Loretti *et al. Journal of Big Data* (2026) 13:84
https://doi.org/10.1186/s40537-026-01418-4

Journal of Big Data

METHODOLOGY Open Access

Investigating the role of conversational agents in augmented cultural heritage tours

Andrea Loretti[1], Valentine Bernasconi[1], Pasquale Cascarano[1*], Alessio Troffei[1], Mengting Lai[1], Luca Vitale[2], Andrea Bortolotti[2] and Gustavo Marfia[1]

*Correspondence:
Pasquale Cascarano
pasquale.cascarano2@unibo.it
[1]Department of the Arts, University of Bologna, Bologna, Italy
[2]Touchlabs, Bologna, Italy

Abstract

Recent technological advancements in generative Artificial Intelligence (AI) models combined with Augmented Reality (AR) systems represent a new opportunity for cultural heritage valorization, particularly in the context of increasingly large, heterogeneous, and multimodal digital cultural repositories that pose challenges in terms of scalable access and semantic retrieval. In this article, we introduce ARtour, an application for cultural tours implementing a navigation system to direct users toward a point of interest and provide information through an interactive Large Language Model (LLM)-based audio system. Beyond enhancing user experience, the system explores the role of LLMs as conversational interfaces for accessing structured and unstructured cultural heritage data within digital twin environments, while also examining how such conversational interaction influences users' experience, including usability, cognitive workload, and affective responses during cultural heritage exploration. A user study comparing LLM-based interaction with traditional web search shows that the system achieves high usability and technology acceptance with low cognitive workload. While task performance remains comparable across conditions, the AI conversational agent enhances user engagement and supports a more exploratory information-seeking behavior, increasing perceived immersion and co-presence and providing insights into the role of conversational agents in human–machine interaction within multimodal AR environments. We finally argue that the integration of fine-tuned or retrieval-augmented models for accessing information on historical and cultural artifacts could have a constructive impact on the promotion of publicly accessible cultural heritage, while raising important considerations regarding data accuracy, provenance, and governance.

Keywords Augmented reality, Extended reality, Digital twin, Guided tour, Artificial Intelligence, Intelligent assistant

Introduction

In many regions of the world, tangible cultural heritage is easily accessible and represents an integral part of our daily visual context. Yet, the inexperienced observer often lacks reading keys, adequate access to information systems, and guidance for self-education.

Unfortunately, the history behind a fresco seen daily on the ceiling of an office or classroom frequently remains unknown.

At the same time, ongoing digitization processes are progressively transforming cultural heritage into large and diverse digital collections, including textual archives, high-resolution imagery, 3D models, and structured metadata, which further increases the complexity of organizing and accessing cultural knowledge [1]. Innovative solutions are required to quickly generate and ease access to this knowledge, and to enrich daily interactions with cultural artifacts in public environments.

Technological advances in Augmented Reality (AR) enable the creation of lightweight applications for portable devices. Such AR applications modernize the engagement with cultural objects while preserving their physical integrity [2, 3]. Through multimodal smartphone capabilities such as image and audio output, more immersive and polyvalent educational systems can be imagined [4]. Users not only perceive the world through their device lenses, gaining new insights, but also receive audio explanations of visual elements, providing guidance. By combining visual perception, spatial context, and audio interaction, AR systems enable new forms of multimodal human–machine interaction for exploring complex cultural information spaces [5].

The recent integration of Large Language Models (LLMs) into these portable systems represents a decisive breakthrough for automated information production and human-machine interactions. LLMs can automatically provide additional information based on the request of the user without the laborious manual annotation of artifacts [6–9]. Despite valid concerns about data accuracy, traditional issues such as big data annotation, specific ontologies, and retrieval systems can thus be mitigated [10–12]. Users can progressively request information which enhances their overall experience and fosters a sense of presence [5, 13]. Consequently, AR systems become a novel window for better understanding everyday environments [2].

From a data-centric perspective, LLM-integrated AR systems can also be interpreted as scalable semantic access layers over heterogeneous cultural heritage repositories, raising questions of information quality, provenance transparency, and governance.

Understanding how users interact with conversational AI in such contexts therefore requires considering not only technical performance but also experiential and affective aspects of the interaction. These include perceived usability, cognitive workload, emotional responses, and sense of presence during human–machine interaction with complex information systems.

In this article, we introduce *ARtour*, an innovative AR application for cultural tours that applies to any cultural site, enabling the creation of scalable and customizable virtual tour experiences. *ARtour* includes a navigation system directing users to nearby points of interest and integrates an interactive audio system powered by a LLM to enlighten users about cultural artifacts. Beyond concerns about information quality from the use of LLM, the system provides a controlled framework to investigate how conversational AI can mediate access to digitally available cultural heritage data while preserving user-centered interaction quality. More specifically, *ARtour* allows us to study how conversational agents embedded in AR environments influence users' interaction experience when navigating and exploring complex cultural heritage information spaces. Through a controlled user study, we evaluate not only usability and information retrieval performance, but also experiential dimensions of the interaction, including cognitive workload,

emotional responses, and perceived co-presence. From the perspective of affective computing, this work contributes to understanding how conversational AI integrated within multimodal AR environments can shape users' affective responses and perceived presence during cultural heritage exploration. At the same time, the study highlights the relevance of ethical AI considerations when generative models are deployed as interfaces to cultural heritage data. In particular, issues related to information reliability, provenance transparency, and responsible AI usage become central when conversational systems mediate access to cultural knowledge.

Specifically, the following Research Questions (RQs) guide this study:

RQ1: How does integrating an LLM into AR-based cultural heritage applications influence users' perceived usability, cognitive workload, and emotional experiences compared to traditional web search methods when accessing digital cultural heritage information?

RQ2: To what extent does interacting with an LLM-driven AR assistant enhance users' sense of co-presence during the exploration of digital cultural heritage content?

RQ3: Does the use of an LLM within an AR application improve information retrieval efficiency and consistency compared to conventional methods in the context of accessing cultural heritage data?

The manuscript is organized as follows. In Sect. "Related work" we review related work on LLM-driven assistants in AR environments and cultural heritage applications. In Sect. "Methodology" we describe ARtour, detailing the system architecture, implementation choices, and the user study design. In Sect. "Results" we report the quantitative results of the comparison between the considered conditions. In Sect. "Discussion, limitations and future works" we discuss the findings in relation to the research questions, outline limitations, and identify directions for future work. Finally, Sect. "Conclusions" concludes the paper.

Related work

Numerous research projects and commercial applications have explored the intersection of AR, indoor navigation, and intelligent assistance for tourism and cultural heritage where large and heterogeneous digital collections must be accessed through intuitive interfaces. Among the most established commercial solutions is Navigine AR Indoor Navigation[1] which leverages AR to deliver indoor wayfinding through interactive maps accessible via QR codes. By integrating technologies such as Ultra-Wideband (UWB), Wi-Fi Round-Trip Time (RTT), and Bluetooth Low Energy (BLE), Navigine achieves high positioning accuracy even in complex or crowded environments. Facility managers are empowered to maintain and update digital maps, ensuring that navigation routes and information remain updated and contextually relevant.

Timelooper Xplore is another commercial platform providing immersive and historically based experiences by combining AR and Virtual Reality (VR). The application enables users to "travel back in time" overlaying 3D reconstructions and interactive audiovisual narratives on real-world locations. This approach is particularly effective for educational tourism and museum experiences [14].

