Information Technology & Tourism
https://doi.org/10.1007/s40558-025-00312-8

ORIGINAL RESEARCH

# Tell me more: integrating LLMs in a cultural heritage website for advanced information exploration support

**Angelo Geninatti Cossatin**[1] · **Noemi Mauro**[1] · **Fabio Ferrero**[1] · **Liliana Ardissono**[1]

Received: 19 June 2024 / Revised: 4 January 2025 / Accepted: 8 January 2025


## Abstract

Cultural Heritage websites' capability to satisfy diverse information needs is limited by their high-quality but constrained knowledge bases. Thus, we investigate their extension with external large language models (LLMs), enriching the provision of cultural content by leveraging LLMs' continuous collection and integration of information from heterogeneous data sources. This extension raises important challenges in synchronizing the LLM's behavior with the user's browsing activity on the website to offer a unified interaction environment. To address these challenges, we propose a loosely coupled integration model that provides users with curated content and an assisted question-answering function to answer information needs that the system's knowledge base fails to cover. Our model is agnostic to the LLM and synchronizes its behavior with the user's browsing activity through implicit prompt engineering. We tested a baseline website without LLM integration, one with free-text interaction with the LLM, and another that combines free-text interaction with the suggestion of context-dependent questions. In a user study involving 44 participants, we found that the LLM-powered website has higher usability and that context-dependent question suggestions further enhance user experience, especially for people with low curiosity levels (according to Curiosity and Exploration Inventory-II - CEI-II) who are guided in formulating effective questions. This shows the potential of LLMs to enrich engagement with existing Cultural Heritage websites.

**Keywords** Cultural heritage · Large language models · Mobile guides · Interactive search

Extended author information available on the last page of the article

Published online: 28 January 2025

## 1 Introduction

The technological advances of the past decades have changed cultural heritage (CH) exploration to address two main challenges: (i) extending the domain information presented by digital (web-based/mobile) guides and (ii) enhancing the user experience by enriching the interaction modalities. The first challenge has been partially addressed by integrating CH guides with semantic knowledge bases (e.g., Europeana 2024; Rinaldi et al. 2022), which extend the available knowledge but still represent relatively static and focused information sources. The second one has involved, among the others, employing chatbots as user interfaces to support dialog-based interaction with users. Chatbots are largely used in Cultural Heritage institutions as "info bots" that operate "as guides or as tools to help users plan their visit" (Tzouganatou 2018) but typically provide basic or logistic information about Points of Interest. Moreover, they are used in microservices architectures to integrate information from different tourist services in tour planning (Sperlí 2021). Large Language Models (LLMs, Jurafsky and Martin 2024) open new possibilities for developing Cultural Heritage guides based on chatbots thanks to their strength in understanding and answering users' questions, and their capability to collect and integrate information from heterogeneous sources. However, using a chatbot as a user interface limits interaction to message-based conversations that lack the richness of the advanced information presentation techniques developed in the adaptive web (Ardissono et al. 2012; De Carolis et al. 2018).

The previous discussion highlights the need for new models that combine the benefits of web-based exploration and natural language conversation in a CH guide while mitigating the issues affecting LLMs. However, in developing these models, further risks must be considered:

- Lay users may struggle to formulate precise or targeted questions when interacting with LLMs, whose knowledge spans uncontrolled topics. Thus, they might receive unexpected answers.
- The website and the LLM are separate systems. Therefore, navigation and conversation might misalign, failing to provide users with a unified interaction environment.

We present a loosely coupled model (McNatt and Bieman 2001) to harmonically integrate an external LLM in a web-based Cultural Heritage guide for improved information exploration. The proposed user interface includes a chatbot within the guide's web pages. The chatbot also supports the user by proposing candidate questions that might be asked to deepen the exploration of the visited content. The integration model synchronizes the LLM with the user's browsing activity to provide a unified interaction environment and complement the information presented on the visited web pages with relevant content. This is achieved by managing an explicit conversation context and steering the LLM's behavior through implicit prompt engineering.

We pose two research questions to investigate the potential of integrating LLMs with Cultural Heritage websites:

(RQ1) *How does the extension of a CH website with question-answering support based on an external LLM impact user experience during information exploration?*

(RQ2) *How does the context-dependent generation of suggested questions impact the user experience in a CH website extended with an LLM for question-answering?*

To answer these questions, we analyzed the user experience with three versions of an existing web-based CH guide. The first (baseline) is the Triangolazioni app described in Mauro et al. (2022a, 2022b), which only presents curated content. In the other two versions, we applied our integration model to use an external LLM for advanced information provision. Specifically, the second version extends Triangolazioni by enabling users to ask free-text questions and generates context-dependent answers based on the content they previously explored. The third version is a further extension that suggests candidate context-dependent questions to help the user formulate queries.

Our integration model is agnostic to the LLM because, with the rapid technological advances of Generative Artificial Intelligence, we want to be able to replace the LLM as needed. In this work, we exploited two LLMs hosted by Perplexity Team (2024) (whose models are named pplx) to develop the context-dependent question-answering functions. Specifically, we used the 7B version of `pplx` model to suggest context-dependent questions the user might ask and the 70B version to answer users' questions.

We conducted a user study that involved 44 participants to test the usability, the user experience, and the performance of the proposed user interfaces. The results show that the LLM-powered chatbot improved the system's usability; moreover, the suggestion of candidate context-dependent questions further enhanced the user experience regarding usability and perceived performance.

In summary, this paper makes the following contributions:

- We propose a loosely coupled model to integrate LLMs into Cultural Heritage websites by synchronizing users' browsing activity with conversational turns for supporting query formulation and context-dependent information provision.
- We present a curated CH website that we extended with LLMs by applying our integration model and we report the user experience results collected through a user study.

In the following, Sect. 2 positions our work in the related research. Section 3 describes the user interfaces we evaluate. Section 4 describes the architecture to integrate the LLMs in the CH website. Section 5 provides the details of the user study and Sect. 6 presents the findings we collected. Section 7 describes our work's implications and Sect. 8 discusses its limitations. Section 9 concludes the paper.

Code and data are available on https://anonymous.4open.science/r/tell-me-more-A4E5.

## 2 Related work in Cultural Heritage exploration

Our work differs from the related one because it fuses web-based information presentation, given by curated content, and natural language conversation with an LLM to enhance content provision and user interaction in a Cultural Heritage guide.

### 2.1 Web-based Cultural Heritage guides

In the early research on digitally-supported Cultural Heritage exploration, researchers investigated recommendation and data presentation techniques for web-based and mobile guides, focusing on enhancing the user experience during the navigation of information catalogs or the visit to physical CH sites. For instance, GUIDE (Cheverst et al. 2000) and Riot! (Blythe et al. 2006) presented location-sensitive information about two British cities. PIL (Kuflik et al. 2011) personalized the interaction with the user during the visit to a museum through ubiquitous user modeling. More recently, Braunhofer and Ricci (2016) and Michalakis et al. (2021) extended tourist guides with context awareness. Other works enhanced the user experience through Augmented Reality or Virtual Reality (Bonis et al. 2009; Fenu and Pittarello 2018; Bekele et al. 2018). All these systems were based on carefully authored, internal data repositories providing the information to be presented.

Having recognized the limitations of closed knowledge bases, further work has been devoted to expanding the applications by integrating external data sources. For instance, Wang et al. (2008)'s pioneer work employed semantic web technologies to enhance item presentations by incorporating information from public ontologies like ULAN, AAT, and TGN Getty vocabularies. Kouretsis et al. (2022) leveraged the Europeana (2024) ontology to establish connections between paintings of the Renaissance era. Faralli et al. (2022) created a knowledge graph linking the Italian Cultural Heritage entities (defined in the ArCo ontology) with the concepts of the DBpedia and Getty ontologies. Kim et al. (2017) created a mobile Augmented Reality system using Semantic Web technology to deliver contextual information about CH sites that aggregate heterogeneous data and semantically connect them. Rinaldi et al. (2022) used Linked Open Data in an Augmented Reality mobile application to enhance the understanding of Cultural Heritage.

Despite the advanced personalization and interaction features, these systems were strictly based on a hypertextual model (Millard et al. 2013; Hargood et al. 2016) guiding the content exploration based on the underlying ontology. Thus, the interaction with users was based on page navigation, possibly aided by a recommender system that suggested relevant pages; however, users were not allowed to ask questions to satisfy specific information interests. Our work differs because we aim to extend curated websites with advanced question-answering support.

### 2.2 Chatbots

During the last decade, chatbots emerged as a solution to allow users to express themselves when interacting with a mobile guide (Lombardi et al. 2019). Several museums now employ this technology to communicate with users. The chatbot is the guide's user interface and works as an "info bot" to help users plan their visit (Tzouganatou 2018) by conversing with them.

For instance, Machidon et al. (2020) proposed a chatbot based on Google's DialogFlow[1] to assist users in exploring the Europeana ontology's content. Moreover, Casillo et al. (2022) developed an ontology-guided chatbot that presents information about the Archaeological Urban Park of Naples. Some chatbots provide advanced functions, like gamification, to enhance users' engagement during the visit (Varitimiadis et al. 2021). Moreover, some research leveraged Virtual Reality to model chatbots as Virtual Humans (Noh and Hong 2021; Sylaiou and Fidas 2022; Chalmers et al. 2021; Machidon et al. 2018) and Audio Augmented Reality to provide an immersive interaction environment (Tsepapadakis and Gavalas 2023). Furthermore, researchers developed embodied chatbots for museums and focused on the impact of the interaction style and modality in learning environments (Noh and Hong 2021).

