SAUC Street Art & Urban Creativity Scientific Journal

Street Art & Urban Creativity | Vol. 11, No. 7, 2025 | ISSN 2183-9956

Street Art & Urban Creativity. Scientific Journal

https://doi.org/10.62161/sauc.v11.6110

OBJECT RECOGNITION AND CONVERSATIONAL AI IN REAL-WORLD CONTEXTS

Enhancing Museum Experiences through Interactive Systems

ADRIÁN ORTIZ RAMÍREZ[1], ÁLVARO ILLANA SÁNCHEZ[1], MARTA SALAS GARCÍA[1]

[1] UNIVERSIDAD FRANCISCO DE VITORIA, SPAIN

KEYWORDS

Artificial Intelligence
Generative AI
Cultural Heritage
Museums
Object Detection
Information Retrieval
Context Awareness
RAG

ABSTRACT

This project addresses the challenge of improving museum visitors' experiences by moving beyond static and traditional information access methods. It presents the design and validation of an interactive system that combines real time object detection with a Retrieval Augmented Generation pipeline to offer a context-aware, personalized and immersive conversational guide.

The results verify an accurate spatial and conversational understanding, and a significant improvement in the veracity and relevance of generated responses in comparison with standard LLM responses. This project demonstrates the system's potential to offer a dynamic and attractive access to cultural heritage.

RECEIVED: 31 / 08 / 2025
ACCEPTED: 08 / 10 / 2025

1. Introduction

The start of the XXI century has been characterized by unprecedented technological advances, driven by the growth of Artificial Intelligence (AI) in an increasingly digitized world. These technologies are more than just tools. They are transforming people's engagement with information, learning, and the way they experience the world around them. This study arises from a fascination with that transformative potential, specifically exploring how AI can revolutionize sectors that are deeply rooted in tradition but eager for innovation, such as the cultural heritage sector. This work investigates the design, development, and validation of a novel system that leverages modern AI solutions to serve as a personal, intelligent guide that enriches the experience of museum visitors.

1.1. The Problem: The Static Museum Experience

Since the establishment of the first known museum, dating to circa 530 BCE, these institutions have been dedicated to the preservation and diffusion of knowledge and culture. Museums, as custodians of history and culture, house immense repositories of knowledge. However, the cultural heritage sector faces the challenge of making that knowledge both accessible and engaging for audiences.

Historically, the methods for communicating this knowledge have been fundamentally static. Traditional museum exhibitions relied on passive communication tools such as artwork labels, explanatory posters, and guidebooks. These methods, while somehow informative, often fail to fully engage visitors on a personal level, offering a one-size-fits-all narrative that cannot adapt to individual curiosity. To enhance interactivity and immersion, these institutions later adopted technologies like audio guides, which first appeared in 1952. While an improvement, they remained informationally static and, thus, unadaptable, restricting pace and autonomy with static recordings. Now, in the 21st century, digital solutions like interactive displays and Augmented Reality (AR) have started to appear. While more engaging, interactive and immersive, these solutions still struggle to accommodate individual interests, often following static narratives that limit personal exploration.

The informational staticity found in current and traditional solutions is what the present work aims to overcome through a more dynamic and customized approach.

1.2. Industry Trends: In Search for Effective Personalization

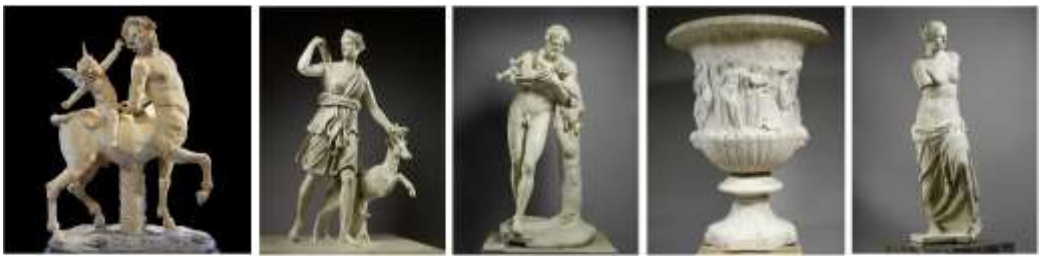
The tourism industry is a major economical driver worldwide. Europe, for instance, received 51.7% of total international tourism in 2024, contributing roughly 10% to the average GDP of European countries, with countries like Spain and Croatia highly exceeding this average. A notable trend of this industry has been the emergence of Tourism 4.0, which focuses on unlocking the potential of innovation to create enriched tourist experiences. As a self-defined driver of Sustainable Development Goals (SDGs), the Tourism 4.0 organization backs up the necessity of this innovation through its recurrent reports.

As main players in cultural tourism, museums report annual innovation trends that reflect this shift. According to Tourism 4.0's 2021 Museum Innovation Barometer, 80% of museums considered important new technologies, with 72% of their data intelligence initiatives aimed at enhancing visitor experiences. AI adoption rose from 3% to 14%, audio guide usage from 5% to 31%, and mobile and web applications from 49% to 70%, highlighting a clear shift towards personalization and interactivity. Nonetheless, from the visitor point of view, studies show a worrying tendency: Artwork engagement remains brief, averaging 27–29 seconds of interest per piece over the past two decades (Smith & Smith, 2001; Smith et al., 2017). These patterns underscore a growing need for personalized, interactive experiences that go beyond traditional museum engagement. Therefore, building on these trends, we developed a Proof of Concept (PoC) to test how AI can provide adaptive, customized interactions for museum visitors.

1.3. Scenario and Scope: A Proof of Concept at the Louvre

The PoC presented in this article takes place on the Department of Greek, Etruscan and Roman Antiquities of the Louvre Museum, more precisely on a few sculptures present in two rooms: The "Salle des Caryatides", and the "Salle de la Vénus de Milo", as illustrated in Figure 1. Being the most visited museum worldwide and a global leader in scale and innovation, the Louvre Museum naturally emerged as the key setting for our PoC. Specifically, the Louvre's publicly available database of over 500,000 artworks made this museum an ideal setting for an AI-based project.

Figure 1. Sculptures from the Salle des Caryatides and the Salle de la Vénus de Milo used in the PoC.



Source: Louvre Museum., 2025.

This work is framed within the education, technology and culture ambits, specifically focusing on innovation inside museum experiences. It is aligned with the Sustainable Development Goals of the 2030 Agenda (United Nations General Assembly, 2015), specifically with goal 4.7, that targets education through culture promotion.

In terms of technical scope, this work focuses on the design, development, and validation of a conversational AI-powered museum guide system. The process encompasses industry and technology research, system and architecture planning, training of the PoC models, and performance validation.

1.4. Objectives: From Vision to Measurable Goals

To achieve the overarching aim of this project, the work focuses on several interconnected goals that address personalization, information quality and knowledge management. These objectives break down the general purpose of the project into specific and, most importantly, measurable and verifiable aspects.

One main objective of the project is to enable real-time gathering of individualized context. Since humans primarily rely on sight, the system will allow identification of nearby artworks using object detection models trained on museum-specific datasets. Additionally, it will incorporate auxiliary information, such as visitor location, past interactions, and other preferences, into the contextual equation. Equally important is the ability of the system to generate answers that are both accurate and contextually relevant. By leveraging museum-curated datasets through a Retrieval-Augmented Generation pipeline, the system can deliver responses that are both meaningful and factually grounded, reducing the risk of confabulations or hallucinations. Finally, the project seeks to establish a dynamic knowledge integration pipeline to maintain the accuracy of the system as well as its relevance over time. As museums constantly update their collections, it is essential to design an automated process for ingesting, preprocessing, and indexing new information, therefore, ensuring that the system remains current, consistent, and reliable.

The effectiveness and reliability of the system will be validated through measurable metrics, such as accuracy of visual recognition, relevance of retrieved information, and consistency of the knowledge integration pipeline.

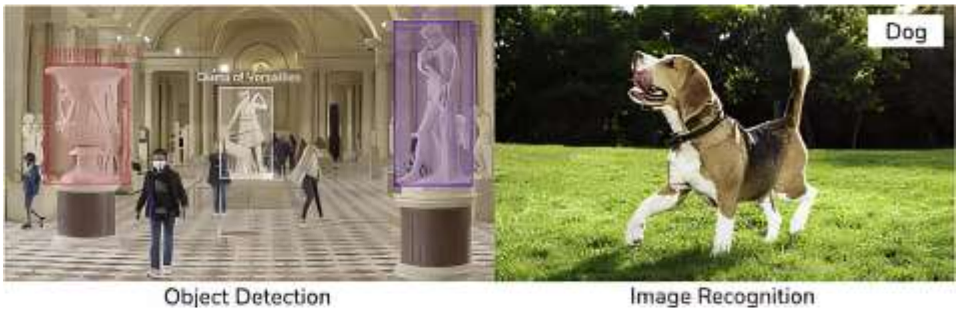
2. State of the Art

Context gathering, information retrieval, response generation, and knowledge integration are areas that have been extensively studied and explored over the years. In the context of museums and cultural heritage, these technologies are being studied and applied to enhance visitors' experiences.

2.1. Computer Vision

Under the computer vision umbrella, tasks such as image recognition and object detection play a critical role in helping machines be able to perceive and understand visual information. As illustrated in Figure 2, image recognition assigns a single label to the entire image, while object detection identifies and localizes several objects within the same image through bounding boxes. Specifically in museums, image recognition has been widely studied for automated artwork classification tasks (Cetinić et al., 2018; Fortuna-Cervantes et al., 2024; Li, 2025). Object detection, on the other hand, has yet to be widely implemented in cultural heritage settings. Among the few current uses of object detection in this sector, we can distinguish inside-painting object exploration and visitor tracking (Breitner & Bandung, 2024; Meyer et al., 2024). However, for our use case, object detection presents a greater potential than image recognition.

