Using Ontologies for LLM Applications in Cultural Heritage

Rocco Loffredo[1] and Massimo De Santo[1]

[1] *DIIN, University of Salerno, Via Giovanni Paolo II 132 84084 Fisciano, Italy*

Abstract

Applying Large Language models (LLMs) offers the potential for transformative change in cultural heritage. This short paper is based on ongoing doctoral research. It examines innovative methodologies for enhancing the accessibility, comprehension, and preservation of cultural heritage by utilizing AI technologies such as LLM. This research aims to improve AI-generated responses' contextual precision and dependability by employing sophisticated knowledge representations, such as ontologies. The approach promises to overcome the challenges associated with data complexity and information retrieval, thereby creating new avenues for heritage documentation, education, and public engagement.

Keywords

large language model, ontology, cultural heritage, retrieval augmented generation

1. Introduction

Cultural heritage represents a series of milestones of human civilization, serving as a collective memory and identity. As technology advances, there is a growing need for innovative approaches to preserving, understanding, and sharing those resources.

Large Language Models (LLMs) and LLM enhancement offer a promising solution for addressing these challenges. However, the practical application of LLMs in cultural heritage domains requires a robust foundation of new external knowledge to specialize LLMs appropriately in that task. This is why, the potential for hallucinations and the necessity for reliable, unbiased and error-free knowledge resources represents a significant current challenge [1].

Among the various techniques for LLM specialization, RAG (Retrieval-Augmented Generation) method is proving particularly successful. It is based on associating the user's prompt with additional documentation that the model can consult to provide a more precise and contextualized answer within the given domain.

This short paper describes our first steps in defining a methodology for enhancing LLMs for cultural heritage applications using ontologies. Ontologies are formal representations of concepts and their relationships, and they can play a central role in this context by being able to provide a structured form of knowledge for LLMs to acquire more detailed information within a specific domain.

Section 2 presents an overview of existing research on the re-training of LLM, with particular attention to its current applications in specific contexts. Section 3 outlines a proposed methodology for enhancing LLM and discusses the competitive potential for its implementation. Section 4 provides details of the proposed method. Finally, Section 5 offers conclusions and a discussion of future steps.

2. Related Works

The scientific community has initiated to explore the applications of Large Language Models in cultural heritage domains.

3rd Workshop on Artificial Intelligence for Cultural Heritage (AI4CH 2024, https://ai4ch.di.unito.it/), co-located with the 23nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024). 26-28 November 2024, Bolzano, Italy

rloffredo@unisa.it (R. Loffredo); desanto@unisa.it (M. De Santo)

The potential of well-known models such as ChatGPT to transform the visitor experience in cultural settings is discussed in [2] and [3], while in [4], ChatGPT's contribution to e-learning is highlighted with a specific focus on the analysis and interpretation of lyrics and poems.
In particular, the utilization of enhanced LLM models through ontologies about cultural heritage domains represents an area of promising research with considerable potential to become a big challenge and one of the future directions of ontologies' possibilities of use [5].

3. Proposed Approach

3.1. The Hallucinations Issue

LLMs, such as GPT-4 or LLaMA 3, are potent tools capable of processing vast amounts of information and generating responses across various topics. However, their generalist nature has some significant drawbacks, which can lead to imprecise or incorrect answers and fabricated statements in specific contexts. Those events in this field are known as "hallucinations," they are pretty standard if the LLM lacks sufficient information to provide accurate responses.

One of the leading causes of hallucinations is the statistical nature of generating the next word by evaluating the likelihood of different possible words based on the context provided by the previous words. The model will always answer in this way, so hallucinations can happen.

3.2. Retrieval-Augmented Generation

Enhancing LLMs with reliable external and new knowledge about well-defined new contexts is essential to making them more focused, thereby mitigating hallucination risk. A promising technique to achieve this is the use of the RAG (Retrieval-Augmented Generation) method [6] shown in Figure 1.