Like web-based guides, chatbots relying on closed repositories might fail to answer users' questions concerning diverse aspects of a tour such as logistics and planning. To improve this ability, some projects used chatbots to acquire information about the user's context and extend natural language interaction with multimodal information to present personalized route recommendations (Casillo et al. 2020). Other projects developed complex software architectures to integrate chatbots with external services providing information about Points of Interest and events (Sperlí 2021). Moreover, Varitimiadis et al. (2021) proposed to exploit graph-based, distributed, and collaborative multi-chatbot conversational AI systems, viewing "knowledge graphs as the key technology for potentially providing unlimited knowledge to chatbot users."

As discussed in Lombardi et al. (2019), a major issue for chatbots has been understanding the human intent during the interaction, with consequences on user experience. To address this issue, chatbots have been tied to a domain of reference for interpreting users' utterances. While this approach improves performance, it ties the answering capability to the reference knowledge base. We aim to overcome these limitations by employing LLMs, which can adapt the conversation to any context without changing the underlying knowledge base.

### 2.3 Large language models

Large Language Models promise to overcome the limitations in chatbots' question-answering capabilities: they can fluently interact with users and seamlessly acquire and summarize information from public repositories to extend their knowledge.

[1] https://cloud.google.com/products/conversational-agents?hl=en.

Therefore, they can overcome the limitations related to the use of closed knowledge bases. Indeed, LLMs cannot guarantee the quality and accountability of provided information because they can extract data from unreliable sources (Dogru et al. 2023; Kim et al. 2023; Mich and Garigliano 2023; Spennemann 2023). Another problem is that LLMs raise accountability issues because they are subject to hallucinations causing the generation of wrong answers. Recent studies show that the factuality of LLMs tends to increase with their size (Tam et al. 2023). While this mitigates the problem in a question-answering task as the one we aim to solve, this is still an open research question.

Currently, LLMs are hardly integrated into Cultural Heritage websites because they are employed as conversational user interfaces to answer users' questions. To the best of our knowledge, the first work in this direction is by Trichopoulos et al. (2023a, 2023b). However, those authors fine-tuned an LLM to provide logistic data about a museum. Fine-tuning techniques and Retrieval-Augmented Generation could leverage authoritative sources to enhance the LLM's knowledge (Yasaka et al. 2024; Ardimento et al. 2024). Nevertheless, we avoid this approach because it would tie our system to a particular LLM, which goes against our goal of developing an LLM-agnostic integration model. As discussed in the introduction, we aim to create a flexible system that can easily adapt to the rapid advancements in LLM technology without being constrained to a specific technological pipeline.

Regardless of LLM's potential for knowledge discovery, we point out that (similar to traditional chatbots) they constrain the interaction with the user by imposing the exchange of messages as the only interaction means. For this reason, they lack the expressiveness of traditional websites, where people can consult rich pages showing multimodal Cultural Heritage information. To overcome these limitations, we aim to integrate the LLMs into websites.

### 2.4 Our work

We adopt a loosely coupled approach based on prompt engineering to:

- Integrate a web-based guide with an external LLM to extend the system's local knowledge through the question-answering mechanism offered by the chatbot.
- Combine free-text questions with the generation of context-dependent questions to support lay users in retrieving the information they are interested in.
- Answer users' questions by considering the explored content to prevent returning redundant or irrelevant information.

## 3 User interfaces

This section describes the three user interfaces we evaluate. These variants of the Triangolazioni guide differ in the level of information exploration support they offer. As Triangolazioni's domain knowledge is written in Italian, this is the target

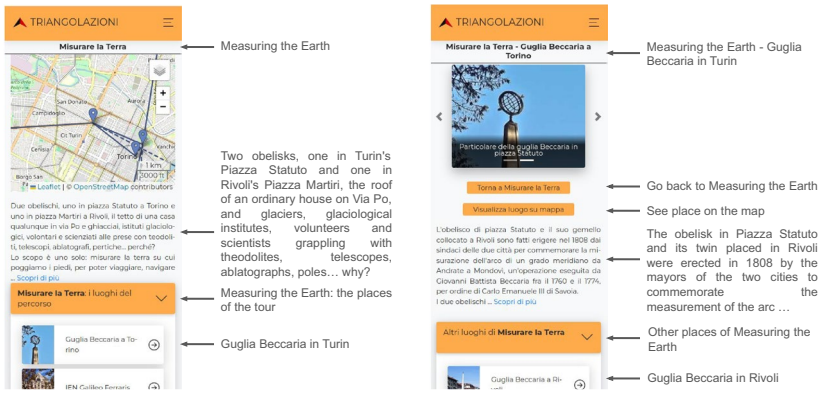


**Fig. 1** The `Baseline` user interface (Triangolazioni CH guide). **a** The user interface shows the information about the "Measuring the Earth" narrative (thematic path). **b** The interface shows the information about the place "Guglia Beccaria" within the path "Measuring the Earth"

language of our work and user study. For this reason, the following figures report the Italian text coupled with a side translation to English.

### 3.1 `Baseline`

This is the original Triangolazioni guide and proposes thematic paths describing Points of Interest in Torino and surroundings. For each thematic path, organized as a geolocalized narrative, the guide provides details of its places, historical personages, monuments, and historical or artistic objects. Moreover, it enables the user to explore interlaced paths that involve the same places or are thematically related to the current narrative to expand the knowledge about places (Mauro et al. 2022a, b).

Figure 1 shows a portion of the user interface presenting the narrative "Misurare la Terra" (Measuring the Earth), which describes methods and tools developed in the past to evaluate the distance between remote places. The guide shows a brief textual introduction, accompanied by a map containing all associated locations, as well as a list enumerating these same places (Fig. 1a, Guglia Beccaria in Torino, IEN Galileo Ferraris, etc.). Each place is visualized in an expandable component that the user can open to visualize the page describing the Point of Interest (Fig. 1b). This page shows some images of the place and a textual description contextualizing it in the selected narrative. The user may switch to other locations of the thematic path by clicking on the geographic map or browsing the list of places.

The content in `Baseline` is human-authored and is the same for each user. Users can explore thematic paths by interacting with the guide to retrieve information about Points of Interest. However, they cannot ask questions, nor explore external content to the authored one.

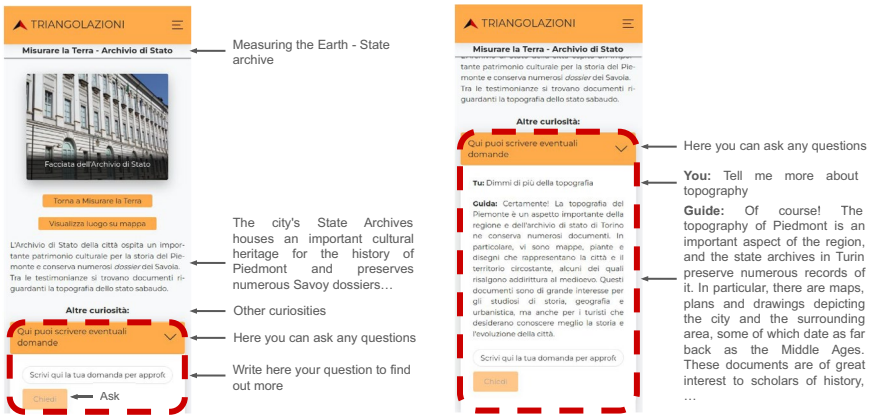


**Fig. 2** The `Chatbox` user interface within the thematic path "Measuring the Earth". The widget devoted to question-answering is highlighted through a red dotted line. **a** User interface showing the information about the State archive in what concerns Measuring the Earth. **b** The chatbox allows the user to question using a free input

### 3.2 `Chatbox`

This user interface extends the `Baseline` one. It integrates an LLM-powered chatbot into the pages about places to answer free-text questions. As shown in Fig. 2, the chatbot is visualized in a widget ("Qui puoi scrivere eventuali domande"— Here you can ask any questions) showing a free text field and including the answer generated by the LLM. In this example, the user asked: "Dimmi di più sulla topografia"—Tell me more about topography (Fig. 2b). The chatbot knows about the content of the current page and the previously browsed ones and answers the inquiries context-dependently. Therefore, its replies are personalized to the user's reading history in the mobile guide. Section 4 describes how we build the context we pass to the chatbot and the prompt to get its answer.

### 3.3 `RecChatbox`

Like `Chatbox`, this user interface enables users to ask free-text questions. Moreover, each web page describing a Point of Interest suggests three questions generated depending on its content and the interaction context, i.e., the previously visited pages. As the user can browse the thematic paths of the CH guide to learn its core content, these questions are intended to satisfy further information needs, inspiring query formulation.

Figure 3 shows a sample interaction with `RecChatbox`. The user is browsing "Istituto Angelo Mosso" (Angelo Mosso Institute, Fig. 3a), and the system suggests three possible questions that may raise her/his curiosity (Fig. 3a, b):


**Fig. 3** The `RecChatbox` user interface within the path Measuring the Earth. **a** The interface shows the information about the Institute Angelo Mosso. **b** The interface shows some suggested questions and the chatbox. **c** The interface shows the answer to one of the suggested questions

1. "Quali sono le attività di ricerca svolte all'istituto Angelo Mosso?" (What research activities are carried out at the Angelo Mosso Institute?),
2. "Come l'istituto contribuisce alla misurazione della Terra?" (How does the institute contribute to measuring the Earth?),
3. "Quali sono i programmi di divulgazione scientifica offerti all'istituto Angelo Mosso?" (What science outreach programs are offered at the Angelo Mosso Institute?).

