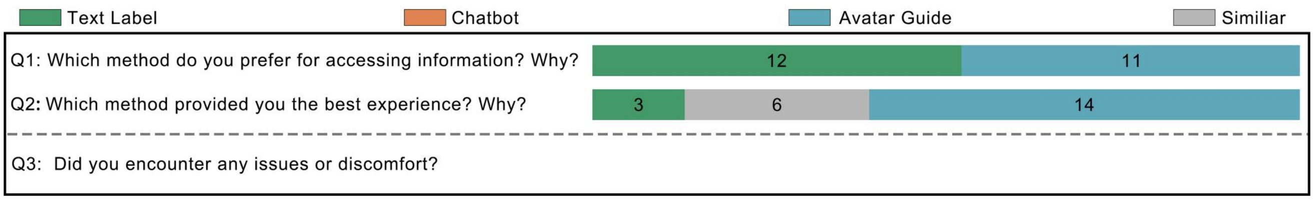


Figure 8. Semi-structured interview questions and results.

Q: Tell me the story about the kneeling archer.

Ans: The kneeling archer is a kind of pottery in ancient China, *also known as* "bronze man," *which was used for ancient military training and exercises*. It was discovered in Pit 2 of the Terracotta Warriors and Horses in 221 BC and is one of the cultural relics collected by the Emperor Qinshihuang's Mausoleum Site Museum. The kneeling archer is wearing a battle robe, a helmet, and holding a crossbow. The vivid shape shows the scene of ancient soldiers preparing for battle.

Q: Why is the statue in the UK?

Ans: Hello, based on the known information, I cannot understand your question. The database does not have such an answer.

Participants also provided some suggestions for improvement. P17 noted that the question-based methods in the LLMs were not very friendly at the beginning because they were too specific. P2 suggested displaying both text labels and avatar guide together. P23 recommended linking artifact information to specific visual features on the 3D model. Using the elephant-shaped wine server as an example, the participant explained that a 3D display with pointing links could be used to highlight the baby elephant standing on a larger elephant's back. Beyond keyboard typing, participants expressed interest in using voice inputs ($N = 4$) and receiving answers through voice outputs ($N = 4$). In terms of the design of the virtual museum environment, some wished to explore a larger exhibition space ($N = 2$). Additionally, one user hoped to see more artifacts in the exhibition room, while another preferred fewer artifacts, as it was hard for him to learn about six artifacts at once.

6. Discussion and implications

In this section, we discuss the effect of LLM-empowered chatbots and avatar guides on engagement, experience, and learning in virtual museums. We also summarize some design implications for future serious games design, and discuss the limitations and future work.

6.1. Interpreting the results

Can LLM-empowered chatbots and avatar guides improve users' engagement in virtual museums? Our study showed the highest level of user engagement with the avatar guide, indicating a positive answer to this research question. Many users found conversing with the avatar to be a novel and captivating experience, which enhanced their motivation to seek out information. For instance, P9 remarked, "The avatar looks cute and I really felt I was talking to it. It's an interesting experience." In addition, participants spent significantly longer time with the two LLM-empowered methods than the text labels. One notable reason is that the interactions with the LLMs require more time than directly triggering the text labels. Still, participants enjoyed the question-driven approach in accessing information. For example, P10 noted, "I prefer to ask my own questions instead of having all the information thrown at me at once." Participants found the quick buttons useful and necessary to get started with. Besides, many participants asked questions beyond what was included in the text descriptions. For example, "Who are you?" "How many people like this bronze mask?" In these cases, the LLMs improvised the answers:

Q: Who are you?

Ans: Hello, I'm ChatGLM, an AI assistant. Nice to meet you! As an AI assistant, my goal is to answer your questions as accurately and clearly as possible and provide useful information. How can I help you?

Q: How many people like this bronze mask?

Ans: According to known information, bronze masks are loved by many people. In different cultures and traditions, bronze masks have important symbolic meanings and are regarded as a symbol of respect and power. It is a very interesting art form that can inspire people's creativity and imagination.