Figure 2. Object detection vs Image recognition



Source: Own elaboration, 2025

Currently, museum guide apps that include computer vision, such as Ask Mona (2025) or Smartify (2025), focus on processing a single captured image of an artwork at a time, rather than supporting real-time detection of multiple artworks simultaneously. This is exactly where image recognition stands out. In contrast, this project aims to move past such limitations by emphasizing real-time object detection, allowing the system to identify several artworks at once within their context.

To achieve real-time object detection, existing solutions have been explored. As the project aims to perform at a real time rate, it prioritizes efficiency over SOTA accuracy. Some existing solutions include You Only Look Once (YOLO) (Redmon & Farhadi, 2017), Faster Region-based Convolutional Neural Networks (Faster R-CNN) (Ren et al., 2015) and Single Shot Multibox Detectors (SSD) (Liu et al., 2016) architectures, among these solutions YOLO surpasses other architectures in terms of efficiency and accuracy for big objects, while Faster R-CNN gets great results on smaller objects but with lower times.

2.2. Retrieval Augmented Generation

The widespread adoption of Large Language Models (LLMs) underscores their significance. Yet, despite advancements, they remain prone to hallucinations from their reliance on parametric memory, making them unreliable in specialized domains. In museums, for example, LLM outputs often show cultural misalignments, which can lead to misleading or distorted interpretations of artworks. Recent studies report that such misalignments can reach up to 65% in the cultural heritage ambit (Bu et al., 2025). To address this, prompt engineering techniques have started to gain attention. An emerging solution is Retrieval-Augmented Generation (RAG), one of the most studied techniques. RAG, introduced in (Sahoo et al., 2024), integrates parametric and non-parametric memory to ground answers in retrieved facts, with its effectiveness in museums shown in (Loffredo & De Santo, 2024; Vastakas, 2024). Its adoption is even being noticed in critical sectors like medicine and security in projects such as (Du et al., 2024; Wu et al., 2024).

Beyond reducing hallucinations, RAG has proven particularly valuable, as it allows knowledge to remain up to date when new information becomes available. This is a crucial aspect, as LLMs alone are only able to generate information up to their training cutoff. In museums, where catalogues, records, and other curatorial files are constantly being updated, RAG can ensure that information remains accurate and relevant.

2.3. Current Solutions

In the current state of the art, Ask Mona stands out as the most complete solution. Its mobile app personalizes visitors experience by combining conversational AI with artwork recognition, allowing users to scan artworks and receive contextual and personalized responses. The system integrates content from some of the world's most renowned museums and is already trusted by over 150 organizations globally. However, the image recognition feature restricts the amount of context the system can capture, requiring users to rely on the app repeatedly whenever they wish to explore new artworks. Our PoC addresses this limitation by incorporating context-streaming functionality and replacing image recognition with object detection, enabling the system to autonomously adapt to visitors' evolving context in real time.

Other solutions, such as Smartify (2025) and Nubart (2025), address certain aspects of our requirements for a personalized and context-aware experience. For example, Smartify offers image recognition and Nubart provides QR scanning, but both lack a conversational personalization and still depend on conventional audio guide systems.

2.4. Differential Value Proposition

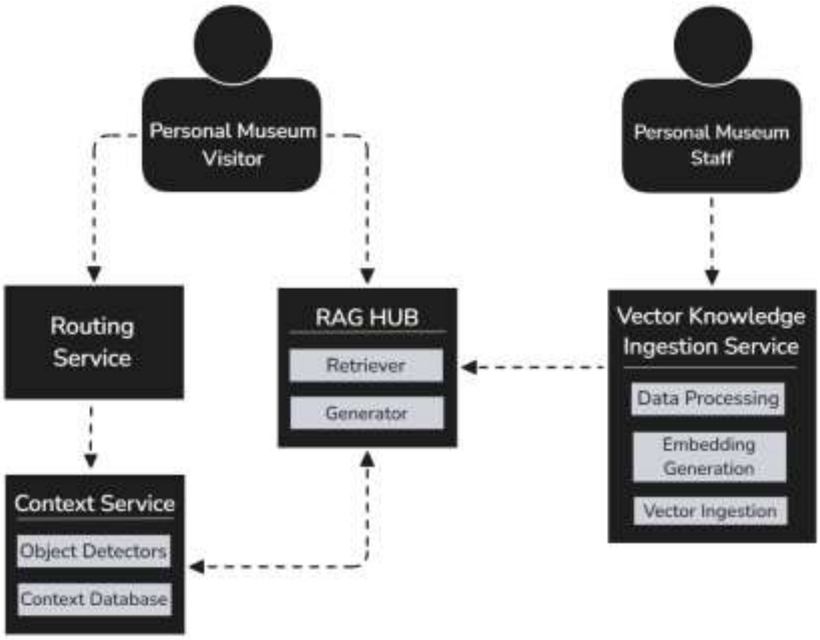
This study differentiates itself by integrating real-time object detection with contextual information retrieval, offering an immersive and personalized museum experience through visual and conversational means. While some existing solutions provide general information about art pieces, they often lack contextual awareness and adaptability to curiosity. By combining real-time object detection with retrieval-augmented generation, our solution can dynamically retrieve relevant documents from a vector database, considering not only the visitor query but also their environment. While RAG technology has proven relevant in sectors like medicine and security, its application can be adapted to other fields where access to accurate information is essential. The lack of documented usage in cultural spaces, where information is crucial, indicates a potential need for its benefits. This project innovates in the way of interacting with museum collections, offering visitors an intelligent and dynamic guide capable of answering questions in a human-like manner while considering their surroundings in real time.

The true value of this project lies in its capacity to change how people interact with culture and history. This approach empowers each user to take an active role in their own learning journey, adjusting to their pace and background. At its core, this project is not just about technology, but about using innovation to connect individuals with culture, preserve cultural heritage, and make knowledge more engaging and easily accessible.

3. System Design

The system's design has been organized as a modular and scalable client-server architecture that connects real-time visual input with knowledge retrieval and language-model reasoning, providing context-aware responses for the user. Figure 3 provides a high-level schematic, highlighting the main modules and the relationships between them.

Figure 3. High-level system architecture.



Source: Own elaboration, 2025.

The *Routing Service* receives the visitor's stream video and location, then applies location-aware routing to forward the video stream into the *Object Detector* trained for the corresponding room. The *Context Service* processes the detections, links them with the ongoing session and conversation history, and stores them in the *Context Database*, making this context available to other modules and recording final answers for continuity. In parallel, the *Vector Knowledge Ingestion Service*, maintained and fed by museum staff, collects and cleans collection sources, splits them into semantically coherent chunks, runs embedding generation, and completes vector ingestion into the vector database. At query time, the *RAG Hub* uses the visitor's question together with recent detections and history to retrieve and re-rank relevant documents from the vector store, then its *generator* produces grounded responses that are returned to the client and written back to the *Context Service*.

This flow supports real-time, sight-aware, and session-aware guidance while remaining scalable across multiple simultaneous visitors.

3.1. Routing Service

For low-latency, real-time communication, the *Routing Service* splits the visitor's input into small packets and organizes them into queues, enabling efficient threaded processing.

The visitor's contextual stream mainly consists of video frames of what the visitor is currently seeing in real time, and the visitor's current physical location inside the museum. Once the client starts sending contextual information packets, the service dynamically routes its video frames to the appropriate object detection model, which has been specifically trained for the corresponding room where the visitor is.

The design, based on queues and location-aware routing, provides high scalability and supports multiple clients simultaneously while keeping low response times. The design simplifies the distribution of workloads across by adapting to different physical spaces.

3.2. Context Service

The *Context Service* is responsible for processing the visual input of visitors, generating predictions about the artworks in sight, and storing this information together with the conversation history, so that it can be reused during retrieval and generation. Its main functionality is to manage the visitor's contextual information relevant to its current visit so that the final LLM responses are as contextually accurate as possible. This service is organized into three main components.

3.2.1. Context Database

The *Context Database* stores contextually relevant information from the visitor's session. For each visitor, a unique identifier is preserved and linked to the predictions generated and to the history of past user queries and system responses, together with their respective queues. This relational design ensures consistency between entities, keeps the information strictly session-related, and allows the stored data to be reused during retrieval and generation processes.

3.2.2. Context Handler

The *Context Handler* is used to ensure that all relevant contextual information is consistently available during a visitor's session, being securely deleted upon session termination to protect visitor privacy. It stores the Object Detector's predictions and conversational exchanges, allowing the RAG pipeline to consider what the visitor has recently seen, enabling more accurate and relevant responses. At the same time, it records visitors' conversational exchanges so that dialogues can continue naturally without losing track of prior messages. Centralizing these tasks in a single service prevents individual microservices from having to manage database logic, which simplifies the architecture, increases modularity, and safeguards data handling. Ultimately, the *Context Handler* is what enables the system to maintain continuity, coherence, and contextual awareness throughout the visitor's experience.

3.2.3. Object Detector

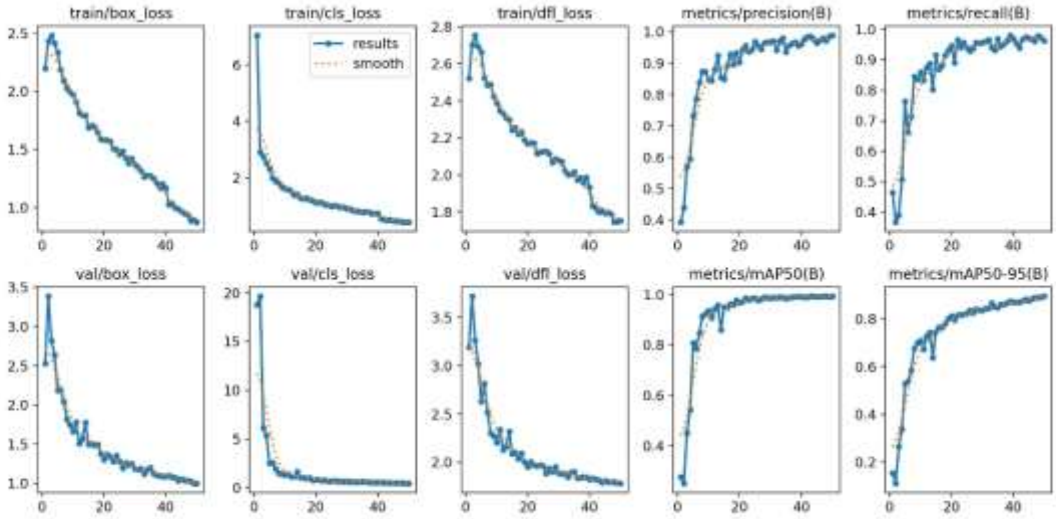
The *Object Detector* is responsible for processing the video frames received from the *Routing Service* and generating predictions about the artworks in sight. Once the frame is sent to it, the YOLO model computes predictions, later stored in the *Context Database*, where they can be combined with the conversation history. This component plays a central role in enabling sight-aware responses, as it links the visitor's real-time visual input with the contextual information managed by the rest of the service.