We suppose the user is interested in the second question and clicks the "Chiedi" button (Ask). Then, the system shows the selected question and the related answer inside another "collapse" component, as though the user initiated a conversation with the guide (Fig. 3c). The user can continue asking other free-text questions or (s)he can select other suggested questions to let the system support the conversation.

## 4 Content generation

To support question-answering, we used the `pplx` model hosted by Perplexity (Perplexity Team 2024), based on Llama-2 (Touvron et al. 2023). As the factuality of LLMs tends to increase with their size (Tam et al. 2023), to answer the users' questions in the `Chatbox` and `RecChatbox` user interfaces, we used the 70B version of the `pplx` model (`pplx-70B-chat`) that exploits the full information available to the LLM for the content generation task. Differently, generating the suggestions of context-dependent questions in `RecChatbox` is creative and does not require deep domain knowledge. Thus, we used the 7B version of the `pplx` model (`pplx-7B-chat`) for this task. The reason for this choice is that the response times of LLMs increase when employing more powerful models; thus, when possible, it is advisable to use the less powerful ones. Despite this, it is important to note that the response times of the models we used were compatible with being applied in a real-world scenario, with the bigger model taking at most a few seconds to answer.

As previously mentioned, we did not fine-tune the LLM because we want to be agnostic to the chosen model, relying on external APIs (Application Programming Interfaces offering a service to a software) to query it. Given the rapid growth and improvement of Generative Artificial Intelligence, this approach allows replacing the LLM used by an application with a small effort.

Based on a preliminary experiment with the 7B and 70B `pplx` models, we set the following parameter values:

- Temperature: 0. We assigned this value for reproducibility as it makes models completely deterministic.
- Top_k: 5. This parameter represents the number of tokens the LLM should keep during content generation. We selected this value because we empirically found out that, for values of Top_k $> 5$, the model was much more likely to generate answers partly in Italian (our desired language), and partly in English.
- Presence penalty: 1. This parameter ranges from $-2.0$ to 2.0. A higher value means repeated tokens will be penalized more. In turn, this encourages the LLM to generate original output. We empirically set this parameter to 1. This value requires a slight novelty in the generated content and, in our tests, it made the models' answers less redundant, without compromising their truthfulness.

### 4.1 Building the conversational context

To synchronize the LLM-powered chatbot with the web-based CH Guide in the `Chatbox` and `RecChatbox` user interfaces, the chatbot needs a representation of the interaction context describing (i) the content the user has explored while browsing the web pages of the guide, and (ii) the information provided while answering the user's questions. We decided to represent this context (CTX in the following) as a *conversation* between the user and the CH guide. In other words, the user's

**Table 1** Prompt templates used to build the conversation context that SC uses to interact with the APIs of the LLM-powered chatbot. The text is translated from the Italian language

| | |
|---|---|
| P1 | Hello, I am your fully Italian-speaking digital tour guide for Torino. How can I assist you? |
| P2a | I am beginning this thematic path in the Torino area: [NAME OF THE PATH]. Please tell me something about it |
| P2b | I am looking at this place in the Torino area: [NAME OF THE PLACE] in the path [NAME OF THE PATH]. Please tell me something about it |
| P3 | You are a tour guide who responds entirely in Italian and is grammatically correct. Please write 3 very short and engaging questions about [PLACE NAME] in the thematic path [PATH NAME], based on our conversation so far, but whose answer is not contained in it. Please, each line should contain only the question, without quotation marks |
| P4 | You are a tour guide of Torino who responds entirely in Italian and you are grammatically correct. Please briefly answer the following question, without repeating what we have said in this conversation: [QUESTION] |
| P5 | Carefully read this text: [TEXT] Please translate it into Italian |

browsing activity is modeled as a dialog where the user asks for information and the guide responds by presenting the requested web pages. These pages include curated, static content and the chatbot's generated content. Moreover, CTX includes dialog turns representing the user's questions for the chatbot and its responses.

The core engines of the chatbot (LLM models) are stateless. Therefore, an external software component must build the *conversation context CTX* by keeping track of the browsing and question-answering activities, and feed the chatbot with CTX at each invocation. The LLM's context window (an intrinsic property of the model) determines the maximum length of CTX. For example, `pplx` has a context window of 4096 tokens.

We integrated the web-based guide and the LLM-powered chatbot through a loosely coupled architecture where a *simulated user (SU)* and a *simulated chatbot (SC)* act as proxies of the *human user* and the *LLM-powered Chatbot* respectively. SU and SC run in the web browser and mediate the communication between these entities. Moreover, they fuse the content of the web pages with the LLM-generated content into the conversation context (CTX). The interaction diagram of Fig. 4 describes the communication flow while the user interacts with the CH guide:

1. At the first interaction (i.e., when the user visits the first web page of the guide), CTX is initialized with a message that describes a welcome dialog act from the simulated chatbot SC to the simulated user SU (P1 in Table 1, Step 1 in Fig. 4).
2. Each time the human user visits a web page of the guide[2]:

    (a) The user interface feeds the simulated user SU with the URL and title of the page, which denotes a topic T like "Istituto Angelo Mosso", or "Misurare la

[2] To avoid redundancies in the conversation context, duplicates are automatically removed. For instance, if the user visits a specific page twice, only the second interaction is retained in CTX.

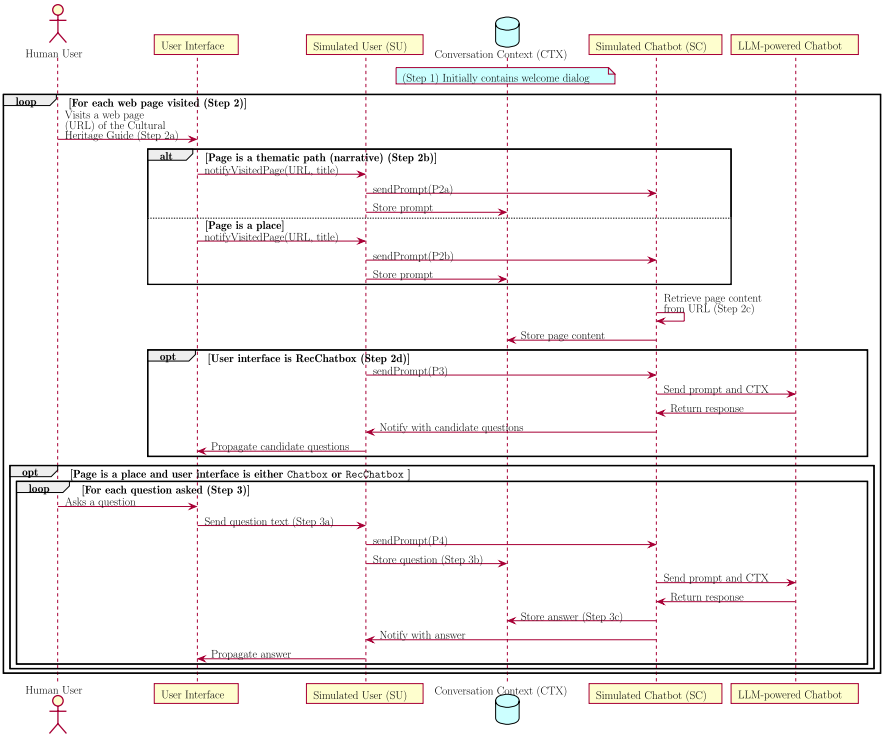


**Fig. 4** Interaction diagram. Message flow to synchronize the web-based CH guide with the LLM-empowered chatbot

Terra". See the message "notifyVisitedPage(URL, title)" of the interaction diagram. This marks the beginning of Step 2b in Fig. 4.

(b) To represent the human user's intent when visiting the web page, SU invokes the simulated chatbot SC through a prompt specifying that SU wants to learn more about T. If the human user has visited a web page describing a thematic path (i.e., a narrative) the message is "sendPrompt(P2a)". If the user has visited a web page describing a place, the message is "sentPrompt(P2b)". Then, SU stores the prompt in the conversation context CTX. Table 1 shows the text of P2a and P2b.

(c) In turn, SC retrieves the content of the visited web page from the URL (Step 2c in Fig. 4) and stores this content in CTX. This is done to fuse the representation of web browsing and question-answering as if the LLM provided the page content to answer SU's request. Thus, CTX maintains the information that SU has received the page content.

(d) Because the `RecChatbox` user interface offers question generation support, it includes the following steps of the interaction flow (Step 2d):

  (i) SU invokes SC (sendPrompt(P3)) asking for three candidate questions to be proposed to the user, based on CTX. In this case, SU does

not store the message in CTX because this portion of the conversation is instrumental in generating candidate questions and does not reflect an exchange of information with the human user. Prompt P3, reported in Table 1, specifies that the questions must be short, engaging, and relevant to the selected place and thematic path. Moreover, it specifies that the questions must consider the conversation context but cover external information items to avoid repetition.[3]

(ii) In turn, SC invokes the APIs of the LLM sending the received prompt, and the content of CTX.

(iii) Then, SC retrieves the response, which consists of three candidate questions, and notifies SU (Notify with candidate questions). The questions are not stored in CTX for the same reason as above.

(iv) Finally, SU propagates the questions to the user interface of the CH guide, which shows them as options for the human user (Propagate candidate questions).

3. Each time the user asks a question (free-text or candidate one):

(a) The user interface sends SU the text of the question (Step 3a in Fig. 4).