[1] see product's page https://navigine.com/ar-indoor-navigation/,

In the academic domain, the Augmented Reality-Based Indoor Navigation system (ARBIN) utilizes low-energy Bluetooth beacons (Lbeacon) and Google ARCore for 3D model placement and localization. Tested in a hospital environment, ARBIN demonstrates reliable wayfinding using direction arrows and short path computation via Dijkstra's algorithm, achieving an average accuracy of 3–5 meters [15]. Another indoor navigation solution, implemented at the Xavier Institute of Engineering, combines LiDAR mapping and strategically placed QR codes to provide high-resolution campus maps and real-time step-by-step directions [16].

The integration of virtual assistants and LLMs with AR is a rapidly growing area. Solutions such as AR Travel Concierges by Designium offer AI-powered tour guides[2] employing models like GPT-4 for user queries and site descriptions, together with advanced visual positioning systems and AR navigation overlays. Google Geospatial APIs are often used to enable seamless navigation and contextual data [17]. VirtuWander enhances virtual museum tours by employing LLMs to deliver personalized guidance via voice narration, avatars, contextual text, and additional visual aids, which increases user engagement and spatial awareness [18]. Similarly, the Live-Guided Tour system enables real-time, multi-user visits of photogrammetry-based digital twins, with features such as stereoscopic rendering, cross-platform access, and integrated e-learning functionalities for remote and inclusive access to cultural sites [19].

Despite the promise of these approaches, most existing solutions remain limited to specific institutional or commercial settings—such as individual museums, heritage sites, or niche markets like hospitality and retail. This often results in limited scalability, reusability, and cross-site integration. Moreover, while these systems enhance navigation and immersion, they rarely address the broader challenge of providing reliable and structured access to diverse cultural heritage data, particularly with regard to information consistency, provenance, and long-term management. Furthermore, many systems are highly dependent on automated content generation, which can compromise the accuracy and contextual relevance of points of interest.

The system proposed in this work aims to overcome these limitations by offering a modular and extensible platform designed to accommodate an expanding catalog of locations and digital twins. Through close integration with the Matterport ecosystem, which provides standardized, high-fidelity 3D scans, the platform enables rapid onboarding of new environments while minimizing resource requirements. Unlike previous systems that depend primarily on automated AI content, all points of interest in the present application are manually curated using Matterport's authoring tools. AI, more specifically the Gemini 2.0 Flash large language model, is used exclusively to provide in-depth insights about human-curated points of interest, thereby ensuring both precision and contextual appropriateness. However, the proposed approach can be interpreted not only as an AR navigation system, but also as a structured framework for conversational access to cultural heritage information, balancing automation with editorial control and supporting integration across multiple sites and digital collections. By combining human oversight with AI-driven elaboration, the system provides a reliable foundation for cross-site deployment.

[2] for more information see https://www.designium.jp/work/ar-travel-concierge,

Methodology

This section presents the methodology adopted in the design, implementation, and evaluation of the ARtour application. It introduces the main technologies employed, describes the system architecture and operational workflow, and details the user study conducted to compare LLM-based assistance with traditional web search within an AR cultural heritage setting.

Technologies used

The development of the application involved a combination of established platforms and innovative solutions to ensure cross-platform compatibility, flexibility, and high-quality immersive experiences.

- **Unity.** Unity[3] was selected as the main development engine. Unity is a widely adopted multi-platform engine, supporting a broad range of environments such as mobile devices, desktop systems, consoles, and virtual/augmented reality devices. The development was carried out using *C#* scripting, fully integrated with *Visual Studio* and *JetBrains Rider*. The structure of the project in Unity was organized using *Scenes*, with assets managed through the *Project Window*, and reusable elements were created as *Prefabs*. Assets integration was streamlined using the Unity Asset Store, which supplies graphical, audio, and scripting resources [20]. Cross-platform deployment was enabled through Unity's built-in support for multiple targets (mobile, desktop, VR/AR, web, consoles), leveraging the Mono Framework for runtime code compatibility [21]. The build process included optimizations for different operating systems and device configurations.
- **Large Language Model.** For natural language processing and conversational features, the application integrates the Gemini 2.0 Flash large language model. Gemini 2.0 Flash is a state-of-the-art transformer-based architecture designed for high-speed generative tasks, question answering, and contextual dialogue management. The model leverages multi-head attention and parallel processing, enabling advanced capabilities in language understanding and text generation [22, 23]. The integration of Gemini 2.0 Flash allows real-time user interaction and dynamic content generation.
- **Matterport.** For the acquisition and digitization of physical spaces, the Matterport technology was used. Matterport enables 3D spatial scans and the creation of immersive digital twins, providing accurate, interactive virtual tours and high-resolution images to be used within the application [24, 25]. Key features such as Mattertags were used to enhance navigation and contextual content delivery.
- **Digital Twin.** The Digital Twin technology was used to represent, synchronize, and simulate real-world environments within the application. Virtual replicas were used for scenario simulation, remote visualization, and data-driven management. The adoption of digital twins follows best practices and current trends in digitizing cultural heritage, as well as healthcare and industrial processes [26–28]. All components were integrated to provide a seamless user experience, using Unity for real-time rendering, Gemini 2.0 Flash for intelligent interaction, Matterport

[3] https://unity.com/

for digital content acquisition, and Digital Twin principles for dynamic scenario management.

Application design

The application was conceived as a modular, extensible platform for AR navigation and digital twin interaction, with a strong focus on maintainability, performance, and user accessibility. The design follows best practices from the Unity ecosystem, introducing architectural patterns and technological choices tailored to hybrid indoor and outdoor AR tourism scenarios. As illustrated in Fig. 1, the user journey is the following: location selection, localization via QR code, virtual guidance to the Point Of Iterest (POI), and access curated or AI-generated content, all in a seamless and accessible interface.

The system adopts a manager-based architecture. Each key functional domain (navigation, UI, input, audio, AI chat, etc.) is managed by a dedicated "Manager" object that follows the Singleton pattern, ensuring a single and persistent instance accessible from any scene. All managers are referenced by a central *MasterManager*, which acts as a unified service locator, simplifying dependencies and facilitating loose coupling among subsystems. The application is structured as a collection of Unity scenes, each corresponding either to the main menu or to a specific digital twin location. Scene transitions are handled so that critical game state and manager objects persist, enabling seamless navigation and efficient resource management. The application relies on local databases to manage information about available locations, POI, and Starting Points (SPs). Three main data repositories are used. The *Location Database* which contains metadata on all locations (name, description, type—indoor/outdoor), the *POI Database* which is populated by parsing CSV files exported from Matterport, holding details (ID, name, description, coordinates) for every POI and finally, the *SP Database* which is loaded from text files, storing information about the unique code, label, and position of each SP.

The 3D models of locations are initially provided in.obj format by Matterport and are converted to.fbx via Blender for optimal Unity integration. This ensures that geometry, hierarchy, and textures are preserved and compatible with Unity's rendering and navigation systems.

The system is designed to explicitly separate spatial representation, metadata management, and interaction logic, following a data-layer abstraction model. Cultural heritage content is ingested through structured CSV exports from Matterport, which are parsed and normalized into internal repositories. This approach enables consistent schema alignment across sites and supports interoperability between different digitized cultural environments. Provided that compatible data formats and metadata standards are adopted, the same architectural framework can incorporate heterogeneous cultural heritage repositories through a scalable ingestion pipeline, enabling unified access across multiple locations within a shared data structure.