Can LLM-empowered chatbots and avatar guides improve users' experience in virtual museums? Overall, the avatar guide demonstrated the greatest user experience, and consistent results were acknowledged in the interview, showing supportive answers to this research question. As P19 mentioned, "I love the companion of the avatar during the visit." P12 echoed, "The avatar guide makes me feel like I'm having a conversation with someone, not just by myself." However, the user experience with the chatbot was below average. Participants identified some limitations of the two LLM-empowered methods that negatively influenced their overall experience. First, those who were goal-oriented preferred to receive all relevant information directly through text labels. They found the LLM-based methods to be more complicated, requiring additional actions such as formulating questions and typing them out, which reduced the efficiency of information retrieval. In addition, the unexpected response time and the occasionally inaccurate answers provided by the two LLM methods further affected user experience. For instance, P14 remarked, "Sometimes the response time was longer than I expected. I was a bit impatient to wait for the answers." In addition, P10 noted, "I remembered I asked the differences of materials between two artifacts, and the AI said the differences between two artifacts. This is apparently not what I asked for. I was doubtful of the accuracy of the AI responses." This highlights the need for improvements in both the speed and reliability of responses to enhance the user experience. While our controlled experimental study examined the independent use of the traditional text labels or the LLMs for a fair comparison, integrating both in the design of virtual museums is a feasible option. This could involve using text labels as the default approach while also providing avatar guides for additional information. As P2 noted, "I hope the system could provide the text labels and avatar guide together and let me decide which one I hope to use."

Can LLM-empowered chatbots and avatar guides improve users' learning in virtual museums? Our results did not indicate a significant impact of LLMs on user's learning. Participants demonstrated similar learning outcomes using the three interaction methods. In addition, they spent significantly less time using the traditional text labels than the two LLM-empowered methods, suggesting that the traditional way is straightforward and efficient. We believe this is relevant to how the information is presented to users. As P2 noted, "The text labels were clear, and I could see all information directly on the screen. I didn't need to think about what questions to ask step by step." Besides, this method was found to be more user-friendly for those who are not familiar with LLMs. Some participants felt they accessed limited information through the LLM methods. For instance, P5 noted, "I had to think about how to ask questions when using AI." P6 echoed, "Sometimes I don't know what question I could ask." Nevertheless, some participants acknowledged the benefits of the chatbot and the avatar guide in providing rich answers beyond the text descriptions. For example, P13 said: "I can always ask whatever I want to know." Participants found this approach more tailored to their interests, like P5 commented, "The information on the text labels was too long to read. I had to read carefully to search for the answers. But with AI, I can get very accurate results." P11 also noted that using AI allowed him to ask the same question multiple times, reinforcing his memory of the information, whereas with text labels, he could easily forget parts after reading. Participants who were not fond of text labels found it challenging to locate specific information with everything displayed on the screen. They also considered the presentation too generic to capture their interest.

6.2. Lessons learned

Based on the study results and users' feedback, we proposed the following suggestions that could implicate the future design of LLM-empowered serious games.

6.2.1. Provide suitable avatars for the LLMs

Given that user interaction with the LLMs is fundamentally conversational, offering an appropriate avatar that aligns with users' mental models significantly enhances immersion. This embodiment fosters a more engaging experience, as evidenced by the positive outcomes associated with avatar guides compared to traditional chatbots. One participant (P4) even expressed a desire for the avatar to actively interact with him, highlighting the potential for dynamic engagement. Such interactive avatars not only enrich the user experience but also create a sense of presence, making users feel more connected to the environment. By integrating movement and responsiveness, avatars can transform conversations into immersive dialogues, reinforcing the effectiveness of LLM-empowered interactions.

6.2.2. Keep the system response concise for sustained engagement

Participants have identified the response time affected their user experience. A closer examination of the system logs showed that some answers were perhaps too long (> 250 Chinese characters). Therefore, we recommend limiting the length of responses to prevent prolonged wait times, with an optimal length being around 150 Chinese characters, equivalent to about 2–3 sentences. Shorter feedback not only reduces wait times but also encourages users to continue interacting with the system without frustration. This approach aligns with best practices in user interface design, where clarity and efficiency are prioritized to enhance user satisfaction.