(b) SU notifies SC (sendPrompt(P4)) and stores the question in CTX (Step 3b). SC sends the LLM-powered chatbot a prompt including the question ("sendPrompt(P4)") and CTX.

(c) When SC receives the chatbot's response, SC stores it in CTX (Step 3c)) and notifies SU. Then, SU propagates the answer to the user interface of the CH guide.

By building the conversation context in this way, our system integrates the user's navigation history within the website and the conversation with the LLM-powered chatbot. This allows the LLM to provide context-aware answers considering the previously explored content.

We remind the reader that SC always prompts the LLM by sending the conversation context (CTX) because the APIs of the LLMs' core engines are stateless. However, the user is unaware of the interaction between SU and SC because CTX is hidden and the answers generated by the chatbot are embedded in the visited web page, as described in Sects. 3.2 and 3.3. Figure 5 in Appendix A depicts the flow of the sample interaction shown in Fig. 3.

### 4.2 Translating English text

Despite our efforts, sometimes the output generated by the LLM is partially or entirely written in English. This is common for Llama models (the base of the

[3] The prompt repeats the chatbot must answer in Italian to enforce the language choice. This is because the Llama model tends to respond in English, see Sect. 4.2.

pplx models) because they have been trained with a corpus almost entirely written in English (Touvron et al. 2023).

To address this issue, we used the languagedetect library (Ribreau and Zavackiy 2023) in the simulated chatbot SC to detect the presence of English content in the answers it receives. In the positive case, SC asks the pplx-70B-chat model to translate the English sentences to Italian using prompt P5 of Table 1. Then, it replaces the original English output with the translated text in the response generated by the chatbot and stores the complete text in CTX. Finally, it sends the overall text to SU that propagates the answer to the user interface.

## 5 User study

We conducted a user study to assess participants' experience with the three interfaces described in Sect. 3; see Figs. 1, 2 and 3. The University of Torino's Ethics Committee approved our experiment (Protocol Number: 0421424).

We developed an extended version of the Triangolazioni CH guide to support the user study. This web-based mobile app guided participants through the experiment, recording their click and scroll actions for subsequent behavioral analysis. No personally identifiable information was collected to safeguard users' privacy, and numerical identifiers were used to tag the acquired anonymous data during interaction sessions.

The user study occurred from December 1st, 2023 to January 5th, 2024, with an average duration of 17.62 min per participant (Standard Deviation = 8.35).

### 5.1 Structure of the experiment

The study employed a within-subjects approach, considering each treatment condition (Baseline, Chatbox, RecChatbox) as an independent variable. We managed the experiment through a master web application that guided the interaction flow using Google Forms to elicit users' responses. Participants experienced all treatments, and the application presented tasks in a counterbalanced order to mitigate fatigue and practice biases. No time constraints were imposed and the study procedure comprised the following steps:

1. The application initially presented the informed consent (available at https://bit.ly/4ak1IF0) and requested participants' explicit agreement. Moreover, it asked them to confirm that they were 18 or older (a mandatory condition to continue the experiment).
2. Participants then answered questions about demographic information, cultural background, and familiarity with Generative AI systems like ChatGPT.[4] Moreo-

[4] https://chatgpt.com/.

ver, they answered the Curiosity and Exploration Inventory-II (CEI-II) questionnaire (Kashdan et al. 2009), which we used to determine the impact of people's curiosity on the user experience with the three user interfaces. CEI-II supports the understanding of people's motivation to seek knowledge and new experiences, and their willingness to embrace everyday life's novel, uncertain, and unpredictable nature. We were interested in this aspect because we hypothesized that users' curiosity might influence their navigation patterns in a tourism application and their perception of functionalities such as having a chatbot to satisfy their curiosity and receiving suggested questions.

3. Then, participants interacted with the three user interfaces in a counterbalanced order. For each interface, the application instructed users by presenting the following text:

   > "Please, explore the system you will find here: <click here>. After reading the introduction to the story, explore 1–2 places related to it. When you are done exploring, click on the following button: <Let's continue>".

   Participants explored the Points of Interest of a different thematic path for each user interface. After each interaction, the master web application administered a post-task questionnaire to assess their experience by asking the questions of the System Usability Scale (SUS) (Brooke 1996) and the questions concerning the Performance section of the Trust Of Automated Systems Test (TOAST) (Wojton et al. 2020). The SUS responses were in the "Strongly disagree" to "Strongly agree" scale, mapped to values between 1 and 5. For the TOAST they were on a scale from 1 to 7. We added attention checks to these questionnaires to verify people's careful engagement in the study.
4. Finally, participants could leave a free-text comment about the interface they used.

### 5.2 Participants

We recruited volunteer participants through public mailing lists and social networks. For statistical significance, the experiment targeted a sample of 42 subjects, determined by power analysis with $\alpha = 0.05$, power $= 0.8$, and effect size $= 0.4$. However, as 44 people answered our invitation and nobody was excluded by the attention checks, our final sample size was 44, slightly larger than the planned one. In the following, we report participants' descriptive statistics.

- **Gender:** 20 females, 24 males, 0 not-binary, and 0 not-declared.
- **Age distribution**: 18–20 (0 participants), 21–30 (29), 31–40 (11), 41–50 (2), and 51–60 (2).
- **Education level:** middle school (1), high school (4), university (34), and Ph.D. (5).
- **Background:** scientific (24), humanistic (14), technical (3), linguistic (2), other fields (1).

- **Familiarity with computers:** 15 participants classified themselves as advanced computer users, 19 as average users, and 10 as beginners.
- **Familiarity with LLMs:** 5 people identified themselves as advanced users of systems like ChatGPT, 9 as average users, 18 as beginners, and 12 said they had no experience with those systems.

## 6 Results

### 6.1 Questionnaires

Table 2 presents the results for the SUS and Performance (TOAST) questionnaires. The Kruskal-Wallis test shows statistically significant differences between the three user interfaces across many of the questions, and this result is confirmed by the Wilcoxon Rank-Sum test performed between the best and second-best conditions. `Chatbox` and `RecChatbox` outperform `Baseline`. This finding suggests that integrating a chatbot enhanced the usability of the Triangolazioni CH guide and users' trust in its performance. Moreover, `RecChatbox` received higher ratings than `Chatbox` across all questions. This indicates that providing suggested questions enhances the usability and interaction with the chatbot.

More specifically, `RecChatbox` significantly outperforms the other interfaces in users' desire to use the system frequently (S1). It is also deemed the one where the functions are best integrated (S5), and the less inconsistent one (S6 and T2), highlighting that the suggested questions improve the clarity of the user interface. Moreover, `RecChatbox` made the user experience less cumbersome (S8) and made users more confident while using the system (S9), helping them achieve their goals (T1). Furthermore, it increased users' trust in the system regarding reliability (T3 and T4), making them feel comfortable to rely on the provided information (T5). Conversely, the results weakly suggest that `Baseline` is the easiest user interface (see S2 and S4).

In most cases, `Chatbox` is rated better than `Baseline`, except for S9 (*I felt very confident using the system*). This further indicates that the use of suggested questions may help balance out the added complexity of using a chatbot, particularly for users who are not familiar with such technologies, thereby providing useful support.

The above findings are reflected in both the SUS score (77/100 for `Baseline`, 80/100 for `Chatbox`, and 88/100 for `RecChatbox`) and the Performance one (TOAST, 4.9/7 for `Baseline`, 5.7/7 for `Chatbox` and 6.2/7 for `RecChatbox`).

To investigate the impact of curiosity on user experience, we divided participants based on their CEI-II scores using the median value of the sample as a threshold (3.5). Table 3 shows the SUS and Performance (TOAST) scores. The difference between the user interfaces is statistically significant across both partitions. As expected, participants with a lower CEI-II score rated `RecChatbox` significantly higher than `Chatbox`, with a difference of 14 points on the SUS scale. This might be because, unlike people with higher curiosity levels (who rated the two interfaces almost equally), these participants preferred to be guided by the system through

**Table 2** Results of the SUS and TOAST questionnaires

| SUS (5-point scale) | `Baseline` | `Chatbox` | `RecChatbox` | *p-value* |
|---|---|---|---|---|
| S1: I think that I would like to use this system frequently | 3.09 (1.20) | 4.14 (0.59) | **4.50 (0.70)** ⋆ | 5e-9*** |
| S2: I found the system unnecessarily complex. † | **1.50 (0.70)** | 1.84 (0.89) | 1.55 (0.76) | 0.067 |
| S3: I thought the system was easy to use | 4.25 (0.65) | 4.34 (0.61) | **4.50 (0.59)** | 0.158 |
| S4: I think that I would need the support of a technical person to be able to use this system. † | **1.39 (0.65)** | 1.61 (0.92) | 1.43 (0.76) | 0.536 |
| S5: I found the various functions in this system were well integrated | 3.52 (0.79) | 4.23 (0.68) | **4.55 (0.66)** ⋆ | 8e−9*** |
| S6: I thought there was too much inconsistency in this system. † | 2.02 (0.95) | 1.73 (0.95) | **1.50 (0.73)** | 0.009*** |
| S7: I would imagine that most people would learn to use this system very quickly | **4.09 (0.83)** | 3.55 (1.25) | **4.09 (0.74)** | 0.072 |
| S8: I found the system very cumbersome to use. † | 1.93 (0.87) | 1.66 (0.83) | **1.27 (0.54)** ⋆ | 1e−4*** |
| S9: I felt very confident using the system | 4.18 (0.72) | 4.05 (0.99) | **4.52 (0.79)** ⋆ | 0.009** |
| S10: I needed to learn a lot of things before I could get going with this system. † | 1.41 (0.73) | 1.39 (0.62) | **1.25 (0.49)** | 0.463 |
| TOAST (Performance, 7-point scale) | | | | |
| T1: The system helps me achieve my goals | 3.82 (1.93) | 5.64 (1.04) | **6.20 (0.82)** ⋆ | 1e−10*** |
| T2: The system performs consistently | 5.16 (0.96) | 5.8 (0.85) | **6.23 (0.86)** ⋆ | 2e−6*** |
| T3: The system performs the way it should | 4.95 (1.40) | 5.68 (0.98) | **6.30 (0.70)** ⋆ | 2e−6*** |
| T4: I am rarely surprised by how the system responds | 5.41 (1.24) | 5.66 (1.41) | **6.34 (1.01)** ⋆ | 2e−4*** |
| T5: I feel comfortable relying on the information provided by the system | 5.25 (1.43) | 5.61 (1.24) | **6.05 (1.01)** ⋆ | 0.013* |