Localization within the digital twin

The application starts with the accurate location of the user within the digital twin. To achieve this, each physical environment features one or more *Starting Points*, which are marked by physical QR codes placed at a known, fixed positions in the real world. Each QR code refers to a unique identifier, a human-readable label, and the 3D coordinates corresponding to its placement within the digital model. As illustrated in Fig. 2, when

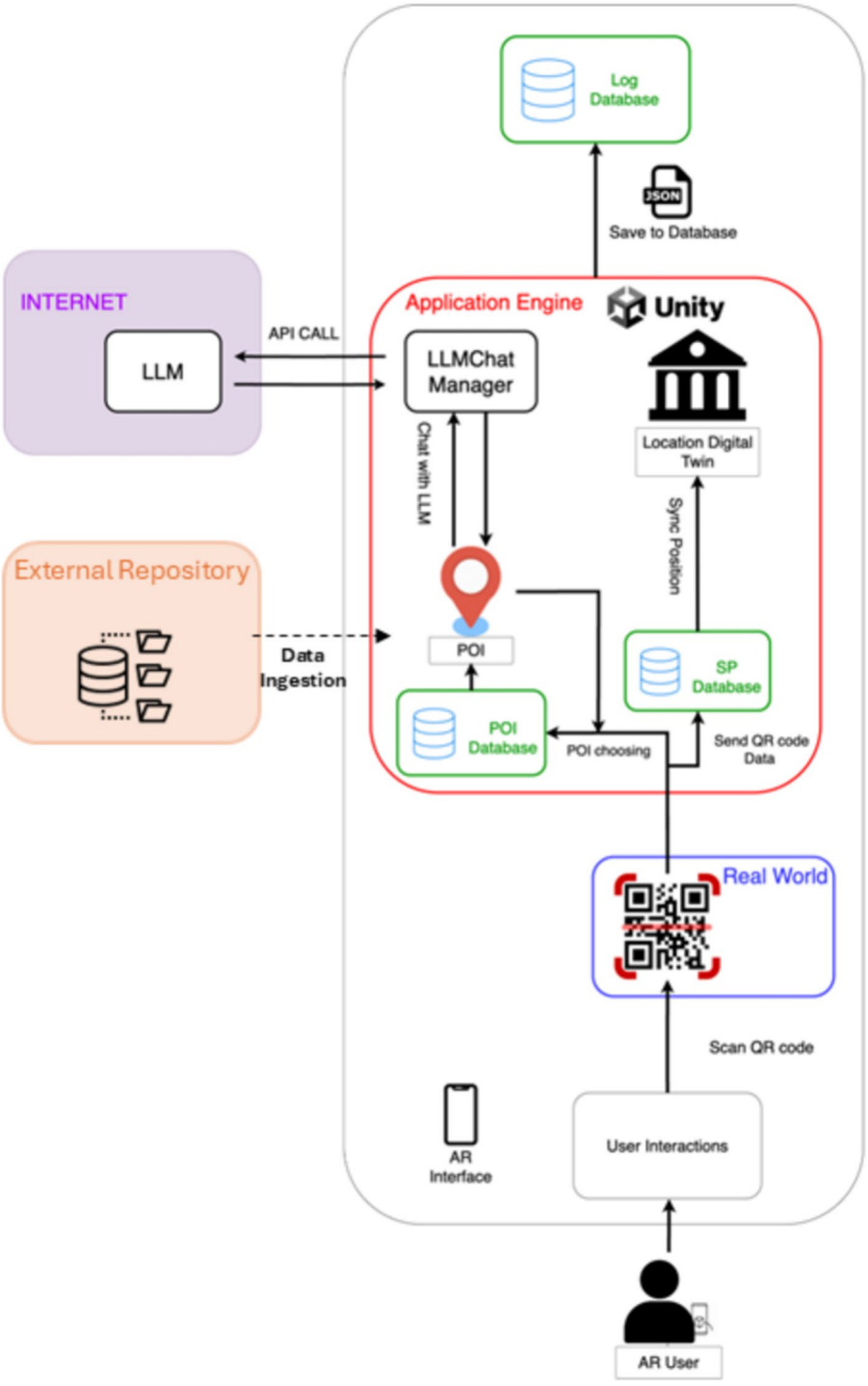

Fig. 1 Overview of the system architecture, illustrating the interaction between the AR application engine, the digital twin environment, the LLM service, and the structured data layer, including POI and SP repositories. The scheme also highlights the data ingestion process from external cultural heritage repositories and the user workflow from QR-code localization to POI-based information retrieval

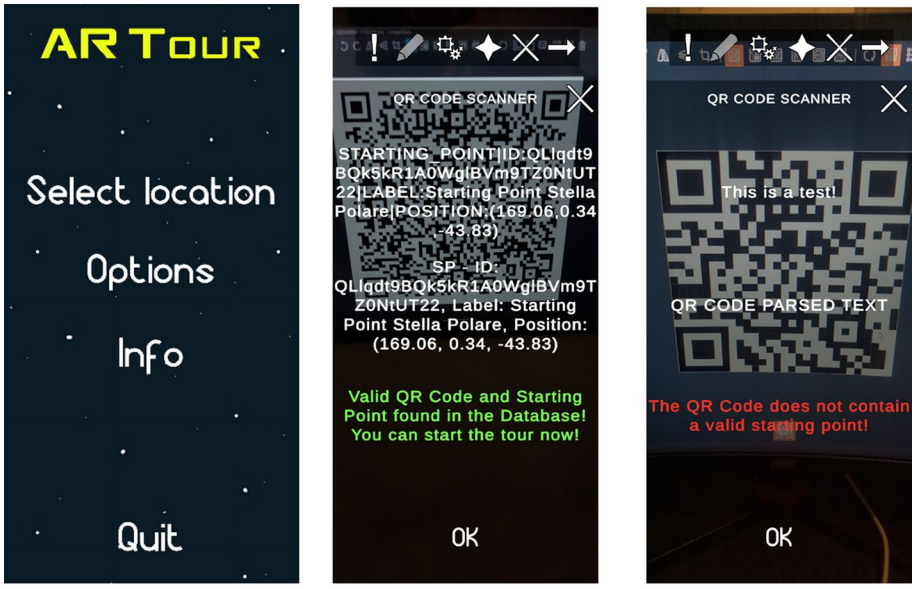

Fig. 2 ARtour interface illustrating the localization workflow: main menu and location selection (left), successful parsing and validation of a QR code associated with a Starting Point (center), and error notification when the scanned QR code does not correspond to a valid Starting Point (right)

the user arrives at the location and scans the QR code using the application, the QR code data are first parsed and validated in order to extract the associated spatial information. Based on this data, the user's virtual avatar is positioned at the corresponding coordinates within the digital twin, thereby synchronizing the physical and virtual locations. Finally, the application loads and presents the set of POIs available for the selected environment.

This workflow guarantees a one-to-one correspondence between the real and virtual worlds, enabling precise AR navigation and content delivery from the very beginning of the user experience.