This technique entails the utilization of a Retrieval Model in conjunction with the LLM. In response to the user's query, the Retrieval Model identifies the documents within an external knowledge base most closely aligned with the request through a semantic similarity calculation. In this manner, the LLM will receive both the user's query, modified as necessary, and the documents the model can reference to provide a more precise and coherent response within the specified context. In this way, without changing the model's internal parameters, it is possible to obtain more context-related, reliable, and up-to-date responses.

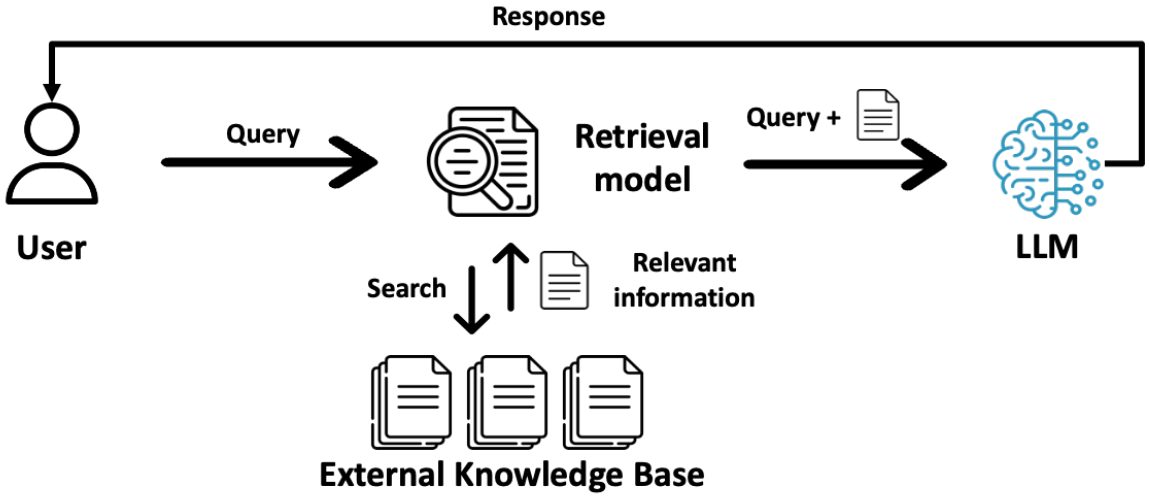

Figure 1: Summary diagram of how the RAG method works

3.3. LLMs Enhancement with Ontologies

Unlike traditional documents, which are often unstructured, ontologies utilize explicit and formalized relationships, such as subject-predicate-object triples, which clarify and formalize the links between concepts. This approach is particularly advantageous for RAG, as it allows the retrieval of relevant text fragments and structured knowledge that can be integrated with the model's generative capabilities, thereby improving the precision and coherence of responses.

Ontologies provide a formal structure for representing relationships between concepts, making it easier to navigate a vast set of information. Compared to standard documents, ontologies offer an organized and hierarchical view of knowledge based on semantic links between entities, facilitating the retrieval of relevant information. In an RAG context, ontologies can serve as structured knowledge bases upon which retrieval methods are applied, ensuring that retrieved information is relevant and accurately connected.

Enhancing LLM systems to comprehend the specifics of an artifact or a heritage site is possible. This enables the provision of a highly accurate research experience for the end user, with the ability to answer complex questions and provide detailed information very quickly.

3.3.1. Possible Advantages

In this way, LLMs can be employed to analyze substantial quantities of textual and visual data [4], identifying patterns and anomalies that may indicate conservation issues or risks to cultural property. Furthermore, they can facilitate the development of innovative heritage documentation and enhancement tools, such as immersive virtual realities or augmented reality applications. Moreover, the capacity of these models to process natural language makes it possible to make cultural heritage information accessible to a broader audience, overcoming language barriers [7], adapting the responses to the user's specific context or knowledge level and simplifying the consultation of complex databases [8].