The best values are in boldface. Stars denote the statistical significance of the difference between the three user interfaces (Kruskal-Wallis test). Significance levels: (***)$p < 0.001$, (**)$p < 0.01$, (*)$p < 0.05$. The questions formulated negatively, for which lower values are the best, are denoted with a †. The best values are marked with a ⋆ if both the p-value of the Kruskal-Wallis test and the Wilcoxon rank-sum test between the best condition and the second-best condition yields $p < 0.05$

the generated questions. This may also explain why people with low CEI-II scores ranked `Chatbox` the worst concerning the trust in its performance: that interface requires the highest exploration effort in formulating questions without offering any support to do that. Conversely, the pairwise statistical test for both SUS and TOAST does not reveal a significant difference between `Chatbox` and `RecChatbox` for the high CEI partition. However, the test yields a p-value $< 0.05$ when

**Table 3** SUS and Performance (TOAST) scores for the whole sample and the subsets based on the CEI-II score

| | | Low CEI-II, 22 people | High CEI-II, 22 people | Overall |
|---|---|---|---|---|
| *SUS* | `Baseline` | 76 | 78 | 77 |
| | `Chatbox` | 75 | 85 | 80 |
| | `RecChatbox` | **89** ⋆ | **87** | **88** ⋆ |
| | *p-value* | 3e−4*** | 0.014* | 2e−5*** |
| *TOAST* | `Baseline` | 5.3 | 5.4 | 4.9 |
| | `Chatbox` | 4.4 | 6.0 | 5.7 |
| | `RecChatbox` | **6.3** ⋆ | **6.2** | **6.2** ⋆ |
| | *p-value* | 7e−7*** | 0.003** | 6e−9*** |

The SUS score is in [0,100], and the TOAST score is in [1,7]. The notation for p-values is the same of Table 2

comparing `Baseline` with both `Chatbox` and `RecChatbox`. This finding suggests that for people with high CEI, the ability to ask questions is valuable albeit they may find less value in the recommended questions and prefer to formulate queries independently.

Table 4 shows the results of the post-task questions related to the generated content. These findings indicate that `RecChatbox` made it easier to ask questions (C1). Moreover, users deemed the generated questions understandable (C4, 4.73/5) and interesting (C5, 4.45/5). Since the LLM model and the context-building method behind the chatbot were the same across `Chatbox` and `RecChatbox`, we did not expect to find a statistically significant difference. The ratings about the understandability of answers (C2) are good (4.44/5 and 4.68/5, respectively).

**Table 4** Evaluation of the questions proposed by `RecChatbox` and the answers provided by `Chatbox` and `RecChatbox` on a 5-point scale

| | `Baseline` | `Chatbox` | `RecChatbox` | *p-value* |
|---|---|---|---|---|
| C1: Asking questions was easy | – | 3.88 (1.22) | **4.57 (0.66)** ⋆ | 0.004** |
| C2: I understood the answers provided by the system | – | 4.44 (0.85) | **4.68 (0.56)** | 0.196 |
| C3: I thought the answers provided by the system were interesting | – | 4.23 (0.87) | **4.55 (0.59)** | 0.090 |
| C4: I understood the questions provided by the system | – | – | **4.73 (0.50)** | – |
| C5: I thought the questions provided by the system were interesting | – | – | **4.45 (0.76)** | – |

The statement about the suggested questions was only asked after the interaction with `RecChatbox` because this is the only user interface providing suggestions. We present results using the same notation as in Table 2

### 6.2 Log data

The master web application collected data on user behavior while participants interacted with the user interfaces.

We consider the average time a user spent interacting with a single interface, including activities such as browsing Points of Interest (POIs) and asking or reading questions: users spent the least time exploring `Baseline` (1 min and 37 s on average), followed by `RecChatbox` (3 min and 41 s) and `Chatbox` (4 min and 19 s). We may infer that the two interfaces with the chatbot were more engaging, leading to more thorough exploration. However, the suggested questions in `RecChatbox` saved users' time because they required less typing and a lower cognitive effort to ask questions.

When using `Chatbox`, users wrote an average of 2.54 free-text questions. In `RecChatbox`, they wrote an average of 2.14 free-text questions. Moreover, they clicked on an average of 2.75 suggested questions. Thus, the mean number of submitted questions was 4.98. These numbers indicate that while taking less time overall, users explored more customized information in `RecChatbox` than in `Chatbox`.

### 6.3 Topic analysis

To understand the performance of the LLM-powered chatbot, we conducted a topic analysis using the `Bertopic` library (Grootendorst 2022) in conjunction with `spaCy` (Explosion 2017):

- First, we extracted the named entities contained in all contents (entered by the user, static - i.e., curated content, or generated by the chatbot). We used these named entities for zero-shot classification, suggesting them as possible topics to the model.
- Because our content is in Italian, we chose model `distiluse-base-multilingual-cased-v1`. We estimated the best parameters for the dimensionality reduction and clustering phases using Bayesian optimization with the silhouette score as a target measurement.
- For each Point of Interest explored during the experiment, we calculated the topic overlap among the different kinds of content related to it as the average cosine similarity between the topic distributions of each pair of documents.

As shown in Fig. 1, there are two categories of static content: the first concerns thematic paths, and the second regards Points of Interest within a thematic path.

Table 5 contains the overall data from the topic analysis. We found a fairly high topic overlap between the generated questions and the static content, both about the Points of Interest (POI) and the thematic paths (62.39 and 64.63% respectively). This suggests that the LLM-powered chatbot leverages the static content to suggest follow-up questions related to what users just read but introduces some diversity.

**Table 5** Overall results from the topic analysis

| | User questions | Generated answers | Generated questions | Static content (POI) | Static content (thematic path) |
|---|---|---|---|---|---|
| User questions | – | 0.4212 | 0.3662 | 0.4402 | 0.5137 |
| Generated answers | 0.4212 | – | 0.5565 | 0.6522 | 0.7006 |
| Generated questions | 0.3662 | – | – | 0.6239 | 0.6463 |
| Static content (POI) | 0.4402 | – | – | – | 0.7736 |
| Static content (thematic path) | 0.5137 | 0.7006 | 0.6463 | 0.7736 | – |

The overlaps between different types of content are represented in [0,1]. Static content (POI) denotes the information provided by the web pages of the CH guide that the user previously visited

Conversely, the overlap between the static content and the questions posed by users is lower (44.02% concerning Points of interest; 51.37% regarding the thematic paths), indicating that users tend to ask about topics partly diverging from those presented by the system. Notice that some participants asked questions unrelated to the current page, apparently to test the system's abilities (for example, "which are the [soccer] teams from Milan?"). This behavior partly skewed the results.

The topics overlap between the generated answers and the static content is 65.22% for POIs and 70.06% for thematic paths. The higher overlap concerning the paths is positive, meaning that the chatbot considers the current thematic path rather than providing general information about a place. This indicates that the context is successfully retained and factored in when generating answers.

Table 6 shows the topic analysis data regarding `Chatbox` and `RecChatbox`:

- For `Chatbox`, the overlap between the user-crafted questions and the generated answers is 43.54%. The static content about places overlaps with the user questions by 44.85% and the one concerning thematic paths overlaps by 51.23%. Both findings align with the overall data (Table 5). Conversely, the overlap between generated answers and static content is slightly lower than in the overall data. This can be explained by the fact that the generated questions are typically more focused than the user-crafted ones.
- Regarding `RecChatbox`, the overlap between user questions and generated answers is 42.43%, in line with the other findings. The overlap between the system and user-generated questions is 35.37%, further demonstrating that users tend to diverge from the presented topics when asking questions autonomously. The overlap between the user questions and the static content is in line with the overall data; however, the overlap between the generated answers and the static content is higher (66.70 and 72.66%), as well as the overlap between generated questions and static content (62.39 and 64.63%).

**Table 6** Topic analysis results for `Chatbox` and `RecChatbox` concerning the overlap between different types of content

| | User questions | Generated answers | Generated questions | Static content (POI) | Static content (thematic path) |
|---|---|---|---|---|---|
| User questions (`Chatbox`) | – | 0.4354 | – | 0.4485 | 0.5123 |
| Generated answers (`Chatbox`) | 0.4354 | – | – | 0.6264 | 0.6555 |
| User questions (`RecChatbox`) | – | 0.4243 | 0.3573 | 0.4303 | 0.5155 |
| Generated answers (`RecChatbox`) | 0.4243 | – | 0.5750 | 0.6670 | 0.7266 |
| Generated questions (`RecChatbox`) | 0.3573 | 0.5750 | – | 0.6239 | 0.6463 |

We use the same notation as in Table 5

These results indicate that the LLM-powered chatbot generated properly focused answers and maintained a certain degree of diversity to avoid repeating what the user previously explored. Moreover, the generated questions helped keep the conversation focused on the content presented by the Triangolazioni CH guide.