6.2.3. Understand the pronouns and user intentions

In our quick buttons, we utilized the term "it" to refer to the artifact of interest. However, our system logs revealed that many users preferred to use the word "you" in their inquiries. For instance, questions like "Who are you?" were common. While the responses addressed the tool behind the avatar guide (i.e., ChatGLM), users likely intended to refer to a specific artifact. This indicates a more personal and engaging interaction style, suggesting that users may perceive the system or avatar as a conversational partner rather than just a tool. This shift in language reflects a desire for a more relatable and interactive experience, highlighting the importance of designing interfaces that resonate with users on a personal level. Such insights can inform future enhancements to improve user engagement and satisfaction.

6.2.4. Text labels and avatar guide complement each other in overview and details, contributing to effective learning

As previously discussed, while the text labels allow for high efficiency, it was challenging for users to locate specific information. In contrast, the avatar guide provides a more engaging and enjoyable experience, making it easier to search for targeted details. By integrating both text labels and the avatar guide, users can choose their preferred interaction method based on their individual needs and preferences. This flexibility enhances the overall user experience, allowing for a more tailored approach to information retrieval. This dual approach can lead to a more immersive and fulfilling experience, ultimately improving user engagement and satisfaction in cultural heritage contexts.

6.2.5. Design for LLM credibility

Serious games enhanced by LLMs have broadened the scope of knowledge by incorporating a larger database and related content. This encompasses not only specific textual descriptions of artifacts but also general information about their craftsmanship, materials, and historical context from the same era. Such information are domain-specific and should be derived from credible research documents that contain specialized knowledge. It can pose challenges for LLMs in accurately matching responses to generic databases if this information is not provided. For instance, users might ask about the differences between various pottery-making techniques (e.g., the distinction between Tang Sancai and grey pottery), the definition of lacquerware, or why a particular artifact is in a foreign country but not China. If the knowledge base lacks this information, the accuracy of the LLM-generated responses cannot be assured. In situations where the system fails to comprehend an user's question, it is preferable to indicate that the question cannot be answered rather than generating irrelevant content, as incorrect or irrelevant answers may detract from the user's experience and willingness to learn. In the meantime, we recommend that content produced by LLMs be clearly labeled with sources. We have observed instances where responses mixed information from the provided knowledge base with that from generic language models. For example, when an user inquires about the use of an artifact (e.g., a bronze ding vessel), the LLM might include characteristics of other ding vessels that do not accurately apply to the specific one being presented. Therefore, it is essential to highlight the sources of information and to differentiate between reliable information and that which is merely for reference.

6.3. Limitations and future work

The study has several limitations that should be addressed in future work. A primary limitation is the separate assessment of LLM-empowered chatbots and avatar guides, apart from text labels. Evaluating these elements individually might have led us to miss important insights into how they could synergistically improve the user experience. In addition, the user characteristics among participants were homogeneous, which may restrict the generalizability of the findings. Additionally, we did not assess users' familiarity with AI technologies, an oversight that could influence their engagement and experience with LLMs. Furthermore, the results were derived solely from computer-based interactions, lacking evaluations in VR environments, thereby restricting the applicability of our conclusions in those contexts. Lastly, it is challenging to ensure the complete accuracy of the responses generated by the LLMs. Participants reported encountering inaccurate answers and instances where their queries were not responded to properly. This issue is particularly critical for serious educational games, where accuracy is essential for effective learning. Improving the reliability of LLM responses is necessary to enhance user experience and engagement. Additionally, incorporating voice inputs could further enrich interactions, making them more intuitive and accessible. Addressing these limitations will enhance the robustness of future research and the effectiveness of LLM-empowered interactions.

7. Conclusion

The motivation for this work stems from identifying the effects of chatbots and avatar guides based on large-language models (LLMs) on users' engagement, experience and learning, offering insights for the future design of LLM-supported interaction methods in virtual museums. Through our empirical evaluation, we observed significantly different engagement and experience among the three interaction methods. Specifically, users felt significantly higher engagement when they used the avatar guide than the text labels and chatbot. Meanwhile, users stayed longer time in the virtual museum when they used the two LLM-empowered interaction methods (chatbot and avatar guide). For the experience aspect, the text labels showed better usability than the chatbot. However, the avatar guide demonstrated significantly greater hedonic quality and overall user experience. For learning outcomes, the avatar guide demonstrated higher motivational levels, although there were no significant differences in knowledge acquisition or test scores among the three methods. Our results and the recommendations derived from the user study can assist researchers and designers in developing future LLM-empowered interaction methods for serious games.