By leveraging ontologies, RAG can provide LLMs with a structured understanding of the domain, enabling them to comprehend complex queries better, provide more accurate and relevant responses to real-world facts and prevent the generation of hallucinations or misleading information [9].

From a competitive standpoint, in contrast to conventional sources of cultural heritage that employ disparate formats, LLMs can leverage ontologies to integrate and normalize diverse datasets, facilitating a unified approach across multiple resources. This would enable more straightforward and consistent general access to content.

4. Methodology

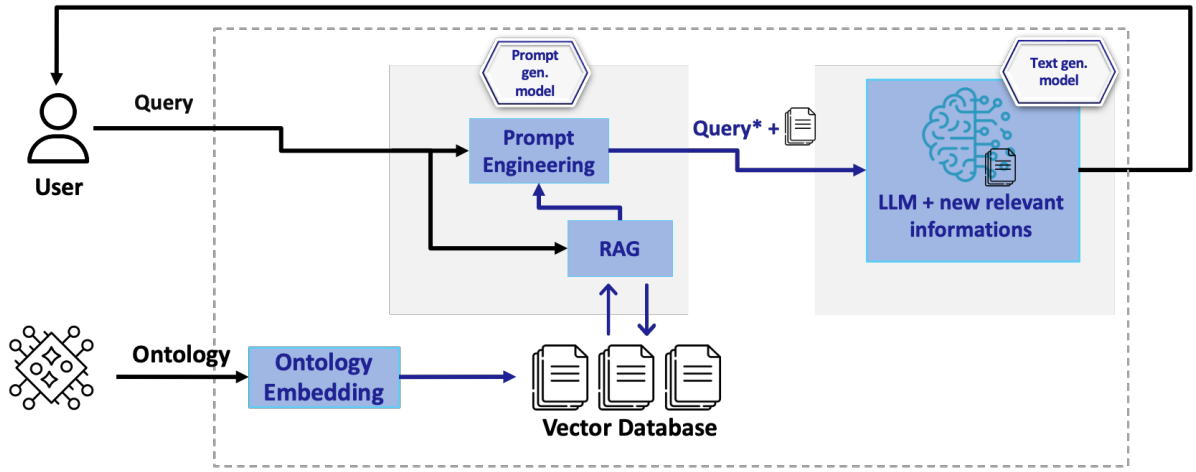

Figure 2: Framework of the proposed methodology

The proposed methodology is designed to enhance the performance of LLM systems using ontology-driven information retrieval techniques within the cultural heritage domain. The use of structured representations of knowledge enables LLMs to gain a deeper understanding of the context of cultural artefacts, historical events and their relationships.

The approach illustrated in Figure 2 involves embedding ontologies, an approach that is being documented in the literature [10], [11]. Embedding these ontologies into vector databases enables LLMs to retrieve more relevant and accurate information in response to user queries.

The system comprises three main parts: the Ontology Embedding, the Prompt Generation Model and the Text Generation Model.

4.1. Ontology Embedding

The Ontology Embedding block is the sole component that does not require repetition during the session between the user and the model. Its responsibility is to convert the input ontology into a vectorized format that can be efficiently queried by the Prompt Generation Model.

This process represents the most critical aspect of the entire framework, as it involves transforming the ontology from an RDF format into a textual representation compatible with the embedding process. The challenge is to ensure that this conversion does not compromise the semantic relationships inherent in the ontology, which are fundamental to its structure and utility.

In particular, the proposed approach involves converting the subject-predicate-object triples within the RDF into a textual format, which is then loaded into a vector database using ChromaDB [12]. ChromaDB is an open-source vector database and it plays a crucial role in this embedding and retrieval process. It supports semantic search using algorithms based on cosine similarity. In this approach, when a query is issued, ChromaDB retrieves the most relevant ontology triples by computing the cosine similarity between the vector representation of the query and the pre-embedded vectors of the ontology. This ensures that even if the exact wording of the query does not match the text derived from the ontology, semantically similar triples will still be retrieved, maintaining the integrity of the ontology's semantic structure during the process. With this approach, semantic relationships are preserved during the conversion from RDF triples to a vector database. The effectiveness of this process depends on the quality of the embedding model, as it must capture the underlying meaning of the relationships within the ontology.