### 6.4 Free-text answers analysis

We identified the main themes emerging from the collected answers through Thematic Analysis (Braun and Clarke 2006). This integrates the quantitative results we collected with further aspects we could not explicitly evaluate through questions. Moreover, it enhances data interpretation using the explanations provided by participants. For this task, we first read and discussed the free-text answers to acquire familiarity with data. Then, we systematically organized the answers, depending on the topics they addressed, and we selected each sentence that provided feedback about the interaction with the app's user interface or with the LLM-powered chatbot. After that, we grouped the selected sentences into themes and discussed whether these themes made sense, revising them until we reached an agreement. We identified five themes, grouped by user interface where they occurred:

- `Baseline`: (1) *Absence of the chatbot*, and (2) *Little information available*.
- `Chatbox`: (3) *Difficulties in formulating the questions*.
- `RecChatbox`: (4) *Motivation to deepen the content through the suggested questions*.
- In both `Chatbox` and `RecChatbox`: (5) *Wrong answers to the users' questions by the chatbot*.

In the following, we describe each theme in detail and we report some sample quotes of participants' answers related to it. All quotes are translated from the Italian language.

1. `Baseline`: *Absence of the chatbot*, mentioned by 6 participants who probably previously interacted with one of the other user interfaces. As shown in the sample below, people missed the possibility to ask questions to the chatbot:

   > *"I missed the opportunity to be able to explore the topics in depth by asking questions."*
   >
   > *"I would have had some curiosities to ask but it was not possible."*
   >
   > *"Some places have a lot of information and others less like the Egyptian Museum. On the Egyptian museum, I wish I could ask for more information."*

This is also reflected in the answers to the post-task questionnaire (SUS and TOAST) where `Baseline` received the worst evaluations.

2. `Baseline`: *Little information available,* mentioned by 6 participants. People complained about the lack of information in the core knowledge of the system. Indeed, Triangolazioni was developed to narrate a story through different places; thus, the curators did not add basic and practical data. Some samples:

   *"There was no basic information about the places such as opening hours or information found on Wikipedia."*
   *"There is too little information. I would have liked to ask for other information."*

   This is reflected in the post-task questionnaire (SUS and TOAST) by the low values that the `Baseline` user interface obtained.
3. `Chatbox`: *Difficulties in formulating the questions*, reported by 7 participants. Some people specified that it was easier when they received suggested questions as in the `RecChatbox` user interface. For instance:

   *"It was more difficult to find something to ask. When there were suggested questions, it was easier."*
   *"The wording of the questions, without guidance from the system, is challenging."*
   *"I think some people may not know which questions to ask, and how."*

   This is consistent with Table 2, where `Chatbox` obtained a lower and statistically significant value concerning the statement "C1: Asking questions was easy".
4. `RecChatbox`: *Motivation to deepen the content through the suggested questions*. 10 people mentioned this theme. Most comments deal with the possibility of finding new curiosities in the suggested questions and to deepen the knowledge:

   *"Having questions available increases my curiosity."*
   *"Very nice to learn new things from the list of suggested questions. I never thought I would ask for some of those things."*
   *"Nice idea to include suggested questions. I think it is important to make users that are less familiar with the technology understand how to ask questions to the AI."*
   *"This is the right middle ground, through the questions I had a starting point to go deeper into the topics."*

   This is also reflected by the log data that reveals more engagement in the interaction with the chatbot. Indeed, people asked on average 4.98 questions using `RecChatbox` while with `Chatbox` they wrote 2.54 questions.
5. `Chatbox` and `RecChatbox`: *Wrong answers to the users questions by the LLM*. This theme appeared 9 times (out of 342 answers generated by the LLM) in the comments of 5 participants when interacting with `Chatbox` and `RecChatbox` user interfaces. People complained that sometimes the system did not understand the question and thus answered wrongly. For instance:

*"The system did not understand one of my questions. I had to write it more specifically by making explicit the place name."*
*"Again, as mentioned before, the system failed to give me the answer several times until I suggested the answer."*
*"I asked the system a question, but I don't think that it got it."*

This is related to the performance of the employed LLM that might hopefully improve with technological maturity.

### 6.5 Discussion

The user study results allow us to answer the research questions we posed.

(RQ1) *How does the extension of a CH website with question-answering support based on an external LLM impact user experience during information exploration?*

The findings reveal that extending a web-based CH guide (Triangolazioni) with an LLM-powered chatbot through our integration model positively impacted user experience. It improved usability and trust in the system's performance compared to a pure hypertext CH website.

Moreover, since the LLM integrates multiple sources of information, it was able to satisfy most users' information needs. This is shown in the questionnaires and free-text comments despite the imperfections in the answers, which users reported very rarely (see the topic (5) of Sect. 6.4). This finding might vary significantly depending on the LLM used in the system. However, this is not a major issue because our integration model supports the replacement of this component with little effort. Noticeably, in the user study, users found answers understandable and interesting, as revealed by questions C2 and C3 of Table 4.

The presence of a chatbox in the CH guide made it more engaging than a traditional web-based guide, as indicated by question S1 in the post-task questionnaire (Table 2) where users declared that they would like to use `Chatbox` and `RecChatbox` more frequently than `Baseline`. This is also supported by the interaction time: users spent about one minute exploring when using `Baseline`, compared to about four minutes when using `Chatbox` and `RecChatbox`.

Furthermore, the conversation context CTX enabled the chatbot to synchronize with the content previously explored by the user while browsing the web pages of the CH guide. The chatbot successfully integrated this content, which was sometimes deemed insufficient by users, as reported in topics (1) (*Absence of the chatbot*) and (2) (*Little information available*) of the thematic analysis concerning the free-text comments. However, topic (3) (*Difficulties in formulating the questions*) shows that sometimes users struggled in the formulation of spontaneous questions.

(RQ2) *How does the context-dependent generation of suggested questions impact the user experience in a CH website extended with an LLM for question-answering?*

Our findings reveal that the context-dependent generation of suggested questions positively impacts user experience. Questionnaire data indicate that `RecChatbox` is better than `Chatbox` concerning usability and trust in performance. `RecChatbox` also made it easier to ask questions (question C1, Table 4), providing better support to users. Interestingly, users with a low CEI-II score rated `RecChatbox` significantly higher, suggesting that it offers superior support for this group. In contrast, users with a high CEI-II score did not report a substantial difference in the perceived usability and trust between the two models; see Table 3.

Participants considered the suggested questions understandable and interesting (questions C4 and C5, Table 4). Moreover, the questions induced users to explore Points of Interest thoroughly. Participants asked about 5 questions in `RecChatbox`, compared to approximately 2.5 in `Chatbox`. This is also reflected in the topic (4) (*Motivation to deepen the content through the suggested questions*) where users said that the suggested questions increased their curiosity and allowed them to learn new things.

Furthermore, the topic overlap between the suggested questions and the content of the web pages browsed by the user (static content) was fairly high, showing the efficacy of the LLM-powered chatbot in generating pertinent but original questions. Finally, the overlap between generated answers and static content was higher in `RecChatbox` than in `Chatbox`, further proving the advantages of generated questions in keeping the conversation focused.

## 7 Implications

### 7.1 Theoretical implications

Regarding Human-Computer Interaction and Information Retrieval, the integration model we propose combines the benefits of web-based and conversational user interfaces into a unified environment that supports advanced information exploration and assisted information search. Moreover, it enables the retrieval of the content collected by the LLM while preserving the integrity of the curated content contained in the web pages of the CH website. This distinction is also enforced in the proposed user interface, which visually separates the chatbot component from the original content of the CH guide.

From the software engineering perspective, the proposed model supports the seamless integration of LLM-powered chatbots in web-based applications, supporting their synchronization with the user's browsing activity through implicit prompt engineering. Specifically, the proxies mediating the interaction between the human user and the chatbot make it possible to manage a conversation context for synchronizing the navigation in the website with an LLM and steering the generation of relevant prompts to extend the website's knowledge.

### 7.2 Practical implications

From the content-provision viewpoint, integrating a Cultural Heritage web-based guide with an LLM-powered chatbot extends the information it can provide beyond its knowledge base. Moreover, it reduces the need to integrate multiple external knowledge bases, leveraging LLMs' capability to discover and digest information collected on the web. This aspect covers three complementary perspectives:

1. Enabling the CH guide to provide users with more information about Cultural Heritage items.
2. Enhancing the automated presentation of detailed information about CH items, advocated in recent work about automated curation support (Dror et al. 2024).
3. Covering a broader spectrum of topics, not only strictly related to Cultural Heritage information but also complementary issues, such as details about Points of Interest's opening hours and locations.

As discussed, e.g., in Trichopoulos et al. (2023b) and Dror et al. (2024), the quality of generated content is key to making CH site directors and curators accept this technology. To address this issue, we separated the generated content from the curated one in the `Chatbox` and `RecChatbox` user interfaces. However, work is needed to address the accuracy issue more broadly.

Another practical implication concerns user experience: integrating an LLM-powered chatbot in a web-based CH guide combines the power of multimodal presentation techniques developed to convey complex and articulated content with a colloquial interaction style supporting question answering. Moreover, it opens the venue to speech interaction, which people might prefer in some circumstances.