Given that the Louvre's collection includes over 500,000 objects with approximately 35,000 on permanent display, the system must be scalable. Although experiments such as YOLO9000 have demonstrated the ability to manage thousands of classes (Redmon & Farhadi, 2017), increasing the number of categories in convolutional architectures normally affects performance (Luo et al., 2018). To address this, the *Routing Service* integrates the previously mentioned location-aware routing approach, therefore, several models are trained with different image sets of artworks located in different rooms of the museum.

The Louvre Museum provides structured access to its collection through JSON endpoints, offering both information and images of the artworks (Louvre Museum, 2025). However, these images were not sufficient to generate a training dataset. Therefore, additional material was obtained from YouTube videos with the permission of the corresponding creators (Robben, 2019). The labelling process included both manual and automatic techniques. Initially, an open-set detection model named Grounding Dino (Liu et al., 2024), was used to take each image together with a class description and detect the corresponding object within the image. This approach introduced issues when dealing with similar sculptures inside the same image. To address this, images with several sculptures were manually labelled to avoid bias in the training data, using the open-source tool CVAT (CVAT, 2025) to annotate videos into YOLO format. The labelling process was preceded by an augmentation phase, where classes were given more diversity through blur, saturation, brightness and smooth rotations to simulate real-world situations. In addition, non-labelled images from other rooms of the museum were used as background images, helping to prevent the appearance of false positives. After this process, the dataset included around 525 annotated instances, divided into training (70%), validation (20%), and testing (10%) subsets, maintaining a balanced class distribution across subsets.

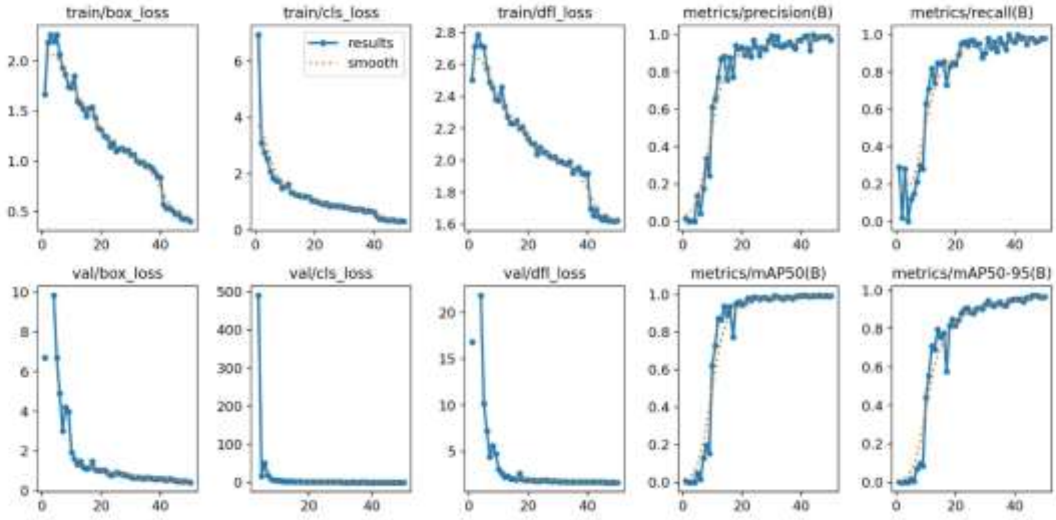
The training process started by selecting a pretrained YOLO model from the official range of models. These detection-based models are pretrained on the COCO dataset (Lin et al., 2014) and available in different versions and sizes. YOLOv10 (Wang et al., 2024) was chosen as a base for fine-tuning, as it provides a significant performance increase over previous versions while keeping a lower computational cost than later ones. Among the available sizes, YOLOv10m.pt was selected as a resource-reasonable option with a balanced speed-accuracy exchange. The fine-tuning process was applied to a pair of models: the Room 1 model trained in four sculptures with a combination of automatic and manual labelling, and the Room 2 model trained in a single sculpture using automatic labelling. The main metrics of the training and performance evaluation over epochs are displayed in Figure 4 and Figure 5, while the F1-confidence curves for both models are presented in Figure 6. From these results, it can be observed that the models accurately and confidently generate predictions without major overfitting on the training set. In addition, Figure 7 summarizes the evaluation of YOLO training performance metrics for both rooms.

Figure 4. YOLOv10m model training metrics over epochs (Room 1).



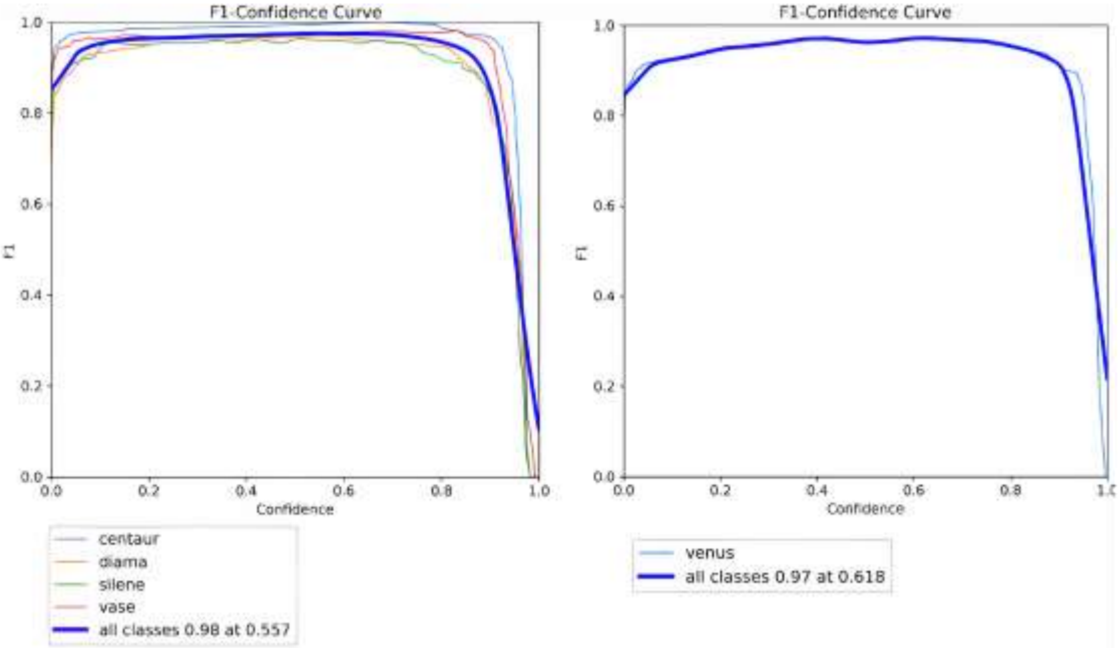
Source: Own elaboration, 2025.

Figure 5. YOLOv10m model training metrics over epochs (Room 2).



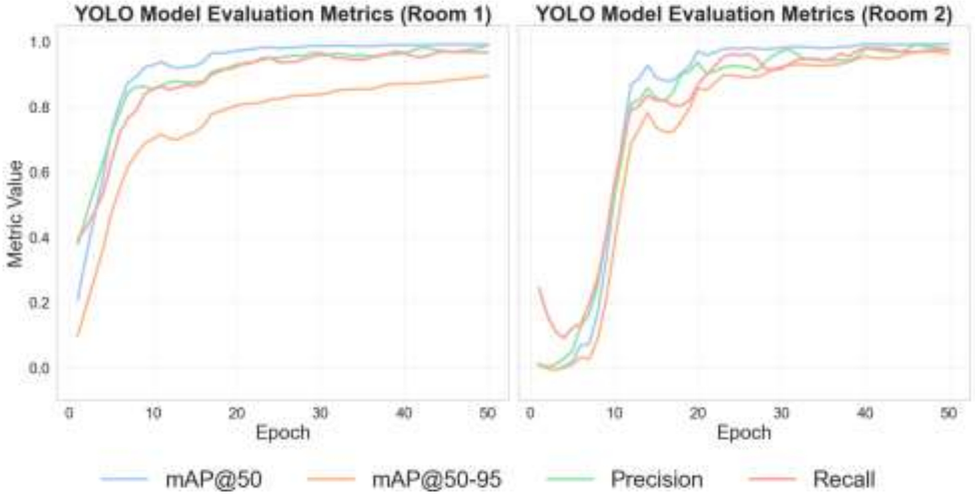
Source: Own elaboration, 2025

Figure 6. F-1 Confidence curve for Room 1 (left) and Room 2 (right) after training.



Source: Own elaboration, 2025.

Figure 7. Evaluation of YOLO training performance metrics for Room 1 and Room 2.



Source: Own elaboration, 2025.