Note

1. https://helixngc7293.itch.io/yandere-ai-girlfriend-simulator.

Acknowledgement

We acknowledge the support from the XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU and SIP AI innovation platform (YZCXPT2022103), Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004), and XJTLU Key Program Special Fund (KSF-A-17). We would like to thank our participants for their time and valuable comments.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work is supported by the National Natural Science Foundation of China [62207022]. This work was also supported by Jiangsu Province Engineering Research Center of Data Science and Cognitive Computing at XJTLU and SIP AI innovation platform [YZCXPT2022103], Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering [SZS2022004], and XJTLU Key Program Special Fund [KSF-A-17].

ORCID

Shuhao Zhang http://orcid.org/0000-0001-8504-8078
Yue Li http://orcid.org/0000-0003-3728-218X
Ka Lok Man http://orcid.org/0000-0002-5787-4716
Yong Yue http://orcid.org/0000-0001-7695-4538

References

Azizi, Z., Alipour, P., Gomez, S., Broadwin, C., Islam, S., Sarraju, A., Rogers, A. J., Sandhu, A. T., & Rodriguez, F. (2023). Evaluating recommendations about atrial fibrillation for patients and clinicians obtained from chat-based artificial intelligence algorithms. *Circulation, Arrhythmia and Electrophysiology*, *16*(7), 415–417. https://doi.org/10.1161/CIRCEP.123.012015

Bernabei, M., Colabianchi, S., Falegnami, A., & Costantino, F. (2023). Students' use of large language models in engineering education: A case study on technology acceptance, perceptions, efficacy, and detection chances. *Computers and Education: Artificial Intelligence*, *5*, 100172. https://doi.org/10.1016/j.caeai.2023.100172

Bönsch, A., Hashem, D., Ehret, J., & Kuhlen, T. W. (2021). Being guided or having exploratory freedom: User preferences of a virtual agent's behavior in a museum. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents* (pp. 33–40). Association for Computing Machinery. https://doi.org/10.1145/3472306.3478339

Capece, N., Manfredi, G., Gargiulo, B., Gilio, G., Erra, U., & Sogliani, F. (2024). A methodological approach for unveiling the evolution of frescoes through a virtual reality. In *2024 IEEE International Conference on Metrology for Extended Reality, Artificial Intelligence and Neural Engineering (Metroxraine)* (pp. 365–370). IEEE. https://doi.org/10.1109/MetroXRAINE62247.2024.10796372

Cheung, M. (2024). A Reality check of the benefits of LLM in business. *ArXiv, abs/2406.10249*. https://doi.org/10.48550/arXiv.2406.10249

Ch'ng, E., Li, Y., Cai, S., & Leow, F.-T. (2020). The effects of VR environments on the acceptance, experience, and expectations of cultural heritage learning. *Journal on Computing and Cultural Heritage*, *13*(1), 1–21. https://doi.org/10.1145/3352933

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization, personalization, and choice. *Journal of Educational Psychology*, *88*(4), 715–730. https://doi.org/10.1037/0022-0663.88.4.715

Di Natale, A. F., Simonetti, M. E., La Rocca, S., & Bricolo, E. (2023). Uncanny valley effect: A qualitative synthesis of empirical research to assess the suitability of using virtual faces in psychological research. *Computers in Human Behavior Reports*, *10*, 100288. https://doi.org/10.1016/j.chbr.2023.100288

Ding, A.-C. E., & Yu, C.-H. (2024). Serious game-based learning and learning by making games: Types of game-based pedagogies and student gaming hours impact students' science learning outcomes. *Computers & Education*, *218*, 105075. https://doi.org/10.1016/j.compedu.2024.105075

Feng, Z., González, V. A., Amor, R., Spearpoint, M., Thomas, J., Sacks, R., Lovreglio, R., & Cabrera-Guerrero, G. (2020). An immersive virtual reality serious game to enhance earthquake behavioral responses and post-earthquake evacuation preparedness in buildings. *Advanced Engineering Informatics*, *45*, 101118. https://doi.org/10.1016/j.aei.2020.101118