Navigating to the point of interest

Once the user is anchored in the digital twin via the Starting Point, the navigation workflow begins. The user selects a POI through the interface so that visual aids, such as arrows and real-time distance indicators, assist the user.

Navigation is powered by Unity's AI Navigation package, which generates and dynamically updates NavMeshes based on the current environment. Separate agent configurations are used for indoor and outdoor scenarios, so that a smaller radius is used for narrow spaces and a larger radius for open areas. An automated check ensures that each POI is reachable; otherwise, a warning is displayed. The identification of POIs in the AR environment is based on raycasting from the camera of the user, which activates the interaction interface when a POI comes into view, without the need for any physical markers.

Additionally, since the digital twin is a 1:1 replica of the real environment, it must remain invisible to the user but still interact with navigation and collision systems. As illustrated in Fig. 3, custom shader renders the digital twin invisible while maintaining

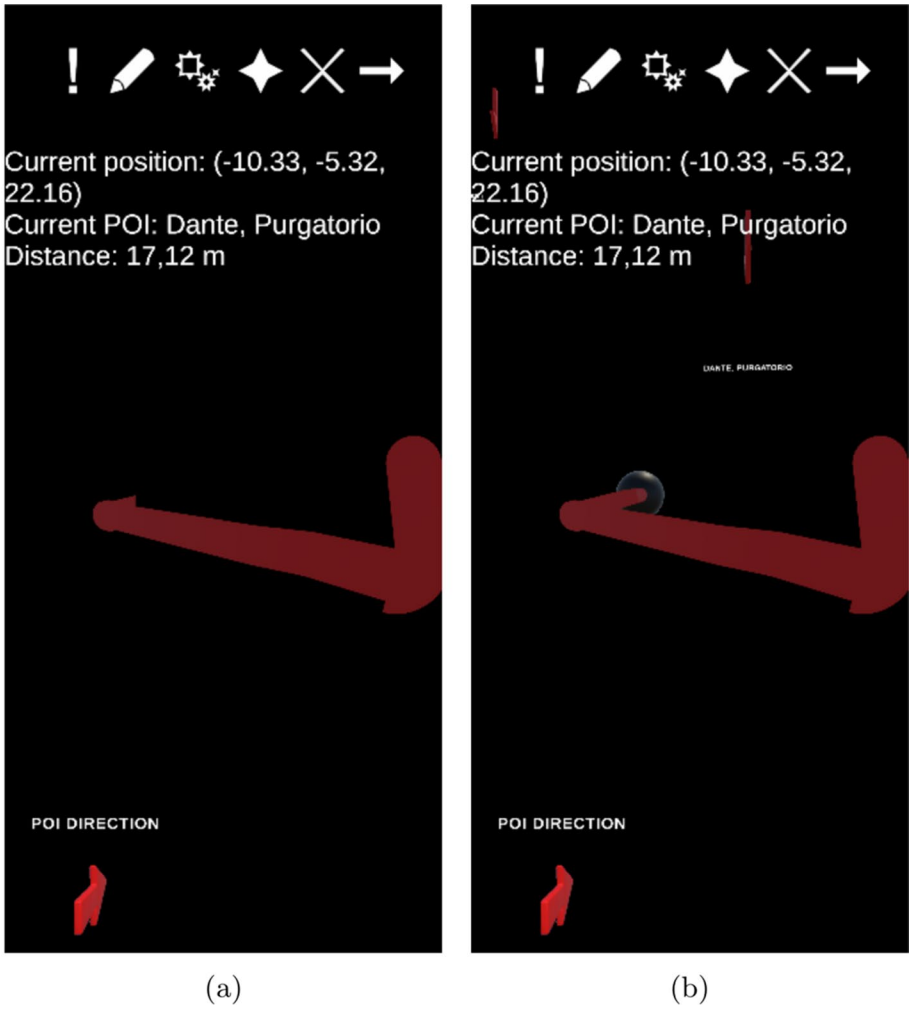

Fig. 3 Comparison of AR scene rendering showing (**a**) correct occlusion achieved with the custom invisibility shader, and (**b**) incorrect visibility of navigation agents through walls when the shader is not applied

occlusion properties for navigation agents and virtual objects. This ensures that the visual cues are hidden when behind walls, just as they would be in the real world.

Interaction with the point of interest

The application displays a system notification to inform the user if a selected POI is unreachable due to navigation constraints (see Fig. 4a). When the user reaches a POI, the application presents a multi-option dialogue, as illustrated in Fig. 4b:

1. Access a curated static description of the POI extracted from the Matterport data.
2. Perform a web-based search to retrieve additional information about the POI.
3. Interact with the integrated AI assistant to obtain further details, contextual explanations, or anecdotes related to the POI.

The app integrates Gemini 2.0 Flash, a state-of-the-art LLM, to function as a virtual tour guide. As can be seen in Fig. 4c, users can interact with the assistant through a chat interface, receiving context-aware responses tailored to the current POI and location.

To retrieve the most accurate information possible and mitigate potential hallucinations from the LLM, the initial prompt was predetermined and carefully constructed as

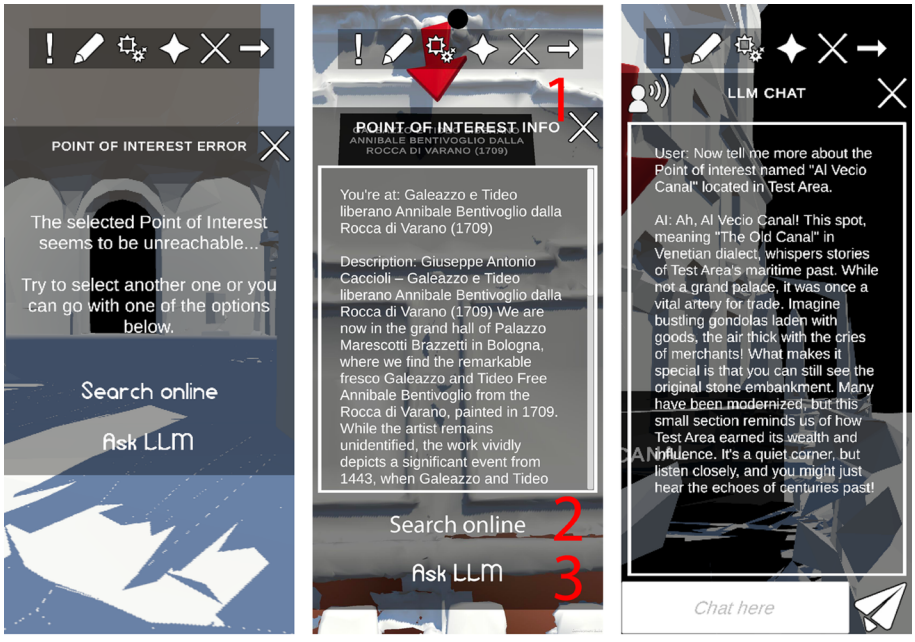

Fig. 4 Screens from the ARtour application illustrating the interaction with a Point of Interest: (**a**) system notification displayed when a selected POI is unreachable due to navigation constraints, (**b**) POI information dialogue providing access to curated static content, web search, and the AI assistant, and (**c**) in-app chat interface with the Gemini 2.0 Flash AI assistant

a question containing precise details about each point of interest. These details included the name of the point of interest, the year of creation, and its specific location, specifying both the building and the city. The initial prompt allows a proper contextualization for the chat bot. Since the present prototype focuses on few points of interest, the primary outputs generated by these initial prompts were double-checked using external sources and verified by an expert. The information provided was generally correct but other prompting techniques or fine-tuning was used. The fixed prompt is used at the beginning to ensure comparable baseline outputs across participants, after which users are allowed to formulate their own prompts freely.