3.3. Vector Knowledge Ingestion Service

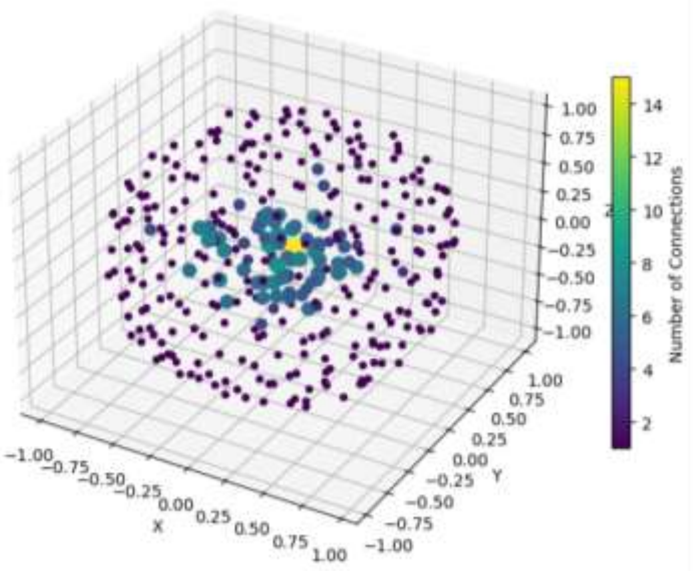
To prepare museum-related documents for retrieval and generation, this service collects, processes and stores information into a vector database. The workflow begins with document collection from available sources, continues with data processing for cleaning and splitting into semantically coherent chunks, and proceeds with the generation of vector representations through embedding models. Once prepared, both chunks and embeddings are ingested into the database, enabling similarity search across the collection.

3.3.1. Document Collection

The quality of the data stored in the vector database determines the effectiveness of future retrieval tasks. Both the structure and contextual significance of the corpus of documents are key for accurate results. Relevant information is gathered from collection catalogues, historic books or visitor-facing contents, but data accessibility is constrained in most cases. Although the Louvre provides access to its collection through public JSON endpoints, including images and descriptions, these materials were not sufficient to meet the system's requirements. Therefore, additional sources were explored, including public domain books about the museum and Wikipedia articles. Despite debates about its reliability, Wikipedia offers well organized and semantically dense content, and its use has already been studied for similar NLP applications (Yano & Kang, 2008), which served as a great testing scenario for this PoC.

To support the selection of Louvre-related Wikipedia articles to be added into the vector database, a depth search was performed starting with the museum's main article "Louvre". From there, the fifteen most frequently linked pages were retained, and in each subsequent step the number of links to retain was reduced until reaching a maximum depth of three. Figure 8 shows the final link graph where the node size and color represent the number of incoming connections; the central yellow node corresponds to the article on the Lescot Wing of the Louvre Museum, which turns out as the most cited when starting from the main Louvre article.

Figure 8. Link graph of Wikipedia articles starting from "Louvre".



Source: Own elaboration, 2025.

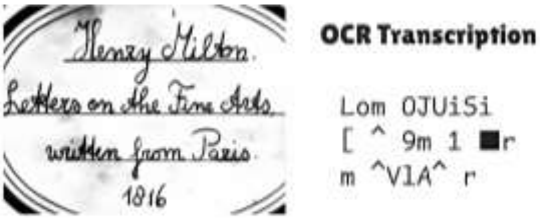
Although not every article in the graph is directly related to the Louvre Museum (e.g. Charles' V Wall in Gibraltar), it served as a useful tool for fast louvre-related article identification. Apart from these articles, some, less popular but still relevant for the demonstration of the project, were included. For instance, the Borghese vase article. This vase is displayed at the salle of the caryatides and has been included in the object detection model's classes of this project, thus some information about it was added into the knowledge of the vector database for validation purposes.

3.3.2. Data preprocessing

The raw data collected in the acquisition phase required preprocessing to ensure its suitability for embedding and ingestion. Preprocessing techniques vary depending on the type of information. In this case, information followed four different formats: HTML codes from Wikipedia article pages, and three types of PDFs: True PDFs, Image-based PDFs and Made-searchable PDFs, which refer to the way text is encoded and accessed within a file.

Out of these formats, Image-based PDFs and Made-searchable PDFs required OCR techniques to extract text. However, OCR proved less accurate on historical texts, where characters follow non-standard typographies and books may have deteriorated over time. After analysis of the raw data from these PDFs, both formats were discarded as the text quality was insufficient. Figure 9 illustrates an unsuccessful OCR transcription from an early nineteenth-century book, while Figure 10 shows irrelevant information extracted from historical sources, such as footnotes. Although training an OCR model on early printings could have improved accuracy on historical texts (Springmann & Lüdeling, 2016), this approach was not pursued due to scope constraints. Instead, information from True PDFs and HTML formats was adopted. HTML data required structural cleaning, which included removing language tags, reference markers (e.g. "[]"), and retaining only the main body content. As an example, Figure 11 displays the original HTML code of a Wikipedia page and the corresponding cleaned text that was used for ingestion. Once the raw data was refined into a clear set of large texts, chunking could be performed.

Figure 9. Unsuccessful OCR transcription from Henry Milton's letter.



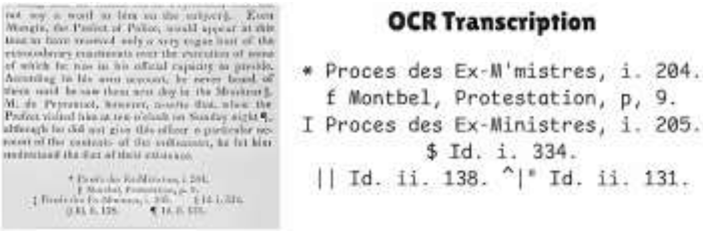
Henry Milton.
Letters on the Fine Arts
written from Paris.
1816

OCR Transcription

```
Lom OJUiSi
[ ^ 9m 1 ■r
m ^V1A^ r
```

Source: Own elaboration, 2025.

Figure 10. Irrelevant information extraction from "Paris, and Its Historical Scenes".

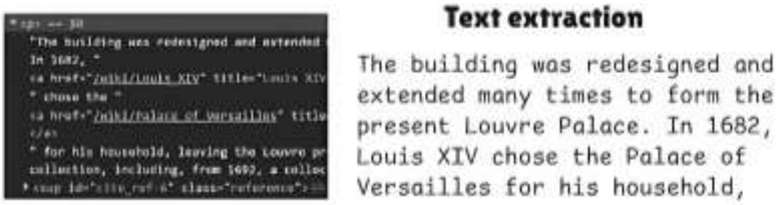


OCR Transcription

```
* Proces des Ex-M'mistres, i. 204.
  f Montbel, Protestation, p, 9.
I Proces des Ex-Ministres, i. 205.
        $ Id. i. 334.
  || Id. ii. 138. ^|" Id. ii. 131.
```

Source: Own elaboration, 2025.

Figure 11. Wikipedia article content extraction.



```
* <p> == $0
  "The building was redesigned and extended 
  In 1682, "
  <a href="/wiki/Louis_XIV" title="Louis XIV
  " chose the "
  <a href="/wiki/Palace_of_Versailles" title
  </a>
  " for his household, leaving the Louvre pr
  collection, including, from 1692, a colloc
 ▸ <sup id="cite_ref-6" class="reference">…
```

Text extraction

```
The building was redesigned and
extended many times to form the
present Louvre Palace. In 1682,
Louis XIV chose the Palace of
Versailles for his household,
```

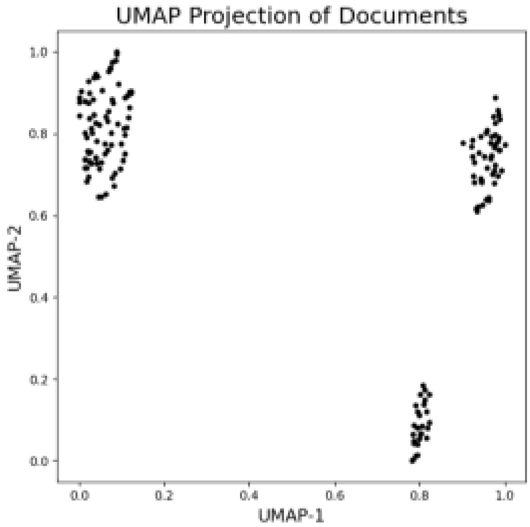
Source: Own elaboration, 2025.

The *Chunk Loader* component of the project manages the splitting of large texts into semantically meaningful units of information, referred to as documents. This step is fundamental in the context of retrieval augmented generation, as language models have token limits and are sensitive to input length (Levy et al., 2024). Moreover, chunking strategies depend on the structure and content of the data, making it not only a technical task but occasionally a domain-sensitive decision process. A detailed analysis of possible methods was conducted, concluding that recursive character text splitting was the best option for our purpose. This method combines simplicity, which transformer-based methods lack, with promising retrieval results. While fixed-size chunking does not look beyond a defined length, recursive character text splitting adapts its length based on text structure, prioritizing natural boundaries such as sentences and paragraphs, and making use of separators (e.g. ["\n\n", "\n", " ", "."]). LangChain, a framework for building LLM-related applications (LangChain, 2025), provides this splitter component and allows tuning chunk size and overlap parameters. Chunk size limits the number of characters in a document, while overlap specifies the number of characters repeated from the end of the last chunk.

3.3.3. Embedding Generation

Embedding models, commonly referred to as bi-encoders, play a key role in semantic retrieval, as they represent the intrinsic meaning of texts as dense vectors. A small and efficient model was preferred to prioritize fast performance. The Sentence-Transformers framework was leveraged with the "all-MiniLM-L6-v2" model, based on the MiniLM architecture (Wang et al., 2020) and derived from BERT (Devlin et al., 2019). This model maps sentences into a 384-dimensional dense vector space with a maximum input length of 256 tokens. Therefore, chunking was adapted to stay within 1024 characters. To manage vectorization, an Embedding API server was created to receive chunks from the *Chunk Loader*, generate embeddings, and load both into the vector database. The same API also processes user queries from the *RAG Hub*, embedding them for semantic search through similarity computations. For visualization, Uniform Manifold Approximation and Projection (UMAP; McInnes et. al., 2018) was applied to reduce the dimensionality of the embeddings and display clusters of documents, as shown in Figure 12, which includes examples from the Mona Lisa, the Cathedral of Notre Dame and the Vénus de Milo.

