Reliability and Domain Limitations of Intelligent Virtual Assistants in Digital Heritage Systems

Milen Kraev[0009-0009-8973-7765]

Institute of Mathematics and Informatics, Bulgarian Academy of Sciences,
8 Acad. Georgi Bonchev Street, 1113 Sofia, Bulgaria
kraevmilen@gmail.com

Abstract. After years of work on cultural heritage and scientific digitization, the world accumulated a vast amount of data and information. Although digital data is far more accessible than the physical one, even in an unstructured way, better interaction is needed by the non-professional users. At the same time AI-related systems gain widespread popularity in all kinds of platforms, easing accessibility. Digital assistants and chatbots are implemented in existing applications and databases. Cultural heritage areas are not falling behind. Plenty of examples of using such systems in the field are already available. Implementation of AI in cultural heritage brings unique challenges along with benefits. The accuracy of the models is highly uncertain, caused by the specifics of the cultural heritage data. Several approaches to overcome the challenges such as parameter-efficient fine-tuning and Retrieval-Augmented Generation are examined.

Keywords: AI Assistant, Cultural Heritage Assistants, RAG, PEFT, LLM Reliability.

1 Framing AI Reliability in Digital Heritage

Digital heritage platforms are increasingly incorporating AI-driven interfaces to facilitate access to cultural heritage knowledge. The process includes a transition in the user interaction from conventional search and browsing to conversational and interpretive systems. Museums, archives, and large heritage sites like Pompeii Archaeological Park, Luigi Einaudi Foundation, National Museum of Ravenna are starting to use AI-based tools that help people find, understand, and interact with digitized collections. This change makes AI systems, especially chatbots and conversational agents, act as middlemen between users and heritage data (Nafis et al., 2022).

Even though artificial intelligence is making quick progress, reliability for the AI-generated answers in digital heritage is still questionable. Current evaluations consider general performance metrics over the relevance, historical accuracy, and contextual suitability of responses to heritage-specific inquiries. This paper shows that the reliability of AI in digital heritage should be perceived as a coupled system problem. The

Digital Presentation and Preservation of Cultural and Scientific Heritage. Volume 16, pp. 221-230, 2026.
DOI: 10.55630/dipp.2026.16.18, ISSN 1314-4006, eISSN 2535-0366.

main issues that are considered are answer relevance, domain generalization, and computational limitations.

2 State of the Art in AI for Digital Heritage

Recent studies on artificial intelligence in digital heritage propose new methods in digitization, semantic indexing, multimodal retrieval, and interactive user interfaces. AI and augmented reality methods have been successfully used for tasks like recognizing artifacts, adding metadata, rebuilding 3D models, and managing large collections (Rachabatuni et al.,2024; Tatić et al., 2025). This has made it easier to process and explore cultural heritage data and information. More and more, conversational interfaces and smart assistants are being marketed as useful tools for getting people involved, teaching, and exploring large collections of heritage items (Sathiyabamavathy & Anju, 2024; Casillo et al., 2022; Caramiaux, 2023).

There are already plenty of working examples in the field. (Natale et al., 2025) present a customized generative-AI chatbot developed to simulate conversations with the Italian economist and political figure Luigi Einaudi. The system is trained on Einaudi's writings and allows users to interact with the historical figure through a conversational interface in a specific historical context. The authors demonstrate that these chatbots can improve public engagement with cultural heritage. In addition, they expose factual inaccuracy and source transparency issues. The study emphasizes the opportunities and risks linked to the implementation of generative AI as an interpretive interface for cultural heritage knowledge. (Casillo et al., 2020) propose a context-aware conversational system designed to support cultural heritage tourism. The chatbot recommends personalized routes and information about heritage sites. It uses pattern recognition and contextual analysis to change its responses based on the user's profile, location, and situation. This provides possibilities for an interactive conversation that is like the one of a digital tour guide. The system was tested at several archaeological sites in the Campania region of Italy, such as Pompeii, Paestum, and Herculaneum. This showed how conversational interfaces can be used explaining heritage sites containing a lot of historical information. (Fabbri et al., 2023) examine the use of AI-based chatbots as storytelling tools for enhancing visitor engagement in museums. Their research centers on the National Museum of Ravenna. In this case a chatbot assists visitors in understanding extensive and diverse collections. The platform is trying to suggest artworks and tailor the narrative to match the visitor interests. The authors conclude that conversational AI can enhance personalized cultural heritage experiences. The approach includes aligning visitor data with interactive dialogue, which enables museums to customize information and narratives for each user. They also stress how hard it is to make sure that AI-driven stories fit with the museum's curatorial structure and institutional identity.