To ensure security, the Application Programming Interface (API) key required for LLM access is encrypted within the app using Advanced Encryption Standard (AES) and managed through a custom Unity editor extension, preventing leakage or unauthorized access. Furthermore, accessibility is prioritized through a built-in Text-to-Speech (TTS) engine: every AI-generated or static POI response can be vocalized, enabling hands-free use and support for users with visual impairments. The TTS system takes advantage of the native Android Java backend, ensuring broad compatibility with the devices and seamless integration. It is invoked from C# via Unity's AndroidJavaObject interface, allowing efficient cross-language communication with minimal overhead.

User testing

The following subsections describe in detail the participants and experimental setting, the task workflow, and the data collection procedure adopted in the user study.

Partecipants and experimental setting

A total of 42 participants were recruited for this study. Women constituted 68.2% of the sample, whereas men accounted for 31.8%. Participants' ages ranged from 20 to 45 years, with 45.2% of the sample aged between 20 and 25 years and the remaining 54.8% aged between 26 and 45 years.

In terms of educational background, the majority of participants held either a Bachelor's degree (45.5%) or a Master's degree (45.5%). Only a small fraction of the sample reported a high school diploma (4.5%) or a doctoral degree (4.5%).

Concerning familiarity with AR/VR technologies, participants generally reported limited to moderate experience. More precisely, 59.5% reported rarely or never using such systems, while only 14.2% indicated that they used them often or regularly.

The experiment was carried out in one of the main lecture room of the Department of the Arts at the University of Bologna called Salone Marescotti. In order to calibrate the application and synchronize the user's real-world position with the digital twin environment, a QR code was printed and placed on a wall outside the Salone Marescotti. The two following frescoes were selected as POIs:

- *Esaltazione della casa Marescotti, Giuseppe and Antonio Rolli, 1686–87.*
- *Galeazzo e Tideo liberano Annibale Bentivoglio dalla Rocca di Varano, Giuseppe Antonio Caccioli, 1709.*

We emphasize that while the empirical evaluation focuses on a single cultural site and two selected POIs, this deployment represents a controlled instance of the broader architectural framework described in the previous sections.

Participants were randomly assigned to one of the two following experimental groups:

- **LLM group**: Users retrieved information about the selected artwork by interacting with the Gemini 2.0 Flash large language model through an in-app chat interface.
- **Web Search group**: Users retrieved information about the selected artwork by using a dedicated in-app button that opened the device's default web browser.

All participants were unpaid volunteers and completed the user test sequentially in a closed room in the Salone Marescotti, ensuring that no one could observe another participant's behavior, with the presence of a facilitator. A smartphone preconfigured with the application was provided for the tests.

Task workflow

As illustrated in Fig. 5, the experimental workflow was conducted as follows:

1. **Briefing:** Each participant was first introduced to the AR application and instructed on its purpose and basic operation.
2. **Calibration:** Participants were asked to scan the QR code placed at the designated Starting Point in front of the Salone Marescotti, initializing the app and aligning their virtual position within the digital twin of the room.
3. **POI Selection:** Participants had to select on the app one of the two available POIs within the environment.
4. **Navigation:** Using the AR interface, participants followed on-screen visual cues (such as directional arrows) to navigate to the selected POI. The application automatically recorded the time taken to reach the target.

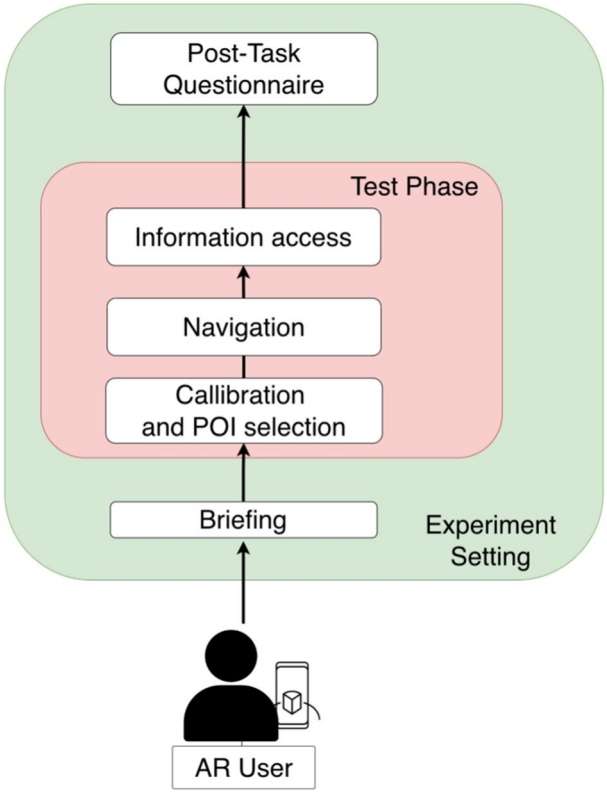

Fig. 5 Overview of the experimental workflow for the AR user

5. **Information Access:**

   - **LLM group:** Upon arrival at the POI, participants could interact with Gemini 2.0 Flash via the in-app (Fig. 4c) chat to ask questions and receive information about the artwork. Conversation logs were saved by the application.
   - **Web Search group:** Upon arrival, participants could tap a button (Fig. 4b) to launch a web search for the name of the artwork using the default browser of the device.

6. **Completion:** After information access, the experimental session concluded for that participant. All relevant interaction metrics (navigation time, chat log, etc.) were stored by the application.
7. **Post-Task Questionnaire:** Immediately after the session, participants were asked to complete a set of standardized questionnaires to assess various aspects of their user experience.

Data collection

To assess participants' user experience, we administered a set of standardized questionnaires. Affective responses were measured using the Positive and Negative Affect Schedule (PANAS) [29], separately analyzing positive and negative affect. Perceived usability was assessed through the System Usability Scale (SUS) [30] using a 5-point Likert scale, with items grouped into positive and negative components. Technology acceptance was evaluated using the Technology Acceptance Model (TAM) [31], also employing a 5-point Likert scale and comprising the perceived usefulness and perceived ease of

Table 1 Summary of subjective measures (Mean, Standard Deviation) comparing LLM and Web Search groups. The table reports SUS, TAM, NASA-TLX, PANAS, and Social Presence scores

Variable	Construct	LLM(M, SD)	Web Search(M, SD)
SUS	SUS Positive	3.89, 0.52	3.80, 0.84
	SUS Negative	2.04, 0.65	2.07, 0.82
TAM	Perceived Usefulness	4.38, 0.70	4.06, 0.72
	Perceived Ease of Use	4.33, 0.83	4.33, 0.64
NASA-TLX	Mental Demand	3.33, 1.17	2.93, 1.48
	Physical Demand	3.79, 2.05	2.88, 1.60
	Temporal Demand	4.05, 2.44	4.67, 2.69
	Performance	7.67, 1.88	7.71, 2.17
	Effort	3.67, 2.20	3.38, 2.14
	Frustration	2.62, 1.96	2.38, 1.69
PANAS	Positive	2.95, 0.83	3.13, 1.03
	Negative	1.46, 0.70	1.41, 0.65
Social Presence	Co-presence	3.79, 0.96	–
	Attentional Allocation	3.64, 0.71	–
	Perceived Message Understanding	2.98, 0.30	–

use constructs. Perceived workload was measured using the NASA Task Load Index (NASA-TLX), considering the mental demand, physical demand, temporal demand, performance, effort, and frustration subscales. Social presence was assessed using the Networked Minds Social Presence Questionnaire [32], particularly we considered three components: perceived co-presence, attentional allocation, and perceived message understanding.