Figure 12. Vector space of the system, including 3 classes.



Source: Own elaboration, 2025.

3.3.4. Vector Ingestion

After collecting, refining, slicing, and embedding historical data, the system must ensure that the processed documents are accessible for retrieval in response to user queries. This functionality is provided by the vector database, which is directly connected to the embedding API. When the embedding API receives a document from the *Chunk Loader* and generates its corresponding embedding, both the document and the embedding are immediately stored in the database.

Chroma was selected as the vector database. It works with collections of documents that can be configured to use different indexing algorithms, but it relies on the Hierarchical Navigable Small World (HNSW) index for approximate nearest neighbor search by default. To prevent duplicate entries, the system generates unique identifiers for each record by hashing the content of the embedded documents. If a document already exists in the database, the insertion is rejected.

3.4. RAG Hub

The *RAG Hub* is responsible for answering visitor questions in natural language while integrating not only the query itself, but also contextual information related to the detected object. It implements a Retrieval Augmented Generation architecture that combines information retrieval from external (non-parametric) memories with response generation. The system is structured around two components: a retriever module, which fetches relevant information from external knowledge sources, and a *Generator* module, which employs a parametric memory for response generation. This design avoids LLM knowledge limitations and minimizes misinformation errors by grounding its responses on retrieved information, aiming to balance response relevance with efficiency.

3.4.1. Retriever

The *Retriever* module orchestrates retrieval, re-ranking, and context preparation. Its pipeline starts as soon as a request is received from a client, which includes not only a question but also the most recent object detections and historical messages (contextual information). The system must first determine the relevance between detected objects and the question. For instance, a user might be looking at a statue but asking something unrelated, therefore if retrieval was made considering the object, wrong documents would be gathered. To mitigate this, a validation mechanism is required to assess the semantic alignment between the question and the detected object.

To calculate the similarity between object labels and user questions, a cross-encoder model was employed, but its performance was uneven. Therefore, the approach was enhanced by reusing the generator model to rephrase the user's question (Ma et al., 2023). By appending previous messages, the detected objects, and the current query, the model generated a reformulated question that was more likely to retrieve relevant and contextually grounded documents.

For the retrieval strategy, several methods were considered. Among the most popular is BM25 (Robertson & Zaragoza, 2009), a keyword-based retriever method that looks for documents that match words from the query, assessing similarity by benefiting uncommon words and repetition while also considering text length. More sophisticated methods include semantic search (Karpukhin et al., 2020), cross-encoders (Rosa et al., 2022) and Maximal Marginal Relevance (MMR) (Carbonell & Goldstein, 1998, pp. 335-336) among others. The *Retriever* module makes use of a combination of all the previously mentioned to select the most promising documents. As part of the vector database retrieval process, a hybrid approach was followed (Bruch et al., 2022), where *k* number of documents were retrieved by each retrieving algorithm (semantic search, BM25 and MMR) and then deduplicated. It is worth noting that semantic search and MMR make use of the embedding API. Consistency comes from using the same bi-encoder model on both the document embeddings and query embeddings.

Once a large list of candidate documents is obtained, a re-ranking step is applied. Re-ranking consists of ordering the results from most to least relevant. While retrieval already provides an initial ranking, re-ranking allows the use of more computationally expensive but also more accurate models. Cross-encoders, although excellent at semantic understanding, are not normally used for first stage retrieval due to their computational cost. These are models that take a query and a document as a combined input and output a very accurate relevance score that reflects how the document matches the query. Cross encoders are commonly used to score the list of retrieved documents to select a smaller list of the most relevant ones.

After re-ranking and keeping the most promising results, a final step prior to sending them to the generator is performed. According to (Liu et al., 2024), LLMs tend to get lost around the middle in long contexts. LangChain (2025) provides a component that reorders documents so that the most relevant ones are positioned at the start and at the end of the LLM's context and the least relevant documents are positioned around the middle.

3.4.2. Generator

The *Generator* module corresponds to the LLM API within the system's architecture. This component hosts an LLM to generate answers. The Ollama framework (Ollama, 2025) was selected for its seamless integration with LangChain (2025) and its support for locally deployable open-source models, some including hundreds of billions of parameters, making them computationally expensive. To maintain a locally functional solution, models with 1 to 8 billion parameters were explored. After an evaluation on a predefined dataset, the "qwen2.5-3b" model was selected, as it was determined to be the most adequate for this PoC.

Also, when the *Retriever* sends to the *Generator*: the documents, past conversations and the query, it also selects a prompt template to use. Apart from a default template, the LLM API component incorporates others that could be chosen by the user, providing more personalized answers. Finally, when the answer is generated, the *RAG Hub* sends it back to the client, which sends it also to the *Context Handler* to store it as a new past message.

4. Evaluation and Results

Several scenarios were evaluated to both improve and verify the system's performance, ranging from the selection of models and justification of techniques to the formalization of test suits and datasets. Tests, validations and results are presented across the system's main modules, reflecting the sequential nature of the architecture from the *Context Service*, through the *Vector Knowledge Ingestion Service*, to the *RAG Hub*, which conveys the final performance of the complete system.

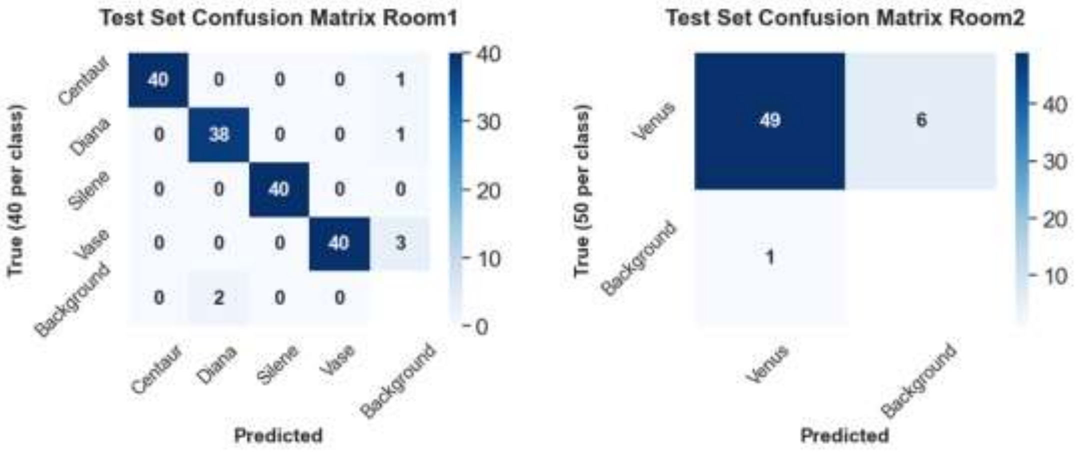
4.1. Context Service

While most of the testing was centered around the RAG components, additional evaluations were conducted to validate the system's communications precision. It is worth noting that the parameter selection was done automatically for the training of object detection models. The YOLO training framework provides an automatic way of selecting the most appropriate optimizer, learning rate and momentum values. Also, the creation of the dataset followed standard procedures on augmentation and image distribution. This resulted in satisfactory initial results from the first training. Therefore, only the process of validation was significant in both communications and object detection training.

As presented in Figure 7, two object detection models were trained on sculptures from two different rooms of the Louvre Museum. This training process was followed by a validation phase. The validation phase included performing inference through a test set for each model and evaluating their performance and generalization capabilities.

The confusion matrices, shown in Figure 13, indicate that both models benefit from an overall good generalization capability without interclass confusion, where the second model could benefit from a higher-diversity training set to reduce false negatives, and where the most common error among them is detection loss. Nonetheless, both models demonstrate strong performance, considering their high precision, recall and F1 score values among all classes, as summarized in Table 1.

Figure 13. Object Detection Performance Across Room1 and Room2 Test Sets.



Source: Own elaboration, 2025.

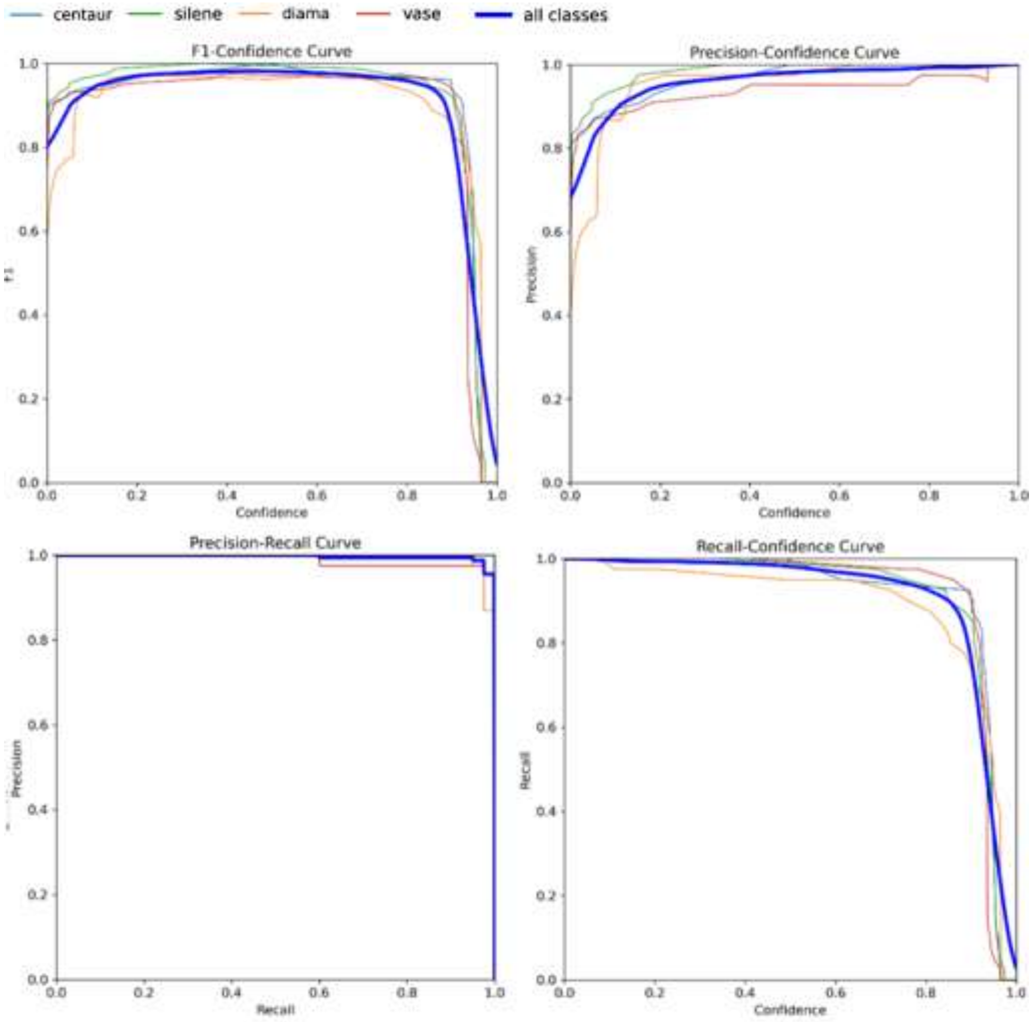
Table 1. Macro-averaged metrics for both object detection models.