Harmon, S., & Rutman, S. (2023). Prompt engineering for narrative choice generation. In *Interactive storytelling: 16th International Conference on Interactive Digital Storytelling, ICIDS 2023, Kobe, Japan, November 11–15, 2023, proceedings, part I* (pp. 208–225). Springer-Verlag. https://doi.org/10.1007/978-3-031-47655-6_13

Hu, S., Huang, T., Ilhan, F., Tekin, S., Liu, G., Kompella, R., & Liu, L. (2024). A survey on large language model-based game agents. *ArXiv*. https://doi.org/10.48550/arXiv.2404.02039

Huang, T., Li, Y., & Liang, H.-N. (2024). Avatar type, self-congruence, and presence in virtual reality. In *Proceedings of the Eleventh International Symposium of Chinese Chi* (pp. 61–72). Association for Computing Machinery. https://doi.org/10.1145/3629606.3629614

Katz, A., Shakir, U., & Chambers, B. (2023). The utility of large language models and generative AI for education research. *Arvix*. https://doi.org/10.48550/arXiv.2305.18125

Kopp, S., Gesellensetter, L., Křamer, N. C., & Wachsmuth, I. (2005). A conversational agent as museum guide: Design and evaluation of a real-world application. In *Lecture notes in computer science* (pp. 329–343). Springer-Verlag. https://doi.org/10.1007/11550617_28

Kumaran, V., Rowe, J., Mott, B., & Lester, J. (2023). SCENECRAFT: Automating interactive narrative scene generation in digital games with large language models. In *Proceedings of the Nineteenth AAAI conference on artificial intelligence and interactive digital entertainment*. AAAI Press. https://doi.org/10.1609/aiide.v19i1.27504

LangChain. (2024). *LangChain*. https://langchain.com/

Lewis, J. R., Utesch, B., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM (Association for Computing Machinery). https://doi.org/10.1145/2470654.2481287

Li, Y., Ch'ng, E., & Cobb, S. (2023). Factors influencing engagement in hybrid virtual and augmented reality. *ACM Transactions on Computer-Human Interaction*, *30*(4), 1–27. https://doi.org/10.1145/3589952

Li, Y., Tennent, P., & Cobb, S. (2019). Appropriate control methods for mobile virtual exhibitions. In M. Duguleaňa, M. Carrozzino, M. Gams, & I. Tanea (Eds.), *VR technologies in cultural heritage* (pp. 165–183). Springer International Publishing. https://doi.org/10.1007/978-3-030-05819-7_13

Liu, Q., Song, J., Huang, Z., & Zhang, Y. (2024). Glide the, & liunux4odoo. *langchain-chatchat*. https://github.com/chatchat-space/Langchain-Chatchat. *GitHub*.

Machidon, O.-M., Tavčar, A., Gams, M., & Duguleană, M. (2020). CulturalERICA: A conversational agent improving the exploration of European cultural heritage. *Journal of Cultural Heritage*, *41*, 152–165. https://doi.org/10.1016/j.culher.2019.07.010

Mirza-Babaei, P., Kamkarhaghighi, M., & Gerling, K. (2014). Opportunities in game-based stroke rehabilitation. In *2014 IEEE games media entertainment* (pp. 1–4). IEEE. https://doi.org/10.1109/GEM.2014.7118436

OpenAI. (2022). *ChatGPT*. https://chat.openai.com/chat

Othman, M. K., Petrie, H., & Power, C. (2011). Engaging visitors in museums with technology: scales for the measurement of visitor and multimedia guide experience. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *Human-computer interaction – INTERACT 2011.* (pp. 92–99). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23768-3_8

Paduraru, C. I., Cernat, M., & Stefanescu, A. (2023). *Conversational agents for simulation applications and video games.* International Conference on Software and Data Technologies. https://doi.org/10.5220/0012060500003538