In addition to questionnaire responses, the following objective measures were collected for each participant:

- Time required to reach the selected POI.
- The number of interactions with the "Search online" and "Ask LLM" buttons.
- Chat interaction logs and system response times during interaction with Gemini 2.0 Flash (only for the LLM group).

This data collection protocol enabled a comparative evaluation of information retrieval performance and user experience between LLM-based assistance and conventional web search within an AR-powered cultural heritage context.

Results

This section reports the quantitative results of the comparison between the LLM and Web Search conditions. For each measure, normality was first assessed using the Shapiro–Wilk test. Subsequently, independent-samples t-tests or Mann–Whitney U tests were applied to assess statistical differences. Effect sizes are reported as Cohen's d, and statistical equivalence was evaluated using the Two One-Sided Tests (TOST) procedure.

Subjective measures

This section reports the results related to the subjective measures. The results for all subjective measures are summarized in Table 1.

SUS

For both positive and negative SUS items, the assumption of normality was not violated in either the LLM-based condition or the Web Search condition ($p > .05$), allowing for the use of parametric tests.

Regarding positive SUS items, no statistically significant difference was observed between the LLM-based system and the Web Search condition, as indicated by a Welch t-test ($p = .69$). However, equivalence testing using the TOST procedure revealed that the two conditions were statistically equivalent within the predefined equivalence bounds ($p = .03$).

A comparable pattern emerged for negative SUS items. No statistically significant differences were found between the LLM-based condition and the Web Search condition ($p = .90$). The TOST procedure further confirmed equivalence between conditions ($p = .01$), indicating similar perceived usability-related difficulties across both interaction modalities.

TAM

Prior to inferential analysis, the assumption of normality was assessed using the Shapiro–Wilk test. For both perceived usefulness (TAM PU) and perceived ease of use (TAM PE), normality was violated in both the LLM-based condition and the Web Search condition ($p < .05$).

Regarding TAM PU, no statistically significant difference was observed between the LLM-based condition and the Web Search condition, as indicated by a Mann–Whitney U test ($p = .11$). Equivalence testing using the TOST procedure did not support equivalence between conditions ($p = .21$).

For TAM PE, no significant difference emerged between the LLM-based condition and the Web Search condition, as indicated by a Mann–Whitney U test ($p = .70$). In contrast to TAM PU, equivalence testing revealed that the two conditions were statistically equivalent within the predefined bounds ($p = .02$), indicating comparable perceived ease of use across interaction modalities.

NASA-TLX

The assumption of normality was not consistently met across NASA-TLX subscales and conditions ($p < .05$ in several cases); therefore, non-parametric tests were employed when appropriate.

For the mental demand subscale, no statistically significant difference was observed between the LLM-based condition and the Web Search condition, as indicated by a Mann–Whitney U test ($p = .18$). However, the magnitude of the difference was small-to-moderate (Cohen's $d = 0.30$). The TOST procedure did not support equivalence between conditions ($p = .41$).

Similarly, no statistically significant difference emerged for physical demand between the LLM-based system and the Web Search condition, as indicated by a Welch t-test ($p = .12$). The effect size was moderate (Cohen's $d = 0.49$). The TOST procedure did not confirm equivalence within the predefined bounds ($p = .76$).

For temporal demand, performance, effort, and frustration subscales, no statistically significant differences were found between the two conditions (all $p > .05$). Effect size estimates suggested small differences for temporal demand (Cohen's $d = -0.24$),

performance (Cohen's $d = -0.02$), effort (Cohen's $d = 0.13$), and frustration (Cohen's $d = 0.13$). In addition, equivalence testing did not support equivalence for any of these dimensions (all TOST $p > .05$).

PANAS

For positive affect scores, normality was satisfied in both the LLM-based condition and the Web Search condition ($p > .05$), whereas for negative affect scores, normality was violated in both conditions ($p < .01$). Accordingly, parametric and non-parametric tests were applied as appropriate.

Regarding positive affect, no statistically significant difference was observed between the LLM-based condition and the Web Search condition, as indicated by a Welch t-test ($p = .64$). Equivalence testing using the TOST procedure supported equivalence between conditions ($p = .04$).

For negative affect, no statistically significant difference emerged between the LLM-based condition and the Web Search condition, as indicated by a Mann–Whitney U test ($p = .51$). Equivalence testing reached statistical significance within the predefined bounds ($p = .04$).

Social presence

The Social Presence questionnaire, which was exclusively answered by the LLM group, assessed participants' perceived awareness of the system, attentional engagement, and message understanding during interaction with the LLM.

Within the LLM-based condition, participants reported a moderate to high sense of presence of the LLM (M = 3.79, SD = 0.96). Similarly, attention-related scores suggested sustained engagement with the LLM (M = 3.64, SD = 0.71). Regarding message understanding, participants reported moderately high scores (M = 2.98, SD = 0.30), suggesting that the interaction with the LLM was generally perceived as understandable.

Objective measures

The time required to reach the selected POI was slightly higher in the LLM condition (M = 41.46 s, SD = 28.33) than in the Web Search condition (M = 34.32 s, SD = 27.02). The two distributions deviated significantly from normality ($p < .001$). Accordingly, a Mann–Whitney U test indicated no statistically significant difference between the two modalities (p =.245). The TOST test did not support the equivalence, and the observed effect size was small (Cohen's d = 0.26).

In contrast, a distinct pattern emerged from the analysis of interaction behavior. In the Web Search condition, only 5 out of 21 participants interacted with the Web Search button, whereas 18 out of 21 participants engaged with the LLM interface. Furthermore, we highlight that the average response latency for LLM interactions was approximately 8 seconds for each question.

Qualitative inspection of the LLM interactions further highlighted an incremental and dialogic use of the model. Participants typically initiated the interaction with descriptive requests (e.g., "Now tell me more about the point of interest..."), which were followed by contextual clarification questions (e.g., "Who is the red one?", "Why was he imprisoned?", "Who painted this?"). Over time, interactions often developed into more exploratory

and spontaneous inquiries (e.g., "How many people are in this fresco?", "Why are they monstrous?"), reflecting a progressive deepening of engagement with the content.

Discussion, limitations and future works

In this section, we provide answers to the research questions, highlight the main limitations of the study, and outline potential directions for future development.

RQ1: How does integrating an LLM into AR-based cultural heritage applications influence users' perceived usability, cognitive workload, and emotional experiences compared to traditional web search methods when accessing digital cultural heritage information?

