



PDF Download  
3756016.pdf  
15 January 2026  
Total Citations: 0  
Total Downloads: 517

 Latest updates: <https://dl.acm.org/doi/10.1145/3756016>

RESEARCH-ARTICLE

## Knowledge Graphs vs. Large Language Models: Competitors or Partners in Supporting Virtual Museums

**IVA VASIC**, University of Fribourg, Fribourg, FR, Switzerland

**HANS GEORG FILL**, University of Fribourg, Fribourg, FR, Switzerland

**RAMONA QUATTRINI**, Marche Polytechnic University, Ancona, AN, Italy

**ROBERTO PIERDICCA**, Marche Polytechnic University, Ancona, AN, Italy

Open Access Support provided by:

Marche Polytechnic University

University of Fribourg

Published: 18 December 2025

Online AM: 29 July 2025

Accepted: 09 July 2025

Revised: 31 May 2025

Received: 20 May 2024

[Citation in BibTeX format](#)

# Knowledge Graphs vs. Large Language Models: Competitors or Partners in Supporting Virtual Museums

**IVA VASIC**, Department of Construction, Civil Engineering and Architecture, Marche Polytechnic University, Ancona, Italy and Department of Informatics, University of Fribourg, Fribourg, Switzerland  
**HANS-GEORG FILL**, Department of Informatics, University of Fribourg, Fribourg, Switzerland  
**RAMONA QUATTRINI** and **ROBERTO PIERDICCA**, Department of Construction, Civil Engineering and Architecture, Marche Polytechnic University, Ancona, Italy

---

Virtual museums are factual means for the dissemination and documentation of Cultural Heritage (CH) content. They are suitable environments for the semantic annotation of artifacts and automatic virtual guides. To this end, we identify and compare Traditional (ontology-based), Large Language Model (LLM)-extended, and LLM-pure methods for the semantic information strategies of digital CH. The traditional method is described through an application prototype, while the methods that involve LLM are tested experimentally. To investigate the integral tasks related to LLMs, our experiments include (i) semantic annotation using the CIDOC Conceptual Reference Model (CRM) and Knowledge Graph (KG) generation with LLMs for a painting sample, and (ii) painting ranking relying solely on LLMs using catalog descriptions as input. The experiments demonstrate the potential of these methods to enhance artwork interpretation, description, and refinement of the results. Based on the relevant literature on traditional semantic annotation and conducted experiments with LLMs, a combination of ontologies and LLMs may provide an optimal approach, as it offers the accuracy of structured knowledge while providing a tool that interprets these elements into natural language and vice versa. Relying solely on LLMs may be risky due to the lack of domain-specific knowledge in the training data of LLMs, whereas traditional methods demand expertise in a specific domain and are more time-consuming. Our approach shows potential in use cases such as guiding museum visitors to artifacts that match their interests, assisting museum curators with documentation, or helping CH researchers identify similarities in artifact collections.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Theory of computation** → **Semantics and reasoning**; • **Human-centered computing** → *Human computer interaction (HCI)*; • **Information systems** → *Digital libraries and archives*;

Additional Key Words and Phrases: Large Language Model, Cultural Heritage, Semantic Annotation, Knowledge Graphs

---

All authors contributed equally to this research.

This research was funded by Virtual Immersion in Territorial Arts (V.I.T.A.) project, funded by POR MARCHE FESR 2014/2020 8OS 22—ACTION 22.1 INTERVENTION 22.1.1—Support for the development and enhancement of social enterprises in areas affected by the earthquake.

Authors' Contact Information: Iva Vasic (corresponding author), Department of Construction, Civil Engineering and Architecture, Marche Polytechnic University, Ancona, Italy and Department of Informatics, University of Fribourg, Fribourg, Switzerland; e-mail: iva.vasic@unifr.ch; Hans-Georg Fill, Department of Informatics, University of Fribourg, Fribourg, Switzerland; e-mail: hans-georg.fill@unifr.ch; Ramona Quattrini, Department of Construction, Civil Engineering and Architecture, Marche Polytechnic University, Ancona, Italy; e-mail: r.quattrini@univpm.it; Roberto Pierdicca, Department of Construction, Civil Engineering and Architecture, Marche Polytechnic University, Ancona, Italy; e-mail: r.pierdicca@staff.univpm.it.



This work is licensed under [Creative Commons Attribution-NonCommercial-NoDerivatives International 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/).

© 2025 Copyright held by the owner/author(s).

ACM 1556-4711/2025/12-ART61

<https://doi.org/10.1145/3756016>

**ACM Reference format:**

Iva Vasic, Hans-Georg Fill, Ramona Quattrini, and Roberto Pierdicca. 2025. Knowledge Graphs vs. Large Language Models: Competitors or Partners in Supporting Virtual Museums. *ACM J. Comput. Cult. Herit.* 18, 4, Article 61 (December 2025), 28 pages.

<https://doi.org/10.1145/3756016>

---

## 1 Introduction

**Knowledge Graphs (KGs)** and **Large Language Models (LLMs)** have recently met at the intersection of artificial intelligence and knowledge representation [38, 45], providing new opportunities for enhancing the user experience in virtual environments. Virtual museums play a pivotal role in providing immersive and interactive experiences for visitors [4, 44]. Thereby, new challenges arise in the field. For instance, the ability of virtual guides to analyze the interests of visitors and provide relevant content is still uncertain. Such analysis includes user behavior assessment and requirements, and matching museum content accordingly. The ecosystem of many virtual museums is intricate and comprises a variety of digitized exhibits, making it difficult to organize data.