Petridis, P., Dunwell, I., Liarokapis, F., Constantinou, G., Arnab, S., de Freitas, S., & Hendrix, M. (2013). The Herbert Virtual Museum. *Journal of Electrical and Computer Engineering*, *2013*, 487970. https://doi.org/10.1155/2013/487970

Pilote, B., & Chiniara, G. (2019). Chapter 2 – the many faces of simulation. In G. Chiniara (Ed.), *Clinical simulation* (Second Edition, pp. 17–32). Academic Press. https://doi.org/10.1016/B978-0-12-815657-5.00002-4

Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S. G., Novikov, A., Barth-Maron, G., & de Freitas, N. (2022). *A generalist agent. Arvix*. http://doi.org/10.48550/arXiv.2205.06175

Rehm, M., & Jensen, M. L. (2015). Accessing cultural artifacts through digital companions: The effects on children's engagement. In *2015 International Conference on Culture and Computing (Culture Computing)* (pp. 72–79). IEEE. https://doi.org/10.1109/Culture.and.Computing.2015.44

Ribeiro, M., Santos, J., Lobo, J. a., Araújo, S., Magalhães, L., & Adão, T. (2024). VR, AR, gamification and AI towards the next generation of systems supporting cultural heritage: Addressing challenges of a museum context. In *Proceedings of the 29th international ACM* *conference on 3D web technology*. Association for Computing Machinery. https://doi.org/10.1145/3665318.3677172

Saito, M. (2023). Effects of presentation modalities in virtual museum guides on agent impressions and painting evaluations. In *Proceedings of the 11th International Conference on Human-Agent Interaction* (pp. 446–448). Association for Computing Machinery. https://doi.org/10.1145/3623809.3623958

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Design and evaluation of a short version of the user experience questionnaire (UEQ-S). *International Journal of Interactive Multimedia and Artificial Intelligence*, *4*(6), 103–108. https://doi.org/10.9781/ijimai.2017.09.001

Seßler, K., Kepir, O., & Kasneci, E. (2024). Enhancing student motivation through LLM-powered learning environments. In R. Ferreira Mello, N. Rummel, I. Jivet, G. Pishtari, & J. A. Ruipérez Valiente (Eds.), *Technology enhanced learning for inclusive and equitable quality education* (pp. 156–162). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-72312-4_21

Tennent, P., Martindale, S., Benford, S., Darzentas, D., Brundell, P., & Collishaw, M. (2020, May). Thresholds: Embedding virtual reality in the museum. *Journal on Computing and Cultural Heritage*, *13*(2), 1–35. https://doi.org/10.1145/3369394

Tian, F., Li, K., Huang, X., Zhang, X., Wang, N., Song, Y., Zhu, Q., & Li, Y. (2024). An empirical study of virtual museum based on dual-mode mixed visualization: The Sanxingdui bronzes. *Heritage Science*, *12*(1), 1–14. https://doi.org/10.1186/s40494-024-01241-1

Tsiantis, V., Konstantinidis, D., & Dimitropoulos, K. (2024). ChatGPT in nutrition: Trends challenges and future directions. In *Proceedings of the 17th International Conference on Pervasive Technologies Related to Assistive Environments* (pp. 548–553). Association for Computing Machinery. https://doi.org/10.1145/3652037.3663898

Tsita, C., Satratzemi, M., Pedefoudas, A., Georgiadis, C., Zampeti, M., Papavergou, E., Tsiara, S., Sismanidou, E., Kyriakidis, P., Kehagias, D., & Tzovaras, D. (2023). A virtual reality museum to reinforce the interpretation of contemporary art and increase the educational value of user experience. *Heritage*, *6*(5), 4134–4172. https://doi.org/10.3390/heritage6050218

Vardhan, H., Saxena, A., Dixit, A., Chaudhary, S., & Sagar, A. (2022). AR museum: A virtual museum using marker less augmented reality system for mobile devices. In *2022 3rd International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICICT55121.2022.10064611

Vasic, I., Fill, H. -G., Quattrini, R., & Pierdicca, R. (2024). LLM-aided museum guide: Personalized tours based on user preferences. In L. T. De Paolis, P. Arpaia, & M. Sacco (Eds.), *Extended reality*. (pp. 249–262). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-71710-9_18