In terms of usability and acceptance, no significant differences emerged between the LLM and Web Search conditions. SUS scores were high and statistically equivalent across modalities, as well as TAM PE, indicating that conversational interaction does not compromise usability. TAM PU showed no significant differences but did not reach statistical equivalence, however, mean scores were higher for the LLM-based modality, suggesting a tendency for users to perceive the conversational approach as more useful. To further contextualize the high levels of usability and acceptance, we also consider the objective performance metrics. The implemented AR navigation system enabled users to reach the selected point of interest efficiently in both conditions, with comparable navigation times across modalities.

Regarding cognitive workload, NASA-TLX results indicate that, overall, workload scores tended to be higher in the LLM condition. This pattern can be explained by the increased level of dialogue and interaction stimulated by the LLM, as also reflected in the objective interaction metrics. Nevertheless, no statistically significant differences emerged between conditions across any of the NASA-TLX subscales. These findings indicate that the use of an LLM does not lead to a significant increase in perceived workload.

Finally, emotional experience was comparable across conditions. Positive and negative affect did not differ significantly and were statistically equivalent, suggesting that LLM-based interaction does not introduce additional emotional strain or discomfort compared to web search.

Beyond usability considerations, these findings suggest that conversational interfaces can support intuitive access to digitally available cultural heritage information without increasing cognitive burden. In contexts where users are required to navigate structured metadata, descriptive archives, or multimodal digital representations, reducing interaction complexity becomes particularly relevant. The comparable workload levels observed in the LLM condition indicate that conversational mediation may offer a scalable interaction paradigm for accessing structured cultural data without compromising user experience.

RQ2: To what extent does interacting with an LLM-driven AR assistant enhance users' sense of co-presence during the exploration of digital cultural heritage content?

Social presence measures indicate that participants perceived the LLM as an attentive and responsive conversational partner, reporting a clear sense of co-presence, sustained attentional engagement, and generally good message understanding, which supported smooth and coherent communication within the AR experience.

These findings are further supported by objective interaction metrics, which reveal a higher level of engagement with the LLM compared to the web search. Participants in

the LLM condition interacted frequently with the conversational interface, whereas only a small number of participants in the Web Search condition sought additional information about the points of interest. This disparity suggests that the LLM encouraged more active and continuous information-seeking behavior.

However, we remark that the comparatively lower ratings observed for message understanding may, at least in part, be attributable to the average response latency of the LLM, which required users to wait several seconds (8 s) before receiving a reply and may have slightly disrupted the perceived fluency of the interaction.

From a broader perspective, we hypothesize that the enhanced engagement observed in the LLM condition may indicate the potential of conversational systems to support exploratory sense-making processes. In data-rich cultural environments, users often need to progressively refine their informational needs rather than retrieve isolated facts. The dialogic nature of LLM interaction may facilitate iterative knowledge construction, particularly when navigating complex or heterogeneous cultural datasets.

RQ3: Does the use of an LLM within an AR application improve information retrieval efficiency and consistency compared to conventional methods in the context of accessing cultural heritage data?

A markedly different pattern emerged when examining interaction behavior. Participants in the LLM condition engaged with the conversational interface far more frequently than participants in the Web Search condition used the search button, indicating a strong preference for and reliance on the LLM as an information access tool. These interactions were often repeated throughout task execution, suggesting a more continuous and iterative information-seeking process.

Qualitative analysis further revealed that LLM interactions followed an incremental and dialogic trajectory. Users typically began with broad, descriptive requests and progressively moved toward more specific clarification questions and exploratory inquiries. This interaction style reflects a deepening engagement with the cultural content, enabled by the conversational nature of the LLM, which supports ongoing refinement of information needs rather than discrete, one-off queries.

However, the analysis of LLM outputs also revealed limitations in terms of information consistency. While a systematic content evaluation was beyond the scope of this study, qualitative inspection of the generated responses highlighted occasional inconsistencies in the information provided by the LLM. This issue is likely related to the fact that the deployed model was not fine-tuned on domain-specific data related to the cultural environment under investigation. As a result, the potential presence of inaccuracies or biases remains a concern, particularly in educational and cultural heritage contexts. Addressing information quality therefore represents an important direction for future work. Fine-tuning LLMs on curated, site-specific cultural heritage datasets, integrating retrieval-augmented generation techniques, or combining AI-generated responses with expert-authored content could help improve accuracy and consistency while preserving conversational flexibility.

These considerations become particularly relevant in contexts where multiple digitized cultural sites are integrated within the same architectural framework. The current implementation already supports the ingestion of additional digital twin models and associated POIs through compatible data formats and metadata schemas. However, the empirical evaluation presented in this study was conducted on a limited number of

curated POIs. As the number of integrated sites and data sources increases, maintaining metadata consistency, provenance transparency, and information reliability becomes increasingly important.

Therefore, RQ3 does not only concern retrieval efficiency, but also highlights the broader implications of conversational systems as interfaces to digitally available cultural heritage data in data-intensive environments.

Additional limitations should be acknowledged. The relatively small number of participants may have reduced statistical power and limited the detection of more subtle effects. Future studies involving larger and more diverse samples are needed to strengthen the robustness and generalizability of the findings.

Finally, the present implementation relied on a text-based conversational interface rather than an embodied LLM. Future work will investigate how avatar-based or multimodal LLM assistants, incorporating voice, gestures, or visual embodiment, affect usability, cognitive workload, emotional experience, and co-presence.

Conclusions

In this paper, we introduced ARtour, an augmented reality application for cultural heritage tours driven by an LLM-based conversational agent. After describing the system architecture and implementation, we evaluated the impact of LLM integration on the AR experience in terms of usability, sense of presence, and information retrieval.

The proposed architecture supports a consistent and high-quality user experience while ensuring extensibility: new locations, digital twins, and content modules can be integrated with minimal changes to the code base. In this sense, the system can also be interpreted as a data-driven framework for conversational access to digitally available cultural heritage information.

The analysis indicates that integrating an LLM within an AR cultural heritage application does not compromise usability, emotional comfort, or task performance. Usability and acceptance were high and comparable to traditional web search, while the conversational interface encouraged more interactive and exploratory information-seeking behavior. The results further showed that participants asked the LLM numerous questions during task execution, indicating a strong preference for and reliance on the conversational assistant as an information access tool. Interaction with the LLM also fostered a moderate sense of co-presence, contributing to immersion during cultural exploration. These findings suggest that conversational agents may facilitate exploratory sense-making in digitally mediated and data-rich cultural environments.

At the same time, qualitative inspection of the generated outputs highlighted occasional issues related to the consistency and veracity of the information provided. Although content accuracy was not systematically evaluated in this work, these observations underline the importance of improving information reliability in future deployments through domain-specific fine-tuning, retrieval-augmented approaches, and stronger editorial control, particularly in educational and cultural heritage contexts. As digital cultural repositories grow in scale and heterogeneity, ensuring data quality and governance will remain central to the responsible deployment of LLM-mediated systems.

Acknowledgements

This study was carried out within the MICS (Made in Italy – Circular and Sustainable) Extended Partnership and received funding from the European Union Next-GenerationEU (Piano Nazionale di ripresa e resilienza (PNRR) – Missione 4

Componente 2, Investimento 1.3 - D.D. 1551.11-10-2022, PE00000004) and the project Virtual Worlds Innovation Masters: Shaping Future Digital Skills Europe (UPRAISE), Grant Agreement No. 10122592, funded under the DIGITAL-2024-ADVANCED-DIGITAL-07-KEYCAPACITY programme.