The studies discussed here, and other such papers in this space, demonstrate very good examples of using AI in cultural heritage. At the same time concerns remain about the reliability and historical accuracy of generated answers. The challenges of adapting AI models for highly specialized cultural and historical terrains are still present and are frequently encountered across the examples. Ensuring that AI-generated narratives stay in line with curatorial interpretations is still a challenge. These issues indicate that providing relevant, trustworthy and contextually appropriate responses continues to be a key challenge for conversational AI in digital heritage systems.

3 Answer Relevance and Reliability in Digital Heritage AI

The relevance and trustworthiness of responses returned by a chatbot are a key concern in the use of AI within digital heritage systems. Modern large language models (LLMs) are indeed capable of producing fluent and grammatically correct text. However, they fail to do this with an adherence to verified cultural heritage material. These can cause hallucinations whereby the model produces plausible but untrue information about artifacts, historical events, or cultural contexts. The LLMs perform even worse when they are asked to list the reference.

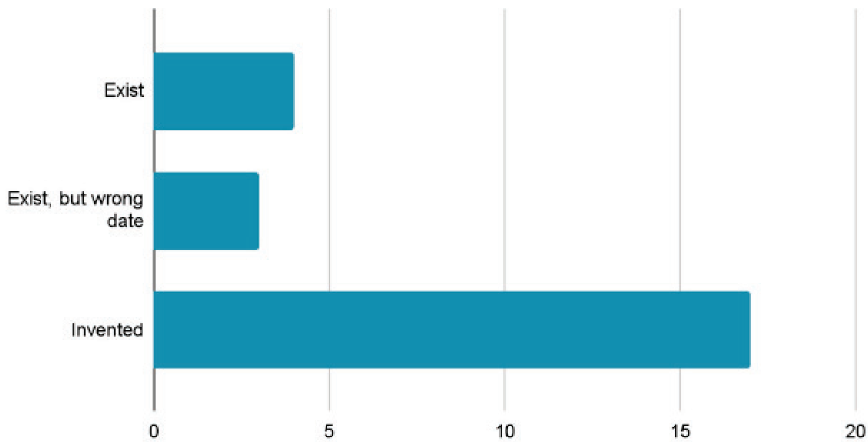

Fig. 1. ChatGPT reference creation experiment (Bouchachi et al., 2025).

Figure 1 presents an experiment where ChatGPT is asked to provide sources about the sources it uses to write a particular text. In majority of cases the reference does not exist at all. Such inaccurate information, whether on the source, dates, stylistic attributions, provenance or conflation of similar figures in different historical settings can deeply skew our understandings of cultural objects and their historical narratives reflected within digital heritage environments. One core technical reason behind this problem is the training paradigm of LLMs, which are usually optimized using probabilistic sequence prediction objectives like cross-entropy loss. A Transformer-based model generates a response drop-by-drop, token by token. In this way it predicts the most likely next word in terms of commonly occurring sequences from its training datasets but not checking facts against authoritative databases (Agbesi et al., 2023). When it answers expert cultural heritage queries, the model may base its inference on general information patterns rather than specific archival knowledge.