Model	Precision	Recall	F1 Score
Room 1 (Centaur, Diana, Silene, Vase)	0.988	0.970	0.978
Room 2 (Vénus)	0.980	0.891	0.933

Source: Own elaboration, 2025.

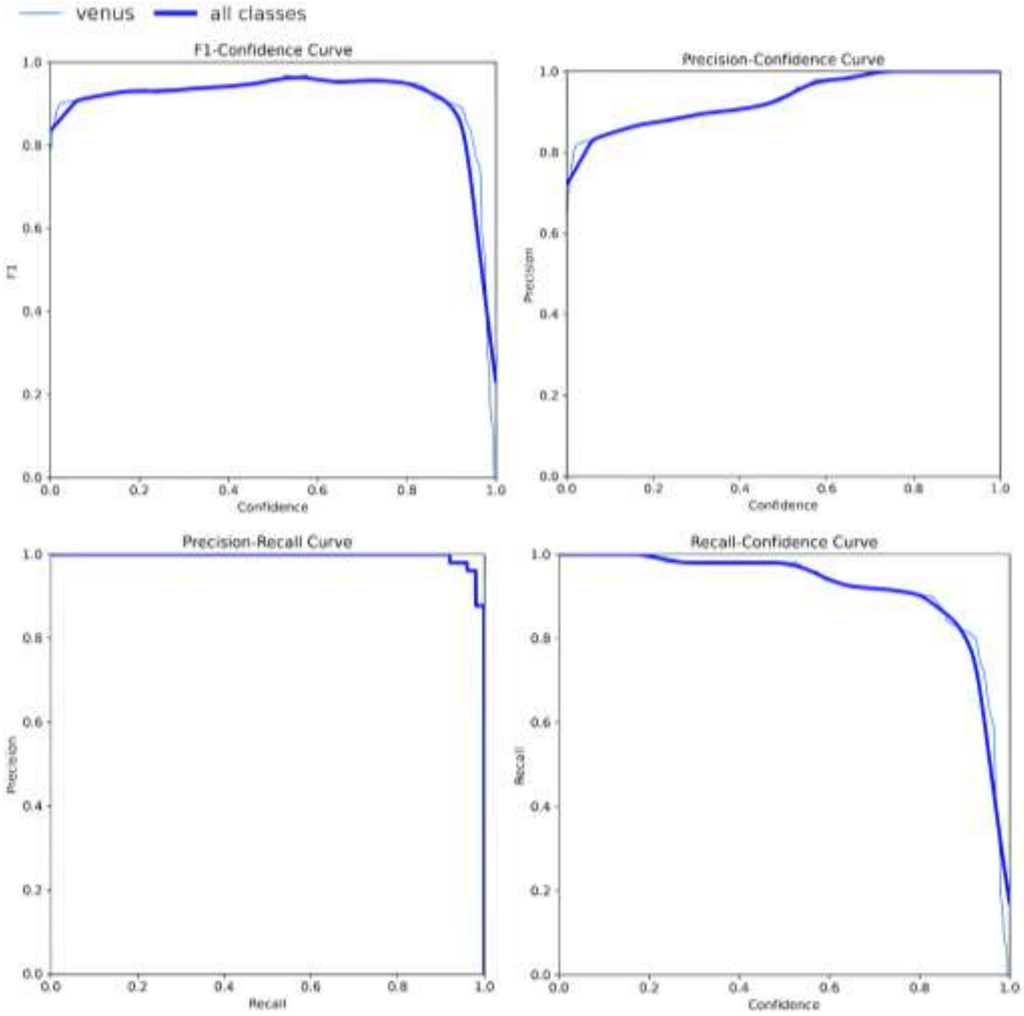
The evaluation curves for Room 1 and Room 2 are presented in Figure 14 and Figure 15, which illustrate that the models achieve a robust performance. These plots suggest a strong balance between false positives and false negatives (precision–recall), and a confident trade-off between prediction reliability and confidence. On both models, the F1 measure suggests that the models perform best around the 0.4–0.6 confidence interval.

Figure 14. YOLO Model Performance Evaluation Curves for Room 1.



Source: Own elaboration, 2025.

Figure 15. YOLO Model Performance Evaluation Curves for Room 2.



Source: Own elaboration, 2025.

The system does not only have to gather contextual information, but it must generate responses based on it. Therefore, the system performs retrieval based on previous interactions and the current view of the visitor and then generates answers based on that retrieved information and on those past interactions.

To verify that this is the case, the system has undergone a series of tests focusing on its capacity to refer to previous conversations and pertinent visual context. Examples of those tests are shown in Table 2 and Table 3.

Table 2. Contextualized generation scenario (Past History Messages).

Model: qwen2.5-3b-instruct + our system		
User Query	**Current Sight of Objects**	**History Messages**
What was I asking about earlier?		**User**: Tell me about the Borghese Vase. **Assistant**: The Borghese Vase is a 1st century BC marble krater.
Rewritten query:	What was I asking about earlier?	
Response: You were asking about the Borghese Vase, and I informed you that it is a 1st century BC marble krater.		

Source: Own elaboration, 2025.

Table 3. Contextualized generation scenario (Objects in sight).

Model: qwen2.5-3b-instruct + our system		
User Query	**Current Sight of Objects**	**History Messages**
When was this created?	Borghese Vase	**User**: **Assistant**:
Rewritten query:	When was the Borghese Vase created?	
Response: The Borghese Vase is believed to have been sculpted in Athens around the second half of the 1st century BC.		

Source: Own elaboration, 2025.

From the first scenario, the system correctly identifies that there is no need to rewrite the information for retrieval, as this question does not require external information. Therefore, when the response generation phase comes, the *Generator* correctly gathers context from previous history messages to respond.

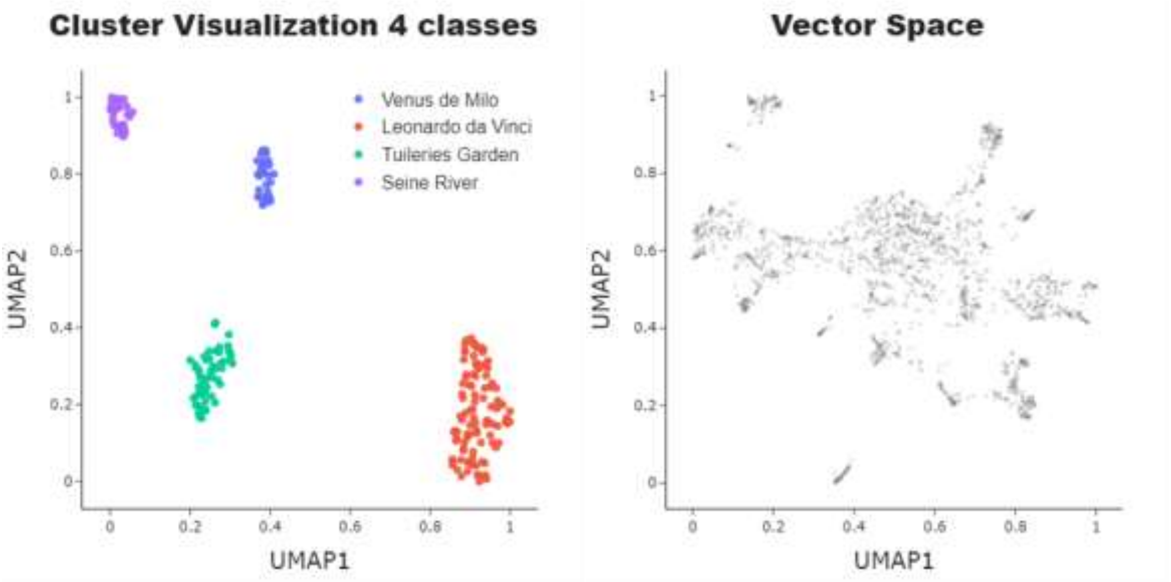
On the second scenario, the system does identify the need to rewrite the question, as the visitor's query is likely referring to its current sight. Therefore, it adds the current view of the user into the query to create a new contextualized query. Following this method, the system can perform contextualized response generation making use of the previously gathered visual context and past conversations.

4.2. Vector Knowledge Ingestion Service

Retrieved documents are automatically integrated into the knowledge base, ensuring consistency and quality. A well-structured vector space should cluster documents of the same class together.

Figure 16 shows that when including documents from 4 different classes, the vector space representation clearly spreads into well-separated clusters of similar documents. As expected from contents that include multiple classes, the final vector representation includes some clear groups of documents with linking documents between them, creating a structured disposition. Although the embedding quality can be clearly seen in the first plot, some metrics were gathered to confirm the overall accuracy of the knowledge integration process, as summarized in Table 4.