Author contributions
A.T. and A.L. and V.B. and P.C. wrote the main mansucript text. A.L prepared figures 1-16. A.T. and A.B. and L.V. designed and coded the main application. A.L. coded the integration of the A.I agent. A.L. and V.B. and M.L. prepared user testings. P.C. and G.M. suppervised the project. All authors reviewed the manuscript.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
This research was approved by the University of Bologna Ethics Committee (Approval number 0159749) on June 23, 2022. The study adhered to all ethical guidelines, and informed consent was obtained from all participants prior to their involvement.

Competing interests
The authors declare no competing interests.

Received: 29 July 2025 / Accepted: 15 March 2026
Published online: 04 April 2026

References

1. Navarrete T. Digital cultural heritage. In: Handbook on the economics of cultural heritage. Edward Elgar Publishing; 2013. p. 251–71.
2. Ribeiro M, Santos J, Lobo JA, Araújo S, Magalhães L, Adão T. 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. Web3D '24. Association for Computing Machinery, New York, NY, USA 2024. https://doi.org/10.1145/3665318.3677172.
3. Hajahmadi S, Calvi I, Stacchiotti E, Cascarano P, Marfia G. Heritage elements and artificial intelligence as storytelling tools for virtual retail environments. Digit Appl Archaeol Cult Herit. 2024;34:00368.
4. Dordio A, Lancho E, Merchán MJ, Merchán P. Cultural heritage as a didactic resource through extended reality: a systematic review of the literature. Multimodal Technol Interact. 2024;8(7):58. https://doi.org/10.3390/mti8070058.
5. Tsepapadakis M, Gavalas D. Are you talking to me? An audio augmented reality conversational guide for cultural heritage. Pervasive Mob Comput. 2023;92:101797. https://doi.org/10.1016/j.pmcj.2023.101797.
6. Bu F, Wang Z, Wang S, Liu Z. An Investigation into Value Misalignment in LLM-Generated Texts for Cultural Heritage 2025. https://arxiv.org/abs/2501.02039.
7. Toth GM, Albrecht R, Pruski C. Explainable AI, LLM, and digitized archival cultural heritage: a case study of the grand ducal archive of the medici. AI Soc. 2025. https://doi.org/10.1007/s00146-025-02238-5.
8. Cossatin AG, Mauro N, Ferrero F, Ardissono L. Tell me more: integrating llms in a cultural heritage website for advanced information exploration support. Inf Technol Tourism. 2025. https://doi.org/10.1007/s40558-025-00312-8.
9. Vallasciani G, Stacchio L, Cascarano P, Marfia G. Creaixr: fostering creativity with generative ai in xr environments. In: 2024 IEEE international conference on metaverse computing, networking, and applications (MetaCom), pp. 1–8 2024. IEEE.
10. Ranjgar B, Sadeghi-Niaraki A, Shakeri M, Rahimi F, Choi S-M. Cultural heritage information retrieval: past, present, and future trends. IEEE Access. 2024;12:42992–3026. https://doi.org/10.1109/ACCESS.2024.3374769.
11. Trichopoulos G. Large language models for cultural heritage. In: Proceedings of the 2nd international conference of the ACM greek SIGCHI chapter. CHIGREECE '23. Association for Computing Machinery, New York, NY, USA 2023. https://doi.org/10.1145/3609987.3610018.
12. Hutchinson D. Mapping the latent past: assessing large language models as digital tools through source criticism. J Digit Hist. 2024;3(1):20230018.
13. Sathiyabamavathy K, Anju KP. Role of chatbots in cultural heritage tourism: an empirical study on ancient forts and palaces. J Heritage Manag. 2024;9(1):9–28. https://doi.org/10.1177/24559296241253932.
14. TimeLooper: TimeLooper – Immersive Historical Experiences. https://www.timelooper.com/. Accessed: 2025-03-02.
15. Huang B-C, Hsu J, Chu E-H, Wu H-M. Arbin: augmented reality based indoor navigation system. Sensors. 2020;20(20):5890. https://doi.org/10.3390/s20205890.
16. Parab DK, Prasanna Deshpande P, Thakur RM, Atul Warke V, Khanvilkar S. Indoor navigation system using augmented reality. In: 2024 international conference on inventive computation technologies (ICICT), pp. 720–725 2024. https://doi.org/10.1109/ICICT60155.2024.10545015. https://ieeexplore.ieee.org/abstract/document/10545015
17. Inc., D. Designium XR. https://www.designium.jp/xr. Accessed: 2025-07-29
18. Wang Z, Yuan L-P, Wang L, Jiang B, Zeng W. Virtuwander: enhancing multi-modal interaction for virtual tour guidance through large language models. In: Proceedings of the CHI Conference on Human Factors in Computing Systems. CHI '24, pp. 1–20. ACM, 2024. https://doi.org/10.1145/3613904.3642235.
19. Gabellone F. Digital twin: a new perspective for cultural heritage management and fruition. Acta IMEKO. 2022;11(1):1–7. https://doi.org/10.21014/acta.

20. Unity Technologies: What is the Unity Asset Store and how do I purchase Assets? https://support.unity.com/hc/en-us/articles/210142503-What-is-the-Unity-Asset-Store-and-how-do-I-purchase-Assets Accessed 2025-03-05.
21. Vishwakarma N. How Unity Supports Cross Platform Feature 2020. https://niraj-vishwakarma.medium.com/how-unity-supports-cross-platform-feature-ae722321cfa Accessed 2025-02-19.
22. Wikipedia contributors: Porting — Wikipedia, The Free Encyclopedia 2025. https://en.wikipedia.org/wiki/Large_language_model Accessed 2025-02-21.
23. Alberts IL, Mercolli L, Pyka T, Prenosil G, Shi K, Rominger A, et al. Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? Eur J Nucl Med Mol Imaging. 2023;50(6):1549–52. https://doi.org/10.1007/s00259-023-06172-w.
24. HomeTrack: What is Matterport? 2024. https://www.hometrack.net/blog/what-is-matterport Accessed 2025-02-20.
25. Scene3D: What is Matterport? Scene3D 2024. https://scene3d.co.uk/what-is-matterport/ Accessed 2025-02-20.
26. Glaessgen EH, Stargel DS. The digital twin paradigm for future nasa and u.s. air force vehicles. In: 53rd AIAA/ASME/ASCE/AHS/ASC Structures, structural dynamics and materials conference 2012. https://doi.org/10.2514/6.2012-1818. NASA. https://ntrs.nasa.gov/citations/20120008178.
27. McKinsey & Company: what is digital twin technology? 2024. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-digital-twin-technology Accessed 2025-02-19.
28. Fortune Business Insights: Digital Twin Market Size 2024. https://www.fortunebusinessinsights.com/digital-twin-market-106246 Accessed 2024-02-19.
29. Watson D, Clark LA, Tellegen A. Development and validation of brief measures of positive and negative affect: the PANAS scales. J Pers Soc Psychol. 1988;54(6):1063.
30. Brooke J, et al. Sus-a quick and dirty usability scale. Usability Eval Ind. 1996;189(194):4–7.
31. Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Q. 1989;13(3):319–40.
32. Harms C, Biocca F. Internal consistency and reliability of the networked minds measure of social presence. In: Seventh Annual International Workshop: Presence, vol. 2004 2004. Universidad Politecnica de Valencia Valencia, Spain.

Publisher's Note