To overcome these constraints, various technical methodologies have been investigated. Retrieval-Augmented Generation (RAG) is a method which is mentioned in a lot of papers for the subject. It combines a language model with an outside information retrieval system. In a standard RAG architecture, embedding models like Sentence-BERT first turn the user's query into a vector representation. The system then uses algorithms like k-nearest neighbor search (k-NN) and vector indexes like FAISS to find documents that are semantically similar in a heritage database or digital archive (Danopoulos et al., 2019). Before generating a response, the system adds the retrieved documents to the model's context window. This lets the system give answers based on specific archival records or metadata descriptions. This method does help with factual grounding, but how well it works depends a lot on how good and well-structured the digital collections are. Another technical issue comes up because heritage datasets are made up of different types of data. Cultural heritage repositories frequently house a combination of structured metadata (e.g., museum catalog records adhering to standards such as CIDOC CRM), semi-structured documentation, and unstructured textual descriptions. Combining these sources, differences in metadata schemas, language, and incomplete records are still a technical burden. This can make it harder to find what a user is looking for and could give them answers that aren't relevant or the ones that are only partially correct. In addition to the technical part there is also an informational difficulty. The questions asked by users regarding cultural heritage are often difficult to answer. They need a lot of context and multi-level reasoning to be answered accordingly. For example, time period, places and author style must be considered. Standard retrieval or ranking algorithms might have trouble understanding these complicated connections. BM25 or basic embedding similarity measures are not always enough to overcome these challenges (Robertson & Zaragoza, 2009).

4 Domain Generalization and Computational Constraints

The issues with chatbot answers not being reliable are closely related to how well the model can generalize and the limited computing power that cultural heritage institutions have. Most conversational AI systems use LLMs that were trained on huge amounts of text from the web using self-supervised learning goals. These datasets mostly show how people use language and what they know in general (Tian & Jiang, 2024). This training is not enough for more specific purposes. Cultural heritage information depends on specialized vocabularies (sometimes obsolete), historical terms, and structured classification systems. This leads to a clear difference in how the training data and the domain-specific data sets used in digital heritage collections are structured. This mismatch causes generalization errors. They could be mixing up different artistic styles or getting confused about historical periods or places (Fu et al., 2025).

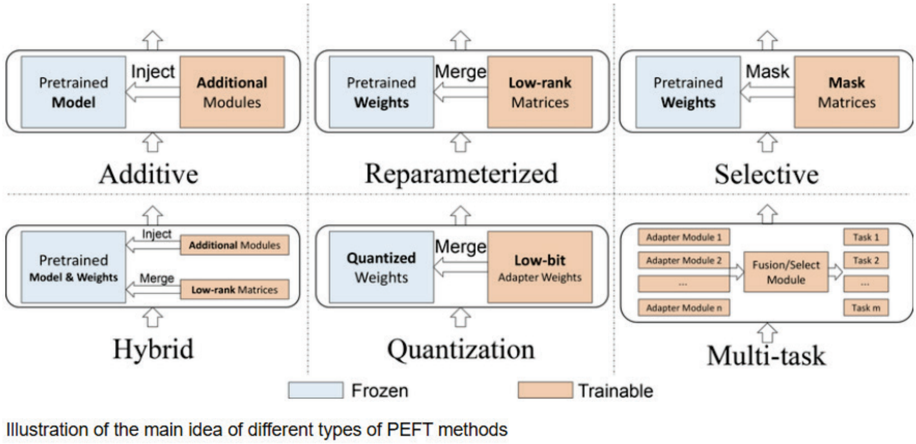

Fig. 2. Parameter-Efficient Fine-Tuning Process (Wang et al., 2025).

The limitations which have been already mentioned require additional training to the LLMs., Fine-tuning, instruction tuning, or parameter-efficient methods like Low-Rank Adaptation and adapter layers are needed in order to achieve good domain reasoning. Figure 2 shows how frozen pretrained weights (blue) and trainable components (orange) interact with each other. This helps to group the main Parameter-Efficient Fine-Tuning (PEFT) methods. The taxonomy separates Additive methods, which add new modules, Reparameterized techniques like LoRA, which use low-rank matrices, and Selective approaches, which mask certain existing weights for updates. It also talks about memory-efficient quantization strategies, hybrid models that combine different techniques, and multi-task frameworks that use fusion modules to manage different adapters. In the end, the diagram shows how PEFT cuts down on computational overhead by focusing training on only a small number of the model's parameters.