Finally, our integration model supports the development of systems agnostic to the LLM to be used. With the quick evolution of Generative Artificial intelligence, this aspect is key to avoid tying a website to a specific technology.

## 8 Limitations and future work

Despite the effort we put into our work, we acknowledge some limitations. From the technological point of view:

- We empirically set the parameters for the LLMs based on preliminary tests we conducted for our specific use case. A more rigorous approach, e.g., based on grid search, would be desirable. However, it would require an annotated dataset of conversations between LLMs and human users in the Cultural Heritage domain. Unfortunately, we could not find this type of resource.
- In a few cases the LLM-powered chatbot generated wrong answers because it did not understand users' questions (see Sect. 6.4). We might fine-tune the LLM for our specific use case to improve performance and reduce the probability of hallucinations, a shortcoming of current technology. However, this approach

would tie our system to a specific LLM. Given the low number of mistakes we observed, we prioritize being agnostic to the technology underlying the chatbot.
- Even though, in the user study, we observed very few mistakes in the information provided by the LLM-powered chatbot, accuracy is a major concern for the acceptability of this type of technology in Cultural Heritage and deserves further investigation.

Regarding the user study:

- The participants' sample (rather young, with probably above-average technology affinity) might not represent all possible groups of users of a typical CH website. Further experiments are needed to investigate user experience with older people.
- The adoption of a within-subjects approach allowed for direct comparisons between user interfaces and reduced the impact of individual differences. However, it might have introduced a comparison bias caused by the order of interaction with the systems. Some participants noticed functionality differences between versions which might have influenced their perceptions and ratings of subsequent interfaces. This awareness could have affected the naturalness of their interactions and judgments, particularly when encountering interfaces with fewer features after experiencing more advanced ones. Our future studies might consider employing a between-subjects design to mitigate these issues and provide an independent evaluation of each interface variant.
- The use of a median-split for the CEI condition may have resulted in a significant loss of information (MacCallum et al. 2002). This approach divides participants into two groups (high CEI and low CEI) based on the median score, which oversimplifies the continuous nature of CEI data. In future studies, we plan to employ continuous models to capture and analyze the full range of CEI scores, preserving the nuanced information within the data.
- Another potential limitation is the risk of users misusing the chatbot functionality. They might attempt to use the system for purposes unrelated to Cultural Heritage exploration, such as testing its limitations or eliciting inappropriate responses. While our integration model aims to keep conversations focused on relevant content through context-aware prompts, it may not entirely prevent all forms of misuse. Future work could explore additional safeguards to mitigate this risk while maintaining the system's effectiveness for genuine CH inquiries.

Finally, our mobile guide is designed to be used before a visit and onsite. Therefore, it is worth exploring the interaction with the LLM via speech recognition and text-to-speech to evaluate whether this improves the user experience and whether the interaction modality might need to be adapted to the context surrounding the user. For instance, a very lightened environment might challenge text reading but a noisy one might harness speech interaction.

## 9 Conclusion

This paper investigated the effects of integrating an LLM-empowered chatbot into a web-based Cultural Heritage guide and the implications of proposing questions to stimulate user engagement during information exploration.

For this analysis, we conducted a user study involving 44 participants and tested three user interfaces. The first one is a curated web-based guide presenting static web pages. The second extends the guide with the possibility of asking free-text questions, answered by the underlying LLM. The third one further enriches the user interface with question-answering support by proposing context-dependent questions that the user can select to receive extra information about the previously explored content. These questions are planned to retrieve information the CH guide cannot provide.

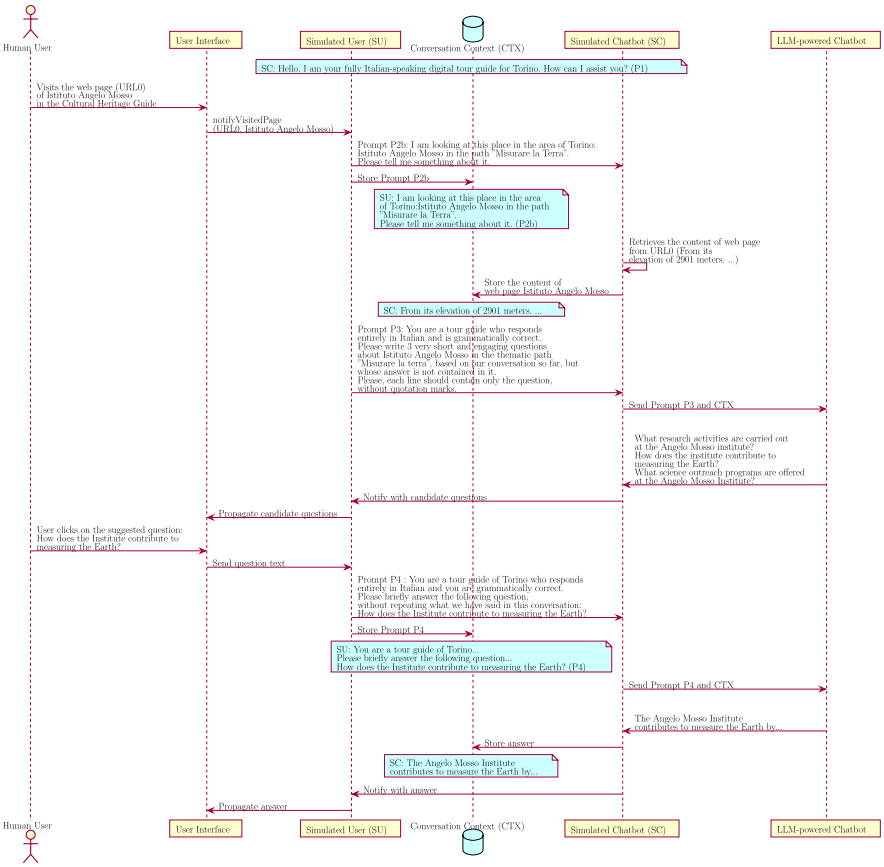


**Fig. 5** Sample interaction based on Fig. 3. The light blue notes represent the content of CTX. URL0 denotes the URL of Istituto Angelo Mosso's webpage in the Triangolazioni CH guide: https://www.triangolazioni.unito.it/app/percorso/114

The findings we collected through (user experience, and performance) questionnaires, and log data show that integrating an LLM-powered chatbot with the user's navigation in the website may greatly benefit the user experience, especially when coupled with the context-dependent suggestion of questions. The people with a low CEI-II (curiosity and exploration) score particularly appreciate this support because it enhances information exploration while providing inspiration to interact with the system.

## Sample interaction with the `RecChatbox` User Interface

See Fig. 5.

**Acknowledgements** The described research has been funded by the University of Turin. It builds on the work done in the Triangolazioni Project (https://www.triangolazioni.unito.it/) of the University's Public Engagement Lab. We are grateful to the anonymous reviewers who helped us improve the article with their insightful comments.

**Author Contributions** Angelo Geninatti Cossatin: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing - original draft, Writing - review and editing. Noemi Mauro: Conceptualization, Data curation, Investigation, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Validation, Writing - original draft, Writing - review and editing. Liliana Ardissono: Investigation, Software, Validation. Fabio Ferrero: Funding acquisition, Supervision, Writing - original draft, Writing - review and editing.

**Funding** Open access funding provided by Università degli Studi di Torino within the CRUI-CARE Agreement. The work has been funded by the University of Turin.

**Availability of data and materials** Code and data are available on https://anonymous.4open.science/r/tell-me-more-A4E5/.

## Declarations

**Conflict of interest** The authors have no conflict of interest to declare that are relevant to the content of this article.

**Ethics approval** The project was approved by the research ethics committee of the University of Turin (protocol number 0421424).

## References


Ardimento P, Bernardi ML, Cimitile M (2024) Teaching UML using a RAG-based LLM. In: 2024 International Joint Conference on Neural Networks (IJCNN), pp 1–8, https://doi.org/10.1109/IJCNN60899.2024.10651492

Ardissono L, Petrelli D, Kuflik T (2012) Personalization in cultural heritage: the road travelled and the one ahead. User Model User-Adapt Interact 22(1–2):73–99

Bekele MK, Pierdicca R, Frontoni E et al (2018) A survey of augmented, virtual, and mixed reality for cultural heritage. J Comput Cult Herit 11(2):1–36. https://doi.org/10.1145/3145534

Blythe M, Reid J, Wright PC et al (2006) Interdisciplinary criticism: analysing the experience of Riot! a location-sensitive digital narrative. Behav Inf Technol 25(2):127–139. https://doi.org/10.1080/01449290500331131

Bonis B, Stamos J, Vosinakis S et al (2009) A platform for virtual museums with personalized content. Multim Tools Appl 42(2):139–159. https://doi.org/10.1007/s11042-008-0231-2

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3:77–101. https://doi.org/10.1191/1478088706qp063oa

Braunhofer M, Ricci F (2016) Contextual information elicitation in travel recommender systems. In: Inversini A, Schegg R (eds) Information and communication technologies in tourism 2016. Springer International Publishing, Cham, pp 579–592

Brooke J (1996) SUS: a quick and dirty usability scale. Usability evaluation in industry 189. Taylor & Francis, New York

Casillo M, Clarizia F, D'Aniello G et al (2020) Chat-bot: a cultural heritage aware teller-bot for supporting touristic experiences. Pattern Recogn Lett 131:234–243. https://doi.org/10.1016/j.patrec.2020.01.003

Casillo M, De Santo M, Mosca R et al (2022) An ontology-based chatbot to enhance experiential learning in a Cultural Heritage scenario. Front Artif Intell 5:808281. https://doi.org/10.3389/frai.2022.808281