Figure 16. Vector space representation with 4 classes and final vector space.



Source: Own elaboration, 2025.

Table 4. Vectorization Quality Metrics on high-dimensional vectors.
Evaluating a labeled subset of documents and an unlabeled vector space.

Metric	Subset (4 classes)	Full Space
Precision@5 (avg)	0.955	-
Silhouette Score (cosine)	0.609	-
Neighborhood Distance k=5 (avg cosine)	0.395	0.380

Source: Author, 2025.

The vector space disposition quality was measured using two label-reliant metrics for the labeled subset and a non-reliant metric for both sets. A high Precision@5 indicates that the five nearest neighbors of a document are normally from the same class, which means that semantic similarity is well captured. A high silhouette score indicates that clusters of similar documents are well separated. A low neighborhood distance, measured between 0 and 2, indicates that neighbors are close in cosine space. These results validate the quality of the cleaning, chunking and embedding processes.

When it comes to consistency, the cleaning and chunking phases are deterministic. In addition, the embedding model's computations were tested by re-embedding the same text file multiple times, achieving a final average Mean Squared Error (MSE) of 0.

4.3. RAG Hub

A ground truth dataset of 126 Louvre-related questions and answers was created to evaluate different language models under the system's constraints. Several open-source models were tested, mostly in the range of 1 to 8 billion parameters, with some larger ones used as baseline comparisons, shown in Table 5.

Table 5. Model Evaluation on Key Generation Metrics (Louvre Dataset, no retrieval, ordered by COMET).

Model	BLUE	BERTScore (F1)	Levenshtein	COMET	SummaCZS
deepseek-r1-671b	0.345	0.942	0.714	0.797	0.337
llama-3.1-70b-instruct	0.234	0.933	0.665	0.781	0.095
qwen2.5-3b-instruct	0.152	0.920	0.616	0.775	-0.389
llama-3.1-8b-instruct	0.224	0.929	0.669	0.775	-0.106
qwen2.5-1.5b-instruct	0.203	0.921	0.661	0.759	-0.358
qwen3-1.7b-instruct	0.260	0.923	0.669	0.758	-0.483
qwen2.5-7b-instruct	0.199	0.920	0.631	0.754	-0.328
llama3.2-3b	0.131	0.909	0.575	0.714	-0.351
gemma-3-4b-it	0.258	0.913	0.583	0.680	0.232
gemma-3-27b-it	0.254	0.913	0.579	0.674	0.206
qwen3-4b-instruct	0.209	0.905	0.578	0.671	-0.314
gemma-3-1b-it	0.247	0.911	0.573	0.668	0.173
deepseek-r1-1.5b	0.095	0.897	0.523	0.665	-0.521
mistral-nemo-12b-instruct	0.166	0.879	0.482	0.586	-0.188

Source: Own elaboration, 2025.

From all the evaluated small models, qwen2.5-3b achieved the best COMET score (Rei et al., 2022). Its semantic fluency together with its low resource requirements made it a suitable choice for a hardware-limited setting, despite its comparatively low factual accuracy, as measured by the

SummaCZS metric (Laban et al., 2022). This limitation made it a valuable baseline for later evaluating the effect of retrieval augmentation.

The same dataset was then used to test retrieval strategies. Combinations of semantic search (Cosine Similarity), BM25, and MMR were evaluated using ms-marco-MiniLM-L-6-v2 as an evaluator. Table 6 summarizes raw scores for each method and their combinations.

Table 6. Retrieval Method Evaluation (Raw Scores).
Using cross encoder as an evaluator with k = 10 best documents and equal weights for hybrid retrievals.

Mean (avg)	Mean (avg)	Max (avg)	Std Dev (avg)
Semantic+BM25+MMR	2.9182	6.8420	2.0081
Semantic+BM25	2.7143	6.8335	2.1603
Semantic+MMR	2.4240	6.7895	2.3418
BM25+MMR	2.3796	6.7131	2.3214
Semantic	1.7783	6.7636	2.9004
MMR	0.6364	6.4247	3.3224
BM25	-0.3909	6.0480	4.0369

Source: Own elaboration, 2025.

From these results, it was observed that combining retrieval methods incentivizes stronger performance, with the combination of all three providing the best balance of quality and consistency. Hybrid retrieval was further tested by adjusting ensemble weights to control the influence of each method. Table 7 presents evaluation metrics for different weight configurations.

Table 7. Evaluation metrics for different retrieval weight configurations (semantic search, BM25, MMR).
Using cross encoder for re-ranking, k = 10 best documents and long context reordering.

Configuration	BLEU	BERTScore (F1)	COMET	SummaCZS
[0.1, 0.6, 0.3]	0.116	0.919	0.749	**0.099**
[0.3, 0.6, 0.1]	0.111	0.918	0.748	0.095
[0.6, 0.3, 0.1]	0.109	0.917	0.742	0.094
[0.3, 0.3, 0.3]	0.108	0.917	0.746	0.085
[0.1, 0.3, 0.6]	0.107	0.915	0.735	0.060

Source: Own elaboration, 2025.

Although semantic retrieval is often predominant in modern pipelines, these results show that the system strongly benefits from BM25, likely due to the fact-heavy nature of the dataset. Using the best-performing configuration, generation quality was tested across different quantities of retrieved documents, as shown in Table 8.

Table 8. Evaluation metrics for different k values.
Using the best-performing configuration, with cross encoder re-ranking and long context reordering.

N Best	BLEU	BERTScore (F1)	COMET	SummaCZS
k5	0.107	0.917	0.739	0.069
k10	0.116	0.919	0.749	0.099
k15	0.110	0.917	0.740	**0.133**
k20	0.105	0.917	0.741	0.092

Source: Own elaboration, 2025.

The effect of the number of retrieved documents depends on the LLM's capacity to manage long contexts. In this case, qwen2.5-3b achieved its best results with around 15 documents, after which performance decreased.

Building on these retrieval results, the system was compared to the base qwen2.5-3b-instruct model without retrieval. As shown in Table 9, the system substantially increased the SummaCZS value, confirming improvements in factuality. Remarkably, the 3 billion parameter model surpassed the factual performance of llama-3.1-70b.

Table 9. Metric comparison of qwen2.5-3b-instruct's performance with (k = 15) and without information retrieval.

Configuration	SummaCZS	COMET	BERTScore (F1)
With retrieval	**0.133**	0.740	0.917
Without retrieval	-0.389	0.775	0.920

Source: Own elaboration, 2025.

A slight decrease in adequacy and fluency measured by COMET, as well as a marginal drop in semantic similarity measured by BERT Score, suggests that small models can be overwhelmed by large contexts. This often produced more verbose answers with occasional redundant information. To further investigate, larger models were also tested, as shown in Table 10.

Table 10. Metric comparison of larger models' performance with and without information retrieval.

Configuration	SummaCZS	COMET	BERTScore (F1)
gemini-2.0-flash Retrieval	**0.245**	**0.785**	**0.934**
gemini-2.0-flash No Retrieval	0.195	0.707	0.918
mistral-nemo-12b-instruct Retrieval	**0.263**	**0.603**	**0.891**
mistral-nemo-12b-instruct No Retrieval	-0.188	0.586	0.879

Source: Own elaboration, 2025.

These results confirm that the system significantly improves factual accuracy while slightly enhancing fluency and semantic similarity. The outcome is therefore a more reliable generation pipeline, suitable for delivering accurate information in the museum scenario. It is important to note that these results are not an upper bound, as the evaluation dataset contains questions whose answers are not always present in the vector database.

Finally, retrieval relevance was validated to assess the grounding of answers. Using the ms-marco-MiniLM-L-6-v2 cross encoder, relevance was evaluated across different retrieval sizes with balanced hybrid weights. The results, presented in Table 11, show that Score@1 relevance remains consistent while mean scores decrease and standard deviation increases as more documents are retrieved. Consistent Top1 scores of 6.8 indicate strong alignment with queries.

Table 11. Retrieval evaluation metrics across different k values using balanced weights for hybrid retrieval.

Strategy	Mean (avg)	Score@1 (avg)	Std Dev (avg)
Semantic + bm25 + mmr (k=5)	4.159	6.713	1.695
Semantic + bm25 + mmr (k=10)	2.918	6.842	2.008
Semantic + bm25 + mmr (k=15)	2.064	6.848	2.203

Source: Own elaboration, 2025.

5. Conclusions and Discussion

This project dived deep into the current state of the art of knowledge accessibility in the museum industry, comparing current improvements with traditional methods of the industry and with modern solutions being applied elsewhere for similar purposes. The solution is designed to focus on providing individual, personal and contextual experience for museum visitors.

After development, the solution incorporates an object recognition module to understand what visitors are looking at. It involves making the system respond naturally and accurately to questions while understanding what the visitor is seeing. Finally, the system incorporates a module focused on the inclusion of new accurate knowledge so that the responses can be up to date with the latest relevant information that the museum institution has.

The result of the project includes a previously inexistent feature in the museum industry (real-time sight-aware responses) and significantly improves the factuality of responses generated by LLMs, making responses more reliable.

5.1. Ethical Considerations and Informed Consent

Considerations regarding ethics and informed consent represent essential factors in deploying AI-based systems for museum visitor interaction. Although this work does not detail the full implementation of such processes, it ensures minimization of data retention, exemplified by the *Context Handler*, which deletes all conversational data upon session termination, maintaining user privacy. Transparent communication about data use and responsible human oversight remains critical to foster trust and acceptance. Future efforts should align with legal frameworks such as the General Data Protection Regulation (European Parliament & Council, 2016) and ethical guidance (European Commission, 2019).

5.2 Sociocultural Dimensions

The work can contribute to a more participatory, inclusive and culturally sensitive museum experience that aligns with contemporary values.