These methods help a pretrained model to add knowledge from a specific domain without the need to retrain the whole architecture. However, effective domain adaptation usually needs carefully chosen data sets and a lot of computing power. Cultural heritage datasets are often small, spread out across different institutions. In addition, they are usually in more than one language, which makes them hard to use directly for training large models. Structured ontologies and knowledge graphs are frequently used to add semantic richness of heritage knowledge. It is not easy to add these kinds of structured data to a standard LLM. The computational needs of modern AI models are another problem. A lot of energy and access to GPUs is needed. It takes a huge amount of specialized hardware to train or fine-tune transformer-based models with billions of parameters (Sinha & Lee, 2024). The museums, archives, and smaller research groups that preserve cultural heritage don't have the resources needed for such training. Even running pretrain models requires a lot of memory, vector databases, and computational resources. Organizations frequently rely on externally hosted APIs or cloud-based AI services to overcome these issues. They are giving up their control over the model's behavior and constrain the possibility for domain-specific customization. General-purpose models do not perform well with heritage-specific classifications and historical research. Systems that are tailored to a specific field or enhancing the first ones is the solution (Peng et al., 2025). Even with a good model, there will be production problems such as model updates, dataset versioning, and reproducibility, all of which need ongoing technical knowledge.

5 Reliability as a Systemic Constraint in Digital Heritage AI

AI reliability in digital heritage is not merely a question of improving model accuracy. It is a systemic design challenge. The interdependence of answer relevance, domain generalization, and computational constraints make it hard to solve problem. The institutions are pushed toward the adoption of general-purpose models due to resource limitation. This leads to poor generalization in the heritage domains and produces unreliable answers (Hamm & Klesel, 2021). These unreliable outputs then shape user perception and engagement with cultural heritage knowledge. In the end the historical narratives are distorted or oversimplified. Reframing AI deployment in digital heritage is an infrastructural issue rather than a purely technical optimization problem. Evaluative frameworks must move beyond generic benchmarks to incorporate heritage-specific notions of relevance, contextual adequacy, and grounding in curated sources. Improvements in model architecture alone are unlikely to resolve the fundamental mismatch between AI systems and the knowledge structures of cultural heritage. The solution must contain middleware such as RAG algorithm.

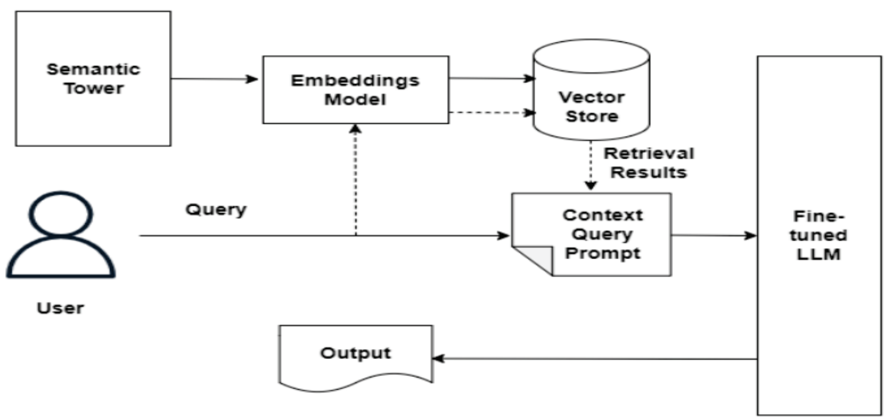

Fig. 3. RAG system architecture (Akl, 2024).

Figure 3 illustrates a RAG framework. This system uses external, authoritative data sources to get around the problems that come with static large language models. The pipeline starts by using a semantic layer and an embedding model to turn raw data into high-dimensional numerical vectors. These vectors are stored in a special vector database (like Pinecone, Milvus, or Weaviate) that uses proximity algorithms like Cosine Similarity or Euclidean distance to make it easy to find them. When a user asks a question, the system creates a query embedding that allows for a semantic search by pulling the most relevant context from the vector store. This information is then combined with the user's input to make a structured prompt. This gives the fine-tuned LLM a solid factual basis for its generation process. This architecture makes it less likely that the model will hallucinate and makes sure that the output is based on the domain-specific context that was given, not just the model's random internal parameters.