Chalmers A, Parkins J, Webb M et al (2021) Realistic humans in virtual cultural heritage. In: Shehade M, Stylianou-Lambert T (eds) Emerging technologies and the digital transformation of museums and heritage sites. Springer International Publishing, Cham, pp 156–165

Cheverst K, Davies N, Mitchell K, et al (2000) Providing tailored (context-aware) information to city visitors. In: Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. AH '00, Springer, London, UK, p 73–85,https://doi.org/10.1007/3-540-44595-1_8

De Carolis B, Gena C, Kuflik T et al (2018) Special issue on advanced interfaces for cultural heritage. Int J Hum-Comput Stud 114:1–2. https://doi.org/10.1016/j.ijhcs.2018.02.007

Dogru T, Line N, Mody M et al (2023) Generative Artificial Intelligence in the hospitality and tourism industry: developing a framework for future research. J Hosp Tour Res. https://doi.org/10.1177/00472875231212996

Dror R, Hutchinson D, Jones M, et al (2024) The curator's helper. In: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization. Association for Computing Machinery, New York, NY, USA, UMAP Adjunct '24, p 496–504, https://doi.org/10.1145/3631700.3664905,

Europeana (2024) Europeana collections. www.europeana.eu/portal/it

Explosion AI (2017) spaCy - industrial Natural Language Processing in python. https://spacy.io/

Faralli S, Lenzi A, Velardi P (2022) A large interlinked knowledge graph of the Italian cultural heritage. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 6280–6289, URL https://aclanthology.org/2022.lrec-1.675

Fenu C, Pittarello F (2018) Svevo tour: the design and the experimentation of an augmented reality application for engaging visitors of a literary museum. Int J Hum-Comput Stud 114:20–35. https://doi.org/10.1016/j.ijhcs.2018.01.009

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:2203.05794

Hargood C, Hunt V, Weal MJ, et al (2016) Patterns of sculptural hypertext in location based narratives. Association for Computing Machinery, New York, NY, USA, HT '16, p 61–70, https://doi.org/10.1145/2914586.2914595

Jurafsky D, Martin J (2024) URL https://web.stanford.edu/~jurafsky/slp3/

Kashdan T, Gallagher M, Silvia P et al (2009) The curiosity and exploration inventory-II: development, factor structure, and psychometrics. J Res Pers 43:987–998. https://doi.org/10.1016/j.jrp.2009.04.011

Kim H, Matuszka T, Kim JI et al (2017) Ontology-based mobile augmented reality in cultural heritage sites: information modeling and user study. Multimed Tools Appl 76(24):26001–26029. https://doi.org/10.1007/s11042-017-4868-6

Kim JH, Kim J, Park J et al (2023) When ChatGPT gives incorrect answers: the impact of inaccurate information by generative AI on tourism decision-making. J Travel Res. https://doi.org/10.1177/00472875231212996

Kouretsis A, Varlamis I, Limniati L et al (2022) Mapping art to a knowledge graph: using data for exploring the relations among visual objects in renaissance art. Future Internet 14(7):206. https://doi.org/10.3390/fi14070206

Kuflik T, Stock O, Zancanaro M et al (2011) A visitor's guide in an "active museum'': presentations, communications, and reflection. ACM J Comput Cult Herit 3(3):175–209. https://doi.org/10.1145/1921614.1921618

Lombardi M, Pascale F, Santaniello D (2019) An application for Cultural Heritage using a chatbot. In: 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS), pp 1–5, https://doi.org/10.1109/CAIS.2019.8769525

MacCallum R, Zhang S, Preacher K et al (2002) On the practice of dichotomizing quantitative variables. Psychol Methods 7:19–40. https://doi.org/10.1037/1082-989X.7.1.19

Machidon OM, Duguleana M, Carrozzino M (2018) Virtual humans in cultural heritage ict applications: a review. J Cult Herit 33:249–260. https://doi.org/10.1016/j.culher.2018.01.007

Machidon OM, Tavčar A, Gams M et al (2020) CulturalERICA: a conversational agent improving the exploration of European cultural heritage. J Cult Herit 41:152–165. https://doi.org/10.1016/j.culher.2019.07.010

Mauro N, Geninatti Cossatin A, Cravero E, et al (2022a) Exploring semantically interlaced cultural heritage narratives. In: Proceedings of the 33rd ACM Conference on Hypertext and Social Media (HYPERTEXT'22). ACM, Barcelona, Spain, p 192-197, https://doi.org/10.1145/3511095.3536366

Mauro N, Geninatti Cossatin A, Cravero E, et al (2022b) A mobile guide to explore interconnections between science, art and territory. In: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct). ACM, Barcelona, Spain, p 397–401, https://doi.org/10.1145/3511047.3537649

McNatt W, Bieman J (2001) Coupling of design patterns: common practices and their benefits. In: 25th Annual International Computer Software and Applications Conference. COMPSAC 2001, pp 574–579, https://doi.org/10.1109/CMPSAC.2001.960670

Mich L, Garigliano R (2023) ChatGPT for e-tourism: a technological perspective. Inf Technol Tour 25(1):1–12. https://doi.org/10.1007/s40558-023-00248-x

Michalakis K, Christodoulou Y, Caridakis G et al (2021) A context-aware middleware for context modeling and reasoning: a case-study in smart cultural spaces. Appl Sci 11(13):5770. https://doi.org/10.3390/app11135770

Millard DE, Hargood C, Jewell MO, et al (2013) Canyons, deltas and plains: Towards a unified sculptural model of location-based hypertext. In: Proceedings of the 24th ACM Conference on Hypertext and Social Media. Association for Computing Machinery, New York, NY, USA, HT '13, pp. 109–118, https://doi.org/10.1145/2481492.2481504

Noh YG, Hong JH (2021) Designing reenacted chatbots to enhance museum experience. Appl Sci 11(16):7420. https://doi.org/10.3390/app11167420

Perplexity Team (2024) Introducing PPLX online LLMs. https://blog.perplexity.ai/blog/introducing-pplx-online-llms

Ribreau F, Zavackiy R (2023) Node Language Detect. https://www.npmjs.com/package/languagedetect/

Rinaldi AM, Russo C, Tommasino C (2022) An approach based on linked open data and augmented reality for cultural heritage content-based information retrieval. In: Gervasi O, Murgante B, Hendrix EMT et al (eds) Computational Science and Its Applications - ICCSA 2022. Springer International Publishing, Cham, pp 99–112

Spennemann DHR (2023) ChatGPT and the generation of digitally born knowledge: how does a generative ai language model interpret cultural heritage values? Knowledge 3(3):480–512. https://doi.org/10.3390/knowledge3030032

Sperlí G (2021) A cultural heritage framework using a deep learning based chatbot for supporting tourist journey. Expert Syst Appl 183:115277. https://doi.org/10.1016/j.eswa.2021.115277

Sylaiou S, Fidas C (2022) Virtual humans in museums and cultural heritage sites. Appl Sci 12(19):9913. https://doi.org/10.3390/app12199913

Tam D, Mascarenhas A, Zhang S, et al (2023) Evaluating the factual consistency of large language models through news summarization. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, Toronto, Canada, pp 5220–5255, https://doi.org/10.18653/v1/2023.findings-acl.322,

Touvron H, Martin L, Stone K, et al (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288

Trichopoulos G, Konstantakis M, Alexandridis G et al (2023) Large language models as recommendation systems in museums. Electronics 12(18):3829. https://doi.org/10.3390/electronics12183829

Trichopoulos G, Konstantakis M, Caridakis G et al (2023) Crafting a museum guide using chatgpt4. Big Data Cogn Comput 7(3):148. https://doi.org/10.3390/bdcc7030148

Tsepapadakis M, Gavalas D (2023) Are you talking to me? An audio augmented reality conversational guide for cultural heritage. Pervasive Mobile Comput 92:101797. https://doi.org/10.1016/j.pmcj.2023.101797

Tzouganatou A (2018) Can heritage bots thrive? toward future engagement in cultural heritage. Adv Archaeol Pract 6(4):377–383. https://doi.org/10.1017/aap.2018.32

Varitimiadis S, Kotis K, Pittou D et al (2021) Graph-based conversational AI: towards a distributed and collaborative multi-chatbot approach for museums. Appl Sci 11(19):9160. https://doi.org/10.3390/app11199160

Wang Y, Stash N, Aroyo L et al (2008) Recommendations based on semantically enriched museum collections. J Web Semant 6(4):283–290. https://doi.org/10.1016/j.websem.2008.09.002

Wojton HM, Porter D, Lane ST et al (2020) Initial validation of the trust of automated systems test (toast). J Soc Psychol 160(6):735–750. https://doi.org/10.1080/00224545.2020.1749020

Yasaka K, Kanzawa J, Kanemaru N et al (2024) Fine-tuned large language model for extracting patients on pretreatment for lung cancer from a picture archiving and communication system based on radiological reports. J Imaging Inform Med. https://doi.org/10.1007/s10278-024-01186-8



## Authors and Affiliations

**Angelo Geninatti Cossatin[1] · Noemi Mauro[1] · Fabio Ferrero[1] · Liliana Ardissono[1]**

✉ Noemi Mauro
noemi.mauro@unito.it

Angelo Geninatti Cossatin
angelo.geninatticossatin@unito.it

Fabio Ferrero
fab.ferrero@unito.it

Liliana Ardissono
liliana.ardissono@unito.it

[1] Department of Computer Science, University of Turin, Corso Svizzera 185, 10145 Torino, Italy