4.2. Prompt Generation Model

The Prompt Generation Model employs the RAG method and prompt engineering techniques [13] to repurpose the user's query and obtain the most accurate answer from the LLM.

Specifically, the Retrieval Model consults the vector database generated from the ontology, and prompt engineering techniques are used to propose a new prompt to the LLM based on the latest relevant information obtained from the database.

In particular, the new prompt could underline the audience to whom the LLM is chatting so it could adapt to the desired knowledge level or set the temperature.

4.3. Text Generation Model

The Text Generation Model is the pre-trained LLM itself. When the query from the Prompt Generation Model is obtained, it generates a response using its intrinsic capabilities and the new information retrieved from the vectorial database.

4.4. Preliminary Experimental Results

To reinforce the effectiveness of this methodology, preliminary tests were conducted to verify that the use of RDF triples instead of actual texts reduces the actual hallucination rate of LLM models. In particular, six works of art at the University of Salerno were taken as case studies.

Three questions were asked for each work to three different models:

- “Who is the author of the work {work name} located at the University of Salerno?”
- “When was the work {work name} located at the University of Salerno created or inaugurated?”
- “What are the materials used and the symbolic or conceptual meaning, if indicated, of the work {work name} located at the University of Salerno?”

The questions were asked in three different situations: in the first situation the models were not given any additional material to consult, in the second situation a document with a description of all the works of art was provided each time, in the third situation a text file was provided with RDF triples inside indicating the same information as the previous text file.

Precision, Recall and F1-Score of the responses were evaluated and compared. The results can be seen in Figure 3.

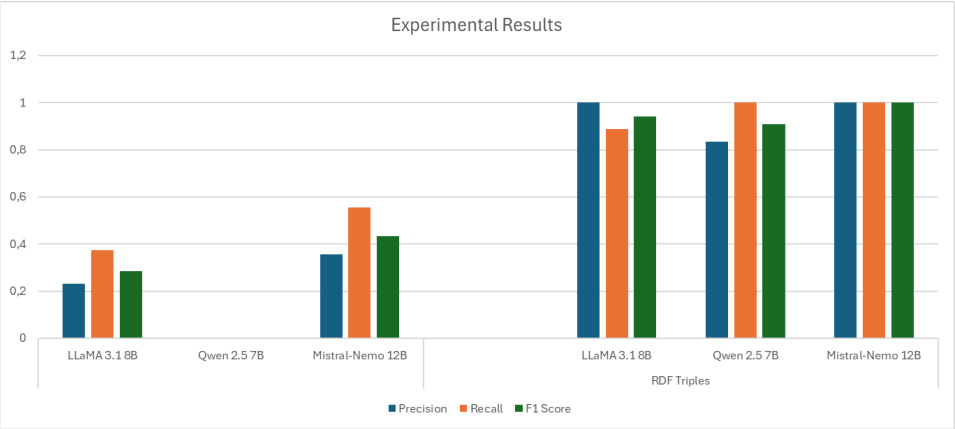

Figure 3: Preliminary Experimental Results

The responses were evaluated as follows:

- True Positive for the answers considered to be correct;
- False Positive for the answers with hallucinations;
- False Negative for the answers in which the model retrieves no information about the subject despite having been provided;

These results show that, at least at this preliminary stage, the use of RDF triplets instead of whole texts significantly improved the reduction in the hallucination rate of the models, inciting further study of this methodology.

These results show that, at least at this preliminary stage, the use of RDF triplets instead of whole texts significantly improved the reduction in the hallucination rate of the models, inciting further study of this methodology.