RAG is a promising way to make AI-generated answers in cultural heritage systems more reliable and relevant to the situation. When dealing with specialized heritage knowledge, traditional large language models make responses based on patterns they learned during training. This can lead to information that is old, incomplete, or wrong in terms of history. RAG gets around this problem by combining generative language models with ways to get information from outside sources. The system doesn't just use its own model knowledge. It also gets documents, metadata, or archival descriptions from curated heritage databases and uses them to help it generate responses (Panda, 2025). In the larger picture of cultural heritage, this lets AI systems use information from digital archives, museum collections, historical records, and academic sources to answer questions from users. As seen so far RAG architecture is very useful for digital heritage infrastructures. Heritage data is often spread across different types of repositories and cultural heritage platforms. RAG-based systems can give answers that are more in line with authoritative heritage sources. Getting information that is specific to the domain before generating an answer could be the solution to the domain data problem. This method makes things clearer and easier to understand. (Mallouhy et al., 2025).

6 Conclusions

This paper assesses AI reliability in digital heritage taking into consideration different examples already developed. A transition from pure performance metrics to a holistic comprehension of how chatbot operates is needed. Answer relevance, domain generalization, and computational limitations collectively influence user-facing results and are still unsolved problems. The main one is that AI-generated answers aren't always relevant or reliable. This is because the training data used for general purposes doesn't always match the knowledge domains used for heritage. These discrepancies are made worse by a lack of resources.

Future research must prioritize evaluation methodologies rooted in heritage specifics. New technical strategies that harmonize domain specificity with computational feasibility could be developed. The solution for more reliable AI usage in digital heritage lays in model design, evaluation methods, and infrastructure work together in a coherent system.

Acknowledgements

This research is supported partly by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, https://clada-bg.eu/bg/, Grant number D01-97/26.06.2025, financed by the Bulgarian Ministry of Education and Science, Funding Procedure: The National Science Infrastructure Roadmap 2020-2027 (NSIF).

References

Agbesi, V., Chen, W., Yussif, S., Hossin, M., Chiagoziem, U., Kuadey, N., Agbesi, C., Samee, N., Jamjoom, M., & Al-Antari, M. (2023). Pre-trained transformer-based models for text classification using low-resourced Ewe language. *Systems*, *12*(1), Article 1. https://doi.org/10.3390/systems12010001

Akl, H. (2024). *DSTI at LLMs4OL 2024 Task A: Intrinsic versus extrinsic knowledge for type classification*. arXiv. https://doi.org/10.48550/arXiv.2408.14236

Bouchachi, M., Jiménez-Delgado, A., De Gracia-Soriano, P., & Nemroudi, R. (2025). Architectural heritage and artificial intelligence: Diagnosis and solutions proposed by ChatGPT for Algerian historical monuments. *Heritage*, *8*(4), Article 139. https://doi.org/10.3390/heritage8040139

Caramiaux, B. (2023). AI with museums and cultural heritage. In *AI in Museums* (pp. 117–130). transcript Verlag. https://doi.org/10.14361/9783839467107

Casillo, M., Clarizia, F., D'Aniello, G., De Santo, M., Lombardi, M., & Santaniello, D. (2020). CHAT-Bot: A cultural heritage aware teller-bot for supporting touristic experiences. *Pattern Recognition Letters*, *131*, 234–243. https://doi.org/10.1016/j.patrec.2020.01.003

Casillo, M., De Santo, M., Mosca, R., & Santaniello, D. (2022). An ontology-based chatbot to enhance experiential learning in a cultural heritage scenario. *Frontiers in Artificial Intelligence*, *5*, Article 808281. https://doi.org/10.3389/frai.2022.808281

Danopoulos, D., Kachris, C., & Soudris, D. (2019). Approximate similarity search with FAISS framework using FPGAs on the cloud. In *Embedded Computer Systems: Architectures, Modeling, and Simulation* (pp. 373–386). Springer. https://doi.org/10.1007/978-3-030-27562-4_27