Volum, R., Rao, S., Xu, M., DesGarennes, G., Brockett, C., Durme, B. V., & Dolan, B. (2022). Craft an iron sword: Dynamically generating interactive game characters by prompting large language models tuned on code. *Proceedings of the 3rd Wordplay: When Language Meets Games Workshop (Wordplay 2022)* (pp. 25–43). Seattle, United States: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.wordplay-1.3

Wang, X., Li, Y., & Liang, H. -N. (2024). *MagicMap: Enhancing indoor navigation experience in VR museums*. In *2024 IEEE Conference Virtual Reality and 3D User Interfaces (VR)* (pp. 881–891). IEEE. https://doi.org/10.1109/VR58804.2024.00107

Wang, Y., Li, Y., & Liang, H.-N. (2023). Comparative analysis of artefact interaction and manipulation techniques in VR museums: A study of performance and user experience. In *2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (pp. 761–770). IEEE. https://doi.org/10.1109/ISMAR59233.2023.00091

Xu, N., Li, Y., Lin, J., Yu, L., & Liang, H.-N. (2022). User retention of mobile augmented reality for cultural heritage learning. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (pp. 447–452). IEEE. https://doi.org/10.1109/ISMAR-Adjunct57072.2022.00095

Xu, N., Li, Y., Wei, X., Xie, L., Yu, L., & Liang, H.-N. (2024). CubeMuseum AR: A tangible augmented reality interface for cultural heritage learning and museum gifting. *International Journal of Human–Computer Interaction*, *40*(6), 1409–1437. https://doi.org/10.1080/10447318.2023.2171350

Xue, Y., Chen, H., Bai, G. R., Tairas, R., & Huang, Y. (2024). Does ChatGPT help with introductory programming? An experiment of students using ChatGPT in CS1. In *Proceedings of the 46th International Conference on Software Engineering: Software Engineering Education and Training* (pp. 331–341). Association for Computing Machinery. https://doi.org/10.1145/3639474.3640076

Yong, Q. R., & Mitchell, A. (2023). From playing the story to gaming the system: Repeat experiences of a large language model-based interactive story. In *Interactive storytelling: 16th International Conference on Interactive Digital Storytelling, ICIDS 2023, Kobe, Japan, November 11–15, 2023,* Proceedings, *Part I* (pp 395–409). Springer-Verlag. https://doi.org/10.1007/978-3-031-47655-6_24

Zhipu, A. I. (2024). *ChatGLM*. https://chatglm.cn/

About the authors

Shuhao Zhang received his MDes degree in Industrial Design from the Xi'an Jiaotong-Liverpool University in 2021. He is currently a PhD student in Computer Science & Software Engineering at the Xi'an Jiaotong-Liverpool University, China. His research interests include human-computer interaction, virtual reality, cross-reality, & asymmetric collaboration.

Mingge Ma received his BSc degree in Information & Computing Science from the Xi'an Jiaotong-Liverpool University in 2024. His research interests include human-computer interaction, generative artificial intelligence, & virtual reality.

Yue Li received her PhD degree in Computer Science from the University of Nottingham in 2020. She is an Assistant Professor at the Department of Computing at Xi'an Jiaotong-Liverpool University. Her research interests lie in the field of Human-Computer Interaction, with a particular emphasis on Virtual Reality and Augmented Reality.

Ka Lok Man received his PhD degree in Computer Science from Technische Universiteit Eindhoven in 2006. He is currently a Professor at the Department of Computing, Xi'an Jiaotong-Liverpool University, Suzhou, China. His research interests include formal methods & process algebras, embedded system design & testing, & photovoltaics.

Jeremy Smith received his PhD degree in Electrical Engineering from the University of Liverpool in 1990. He has held a professorship position since 2006 in Electrical Engineering with the University of Liverpool. His research interests include automated welding, robotics, vision systems, adaptive control, & embedded computer systems.

Yong Yue is a Professor at the Department of Computing & the Director of the Virtual Engineering Centre (VEC) & the Suzhou Municipal Key Lab for Intelligent Virtual Engineering at Xi'an Jiaotong-Liverpool University, Suzhou, China. His current research interests include virtual reality, computer vision, robot applications, & operations research.