5. Conclusion

Compared to traditional solutions, such as essential databases or non-semantic search engines, LLMs enhanced with ontologies offer a more dynamic, intelligent, and accurate way to navigate and interact with cultural heritage data. This enables institutions and businesses in this domain to provide a richer, more informative, and user-friendly experience, driving innovation in research, education, tourism, and cultural industries.

This proposed approach has considerable potential for enhancing the comprehension of cultural artifacts and for addressing significant challenges, including reducing hallucinations and enhancing the factual reliability of AI-generated outputs through the RAG method.

Furthermore, there needs to be more literature on work related to the RAG approach using ontologies devoted to cultural heritage. This makes the scope of this research exciting from the standpoint of its potential for innovation and usefulness.

The next phase of the project will entail continuous updating from the literature and the parallel development of an efficient solution capable of efficient ontology embedding without sacrificing the links that characterize these formal knowledge representations.

References

[1] G. M. Currie, "Academic integrity and artificial intelligence: is ChatGPT hype, hero or heresy?," 2023. doi: 10.1053/j.semnuclmed.2023.04.008.

[2] G. Trichopoulos, "Large Language Models for Cultural Heritage," in *ACM International Conference Proceeding Series*, 2023. doi: 10.1145/3609987.3610018.

[3] N. Constantinides, A. Constantinides, D. Koukopoulos, C. Fidas, and M. Belk, "CulturAI: Exploring Mixed Reality Art Exhibitions with Large Language Models for Personalized Immersive Experiences," in *UMAP 2024 - Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, 2024, pp. 102 – 105. doi: 10.1145/3631700.3664874.

[4] M. Virvou, G. A. Tsihrintzis, D. N. Sotiropouloss, K. Chrysafiadi, E. Sakkopoulos, and E. A. Tsichrintzi, "ChatGPT in Artificial Intelligence-Empowered E-Learning for Cultural Heritage: The case of Lyrics and Poems," in *14th International Conference on Information, Intelligence, Systems and Applications, IISA 2023*, 2023. doi: 10.1109/IISA59645.2023.10345878.

[5] J. Chen, O. Mashkova, F. Zhapa-Camacho, R. Hoehndorf, Y. He, and I. Horrocks, "Ontology embedding: A survey of methods, applications and resources," 2024.

[6] Y. Gao *et al.*, "Retrieval-Augmented Generation for large Language Models: A survey," 2023.

[7] "Breaking language barriers: The role of large language models in multilingual communication," *International Research Journal of Modernization in Engineering Technology and Science*, Jun. 2024.

[8] J. Li *et al.*, "Can LLM Already Serve as A Database Interface? A BIg Bench for Large-Scale Database Grounded Text-to-SQLs," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., Curran Associates, Inc., 2023, pp. 42330–42357. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/83fc8fab1710363050bbd1d4b8cc0021-Paper-Datasets_and_Benchmarks.pdf

[9] J. Li, Y. Yuan, and Z. Zhang, "Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases," 2024.

[10] J. Chen, P. Hu, E. Jimenez-Ruiz, O. M. Holter, D. Antonyrajah, and I. Horrocks, "OWL2Vec*: embedding of OWL ontologies," *Mach Learn*, vol. 110, no. 7, 2021, doi: 10.1007/s10994-021-05997-6.

[11] T. G. M. M. S. M. E. C. Darya Shlyk, "REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition," *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 380–389, Aug. 2024.

[12] "https://github.com/chroma-core/chroma."

[13] O. Fagbohun, R. M. Harrison, and A. Dereventsov, "An Empirical Categorization of Prompting Techniques for Large Language Models: A Practitioner's Guide," *Journal of Artificial Intelligence, Machine Learning and Data Science*, vol. 1, no. 4, 2023, doi: 10.51219/jaimld/oluwole-fagbohun/15.