Fabbri, F., Collina, F., & Barzaghi, S. (2023). *AI and chatbots as a storytelling tool to personalize the visitor experience: The case of National Museum of Ravenna* [Conference poster]. ExICE 2023: Extended Intelligence for Cultural Engagement, Bologna, Italy. Zenodo. https://doi.org/10.5281/zenodo.7973638

Fu, Q. Y., Dong, S. H., & Yuan, C. H. (2025). The current status and challenges of artificial intelligence in the digital preservation of cultural heritage. *Journal of Artificial Intelligence & Robotics*, *2*(1), Article 1018. https://doi.org/10.52768/3067-7947/1018

Hamm, P., & Klesel, M. (2021). Success factors for the adoption of artificial intelligence in organizations: A literature review. In *AMCIS 2021 Proceedings*. Association for Information Systems. https://aisel.aisnet.org/amcis2021/art_intel_sem_tech_intelligent_systems/art_intel_sem_tech_intelligent_systems/10

Mallouhy, R., Fallatah, K., & Alkhateeb, M. (2025). RAG-driven AI for architectural heritage management: Optimizing renovation and preservation of Saudi traditional identity. In *Dakam's spring 2025 architecture and urban design conferences proceedings* (pp. 131–141). Dakam Books. https://www.dakamconferences.org/_files/ugd/bac820_d171ab8323a24697b44b78c8a0e7129d.pdf

Nafis, F., Yahyaouy, A., & Aghoutane, B. (2022). Chatbots for cultural heritage: A real added value. In *Proceedings of the 2nd International Conference on Big Data, Modelling and Machine Learning (BML)* (Vol. 1, pp. 502–506). SciTePress. https://doi.org/10.5220/0010737700003101

Natale, S., Surace, B., Mensa, E., & Befera, L. (2025). ChatGPT for cultural heritage and the customization of generative AI: A talkthrough analysis of the Luigi Einaudi chatbot. *New Media & Society*. Advance online publication. https://doi.org/10.1177/14614448251384258

Panda, M. (2025). Agentic RAG: Redefining retrieval-augmented generation for adaptive intelligence. *International Research Journal of Engineering and Technology*, *12*(1), 731–739. https://www.irjet.net/volume12-issue01

Peng, Y., Heyn, H., & Horkoff, J. (2025). *Data challenges in AI systems and their solutions: A requirements and AI engineering systematic literature review and comparison* [Preprint]. Research Square. https://doi.org/10.21203/rs.3.rs-8077403/v1

Rachabatuni, P. K., Principi, F., Mazzanti, P., & Bertini, M. (2024). Context-aware chatbot using MLLMs for cultural heritage. In *Proceedings of the 15th ACM Multimedia Systems Conference* (pp. 459–463). https://doi.org/10.1145/3625468.3652193

Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Sathiyabamavathy, K., & Anju, K. (2024). Role of chatbots in cultural heritage tourism: An empirical study on ancient forts and palaces. *Journal of Heritage Management*, *9*, 9–28. https://doi.org/10.1177/24559296241253932

Sinha, S., & Lee, Y. (2024). Challenges with developing and deploying AI models and applications in industrial systems. *Discover Artificial Intelligence*, *4*, Article 55. https://doi.org/10.1007/s44163-024-00151-2

Tatić, D., Stanković, R., & Goynov, M. (2025). Usage of geospatial augmented reality for the representation of national heritage. *Digital Presentation and Preservation of Cultural and Scientific Heritage*, *15*, 99–106. https://doi.org/10.55630/dipp.2025.15.9

Tian, L., & Jiang, N. (2024). Research on detection methods for text generated by large language models based on multi-model ensemble. *Applied and Computational Engineering*, *106*, 59–67. https://doi.org/10.54254/2755-2721/106/20241331

Wang, L., Chen, S., Jiang, L., Pan, S., Cai, R., Yang, S., & Yang, F. (2025). Parameter-efficient fine-tuning in large language models: A survey of methodologies. *Artificial Intelligence Review*, *58*, Article 227. https://doi.org/10.1007/s10462-025-11236-4

Received: March 15, 2026
Reviewed: April 21, 2026
Finally Accepted: May 05, 2026