Digital humanities encourages responsible use of AI to promote inclusiveness, cultural diversity, and equitable access (Güven et al., 2025). To align with these principles, the system would add speech for more intuitive and inclusive interaction (text-to-speech and speech-to-text), could support multilingual interaction, and would provide plain-language answers calibrated to visitors’ prior knowledge and language proficiency. At ingestion, museum staff are expected to curate and review materials to mitigate bias by diversifying documents, recording sources and confidence levels, and incorporating community input before content enters the vector store.

Critical museology urges moving beyond static museum narratives to foster participatory, dialogic, and reflective engagement, which aligns with this approach by asking the system to support active visitor involvement rather than passive consumption (Boulakal & Hadi, 2025; Lundgren et al., 2019). Designing conversational interactions that invite reflection, and diverse narratives would enhance visitor empowerment and cultural awareness (Damiano et al., 2022). The system should incorporate dialogue prompts that encourage visitors to reflect and share perspectives, it could enable feedback mechanisms to collect input for iterative improvement and would offer responses that draw on multiple sources to present varied interpretations. Together, these features would turn guidance into a dynamic, inclusive dialogue that respects cultural plurality and promotes critical thinking.

5.2. Future Technical Directions

At the interaction layer, response quality could also improve through prompt template analysis and parameter tuning for sampling/decoding parameters. The *Context Service* could integrate depth estimation models and considering inside-frame object position could improve relevance of the detected artworks on, while a more flexible object detection architecture would also allow moving art pieces between rooms without reducing system performance.

In the *Vector Knowledge Ingestion Service*, migrating from Chroma to a managed, production-grade vector database, like Pinecone, could increase efficiency; ingestion could also enrich documents with metadata automatically, and a specialized OCR model would widen the range of usable PDF sources.

Retrieval quality could improve though agentic chunking for semantic relevance and specialized models for embedding and reranking in art/history. Response quality could also benefit from prompt engineering, parameter tunning, or finetuning. Techniques like ReAct could also help include functionalities such as accessing the internet, avoiding retrieval if not needed, or searching by keywords.

Finally, expanding the model with more art pieces and information could provide a better simulation of a real scenario.

References

Ask Mona. (2025). *Ask Mona*. https://www.askmona.fr

Breitner, A. R., & Bandung, Y. (2024). Development of visitor interest detection and tracking system in the museums. *Journal of Sustainable Engineering: Proceedings Series, 2*(1), 7–12. https://doi.org/10.35793/joseps.v2i1.1279

Bruch, S., Gai, S., & Ingber, A. (2023). An analysis of fusion functions for hybrid retrieval. *ACM Transactions on Information Systems, 42*(1), 1–35. https://doi.org/10.1145/3596512

Boulakal, F., & Hadi, W. M. E. (2025). Cultural & Knowledge Spaces: the Immersive Museums as a Challenge for KO and the Digital Humanities. *Informatio*, *30*(1), e205. https://doi.org/10.35643/info.30.1.10

Bu, F., Wang, Z., Wang, S., & Liu, Z. (2025). An investigation into value misalignment in LLM-generated texts for cultural heritage. *arXiv*, *1*. https://doi.org/10.48550/arXiv.2501.02039

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 335–336). ACM. https://doi.org/10.1145/290941.291025

Cetinić, E., Lipić, T., & Grgić, S. (2018). Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications, 114*, 107–118. https://doi.org/10.1016/j.eswa.2018.07.026

CVAT. (2025). *CVAT: Computer Vision Annotation Tool* [Computer software]. https://github.com/opencv/cvat

Damiano, R., Kuflik, T., Wecker, A. J., Striani, M., Lieto, A., Bruni, L. E., Kadastik, N., & Pedersen, T. A. (2022). Exploring values in museum artifacts in the SPICE project: A preliminary study. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22 Adjunct)* (pp. 391–396). Association for Computing Machinery. https://doi.org/10.1145/3511047.3537662

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. https://aclanthology.org/N19-1423/

Du, X., Zheng, G., Wang, K., Feng, J., Deng, W., Liu, M., Chen, B., Peng, X., Ma, T., & Lou, Y. (2024). Vul-RAG: Enhancing LLM-based vulnerability detection via knowledge-level RAG. *arXiv*, *1*. https://doi.org/10.48550/arXiv.2406.11147

European Commission. (2019). *Ethics guidelines for trustworthy AI*. Independent High-Level Expert Group on Artificial Intelligence set up by the European Commission. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Parliament & Council. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Official Journal of the European Union, L119, 1–88. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Fortuna-Cervantes, J. M., Soubervielle-Montalvo, C., Puente-Montejano, C. A., Pérez-Cham, O. E., & Peña-Gallardo, R. (2024). Evaluation of CNN models with transfer learning in art media classification in terms of accuracy and class relationship. *Computación y Sistemas, 28*(1), 233–244. https://doi.org/10.13053/cys-28-1-4895

Güven, Ç., Alishahi, A., Brighton, H., Nápoles, G., Olier, J. S., Šafář, M., Postma, E., Shterionov, D., De Sisto, M., & Vanmassenhove, E. (2025). *AI in support of diversity and inclusion*. *arXiv*. https://doi.org/10.48550/arXiv.2501.09534

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

LangChain. (2025). *LangChain* [Computer software]. https://github.com/langchain-ai/langchain

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15339–15353). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.818

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision (ECCV 2014)* (pp. 740–755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2016). SSD: Single shot multibox detector. In *European Conference on Computer Vision (ECCV 2016)* (pp. 21–37). Springer. https://doi.org/10.1007/978-3-319-46448-9_2

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://aclanthology.org/2024.tacl-1.9/

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., & Zhang, L. (2024). Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. *arXiv*. https://arxiv.org/abs/2303.05499

Loffredo, R., & De Santo, M. (2024). Using ontologies for LLM applications in cultural heritage. In *CEUR Workshop Proceedings* (Vol. 3865). https://ceur-ws.org/Vol-3865/06_paper.pdf

Louvre Museum. (2025). Collections site JSON documentation [Website]. Retrieved January 20, 2025, from https://collections.louvre.fr/en/page/documentationJSON

Lundgren, L., Stofer, K., Dunckel, B., Krieger, J., Lange, M., & James, V. (2019). Panel-based exhibit using participatory design elements may motivate behavior change. *Journal of Science Communication*, *18*(02), A03. https://doi.org/10.22323/2.18020203

Luo, C., Li, X., Wang, L., He, J., Li, D., & Zhou, J. (2018). How does the data set affect CNN-based image classification performance? In *2018 5th International Conference on Systems and Informatics (ICSAI)* (pp. 361–366). IEEE. https://doi.org/10.1109/ICSAI.2018.8599448

Ma, X., Gong, Y., He, P., Zhao, H., & Duan, N. (2023). Query rewriting for retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 5303–5315). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.322

McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform manifold approximation and projection. *Journal of OpenSource Software, 3*(29), 861. https://doi.org/10.21105/joss.00861

Meyer, L. S., Engel Aaen, J., Tranberg, A. R., Kun, P., Freiberger, M., Risi, S., & Løvlie, A. S. (2024). Algorithmic ways of seeing: Using object detection to facilitate art exploration. In *CHI '24: Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642157

Nubart. (2025). *Nubart* [Mobile application]. https://www.nubart.eu

Ollama. (2025). *Ollama* [Computer software]. https://github.com/ollama/ollama

Redmon, J., & Farhadi, A. (2017). YOLO9000: Better, faster, stronger. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 6517–6525). IEEE. https://doi.org/10.1109/CVPR.2017.690

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*. https://arxiv.org/abs/1506.01497

Robben, H. [Wanderlust Travel Videos]. (2019, May 31). Louvre Museum Paris – Mona Lisa – walking tour | 4K [Video]. YouTube. https://youtu.be/6vuFh6NNa70

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

Rosa, G. M., Bonifacio, L. H., Jeronymo, V., Abonizio, H. Q., Fadaee, M., Lotufo, R. A., & Nogueira, R. (2022). *In defense of cross-encoders for zero-shot retrieval*. arXiv. https://arxiv.org/abs/2212.06121

Sahoo, P. K., Sharma, N., Mehta, P., Kumar, S., Garg, A., … & Pratama, M. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv*. https://doi.org/10.48550/arXiv.2402.07927

Smartify. (2025). *Smartify* [Mobile application]. https://smartify.org

Smith, J. K., & Smith, L. F. (2001). Spending time on art. *Empirical Studies of the Arts, 19*(2), 229–236. https://doi.org/10.2190/5MQM-JWH6-V2P4-7DLK

Smith, J. K., Smith, L. F., & Tinio, P. P. L. (2017). Time spent viewing art and reading labels. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 77–85. https://doi.org/10.1037/aca0000049

Smith, B., & Troynikov, A. (2024, July 3). Evaluating chunking strategies for retrieval (Chroma Technical Report). Chroma. https://research.trychroma.com/evaluating-chunking-strategies-in-retrieval

Springmann, U., Lüdeling, A., & Ernst, F. (2017). OCR of historical printings with an application to building diachronic corpora: The RIDGES herbal corpus. *Digital Humanities Quarterly, 11*(2). http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html

United Nations General Assembly. (2015). *Transforming our world: The 2030 agenda for sustainable development* (A/RES/70/1). https://sustainabledevelopment.un.org/post2015/transformingourworld/publication

Vastakas, L. (2024). *Cultural heritage search with large language models: Enhancing the discoverability of cultural heritage artifacts through large language model-based search systems* [Master's thesis, Linnaeus University]. DiVA portal. https://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-132431

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. *arXiv*. https://arxiv.org/abs/2405.14458

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *NeurIPS 2020*. https://proceedings.neurips.cc/paper/2020/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Wu, J., Zhu, J., Qi, Y., Chen, J., Xu, M., Menolascina, F., & Grau, V. (2024). Medical graph RAG: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv*. https://doi.org/10.48550/arXiv.2408.04187

Yano, T., & Kang, M. (2008). Taking advantage of Wikipedia in natural language processing. Language Technologies Institute, Carnegie Mellon University. https://www.cs.cmu.edu/~taey/pub/wiki.pdf