The three main architectural components for a virtual museum creation are Content Production, Content Management, and Content Visualization [56]. The final output is a digital 2D or 3D model of the contents, and sometimes virtual tours. From the standpoint of visitors, digital content should be entertaining and offer a concise presentation. This is not always the case, as it tends to lead to an excessive amount of scattered data that leaves users uncertain of where to direct their attention. However, obtaining relevant knowledge from the museum collection databases not only depends on the developed system and the quality of the digitized data, but also on the quality of the content information. Apart from the spatial features, digital exhibits are annotated with *semantic information* that should be understood by the machine and presented to the user comprehensively.

Traditionally, *semantic annotation* can be accomplished by adding machine-processable information to the spatial representation of artworks. FAIR Data Principles serve as well-established guidelines for improving data findability, accessibility, interoperability, and reuse, making the annotations interpretable by machines [2]. Such semantic information can take the form of KGs based on the ontology schema. In the context of **Cultural Heritage (CH)** and museums, the **International Committee for Documentation (CIDOC) Conceptual Reference Model (CRM)** is often used as a foundational framework for semantic information modeling, offering concepts and guidelines for documentation. CRM is closely aligned with the **Resource Description Framework (RDF)**, while maintaining logical independence, and allows for integration with other ontologies and respective implementations [18]. Note that CIDOC-CRM is often integrated with other ontologies when needed, depending on the use case, but we do not explore this aspect in our research. Modeling KGs with an ontology poses a challenge, as it requires specific skills and familiarity with various aspects of the Semantic Web framework.

Since semantic annotation may often be demanding in both time and expertise, using automated approaches for ontology construction through **Natural Language Processing (NLP)**, for instance, offers a potential alternative. Such an example is DBpedia, where the structured content is extracted from the Wikipedia resources [14]. Similarly, FRED [23] is a machine reader that parses natural language text and transforms it into structured linked data. On the other hand, the advent of LLMs provides new opportunities to support traditional techniques within semantic information strategies, including modeling, reasoning, querying, and representing data, due to their ability to process natural language. For instance, they are capable of processing complex textual input [5], while others integrate vision image capabilities into LLMs [40]. However, relying only on such stochastic-based tools cannot ensure sufficient accuracy compared to ontology-based methods. This discrepancy prompted our exploration of various approaches to semantic information strategies and influenced us to assess more thoroughly

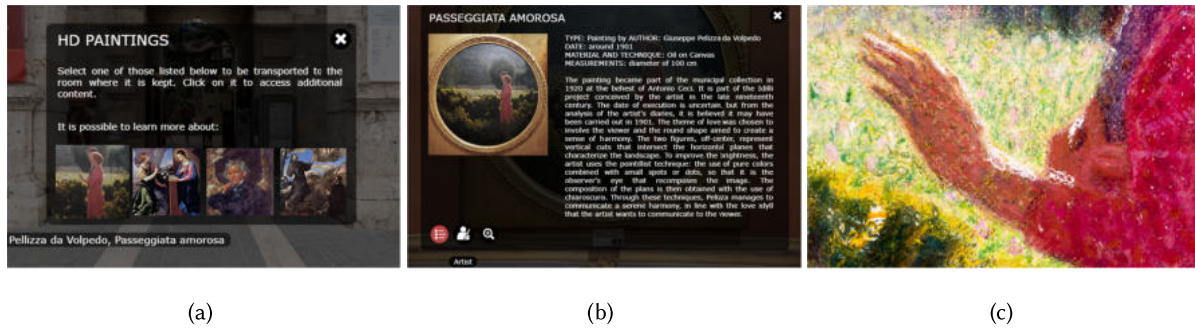


Fig. 1. The virtual tour interface of the Civic Art Gallery in Ascoli: (a) “HD Paintings” panel with the available digitized paintings, (b) catalog description of the painting “Amorous Walk” by Giuseppe Pellizza da Volpedo, and (c) 2D detail-oriented representation of the observed part of the same painting.

some of the key operations within the workflow that strives to connect a user requirement and virtual museum content.

### 1.1 Motivation

Our motivation stems from previous work involving the digitization and virtualization of the Civic Art Gallery of Ascoli in Italy.<sup>1</sup> The project is designed for visitors to explore the museum, presenting numerous pieces of art in a virtual panoramic tour. Figure 1 depicts some of the functionalities, including exhibit descriptions and their 3D visualization. However, the descriptions correspond to those from catalogs as they have been provided by the museum curators in a natural language format. The digitized paintings are represented as high-resolution images created utilizing the photogrammetry technique [10], available for observing their millimetric details. However, the descriptions of the content are not structured according to any ontology, nor are the visual parts of the digitized content embedded with semantic information. To tackle these problems, we have theoretically investigated different solutions in the literature and provided some solutions on how this process can be supported with LLMs.

Based on the former solutions in CH and new LLM technologies, we identified three potential methods for semantic information strategies with highlighted main features (see Figure 2):

- (1) *Traditional*—involves semantic annotation using ontologies and searching the content into KGs formed by experts.
- (2) *LLM-extended*—involves LLMs in KG creation from the natural language descriptions, like those in catalogs. Similar to the traditional method, it incorporates querying KGs.
- (3) *LLM-pure*—relies solely on LLMs for the interpretation of catalog descriptions and visitor requirements in natural language.

The Traditional method reflects expert-driven practices which we implemented through the simple tool to demonstrate its general workflow and potential capabilities. This method is not empirically tested, as it relies on expert input and serves primarily as a conceptual baseline to contextualize the comparison. On the other hand, the LLM-extended and LLM-pure methods are empirically tested, highlighting the transformative potential of automation in semantic annotation tasks, such as generating KGs and ranking paintings based on a given input. Key parameters—user experience, efficiency, accuracy, data richness, and input precision—are extracted throughout the research process and used as variables in the comparative analysis. This comparative analysis lays the foundations

<sup>1</sup><https://dhekalos.it/tour/iva/pinacoteca/index.html>.

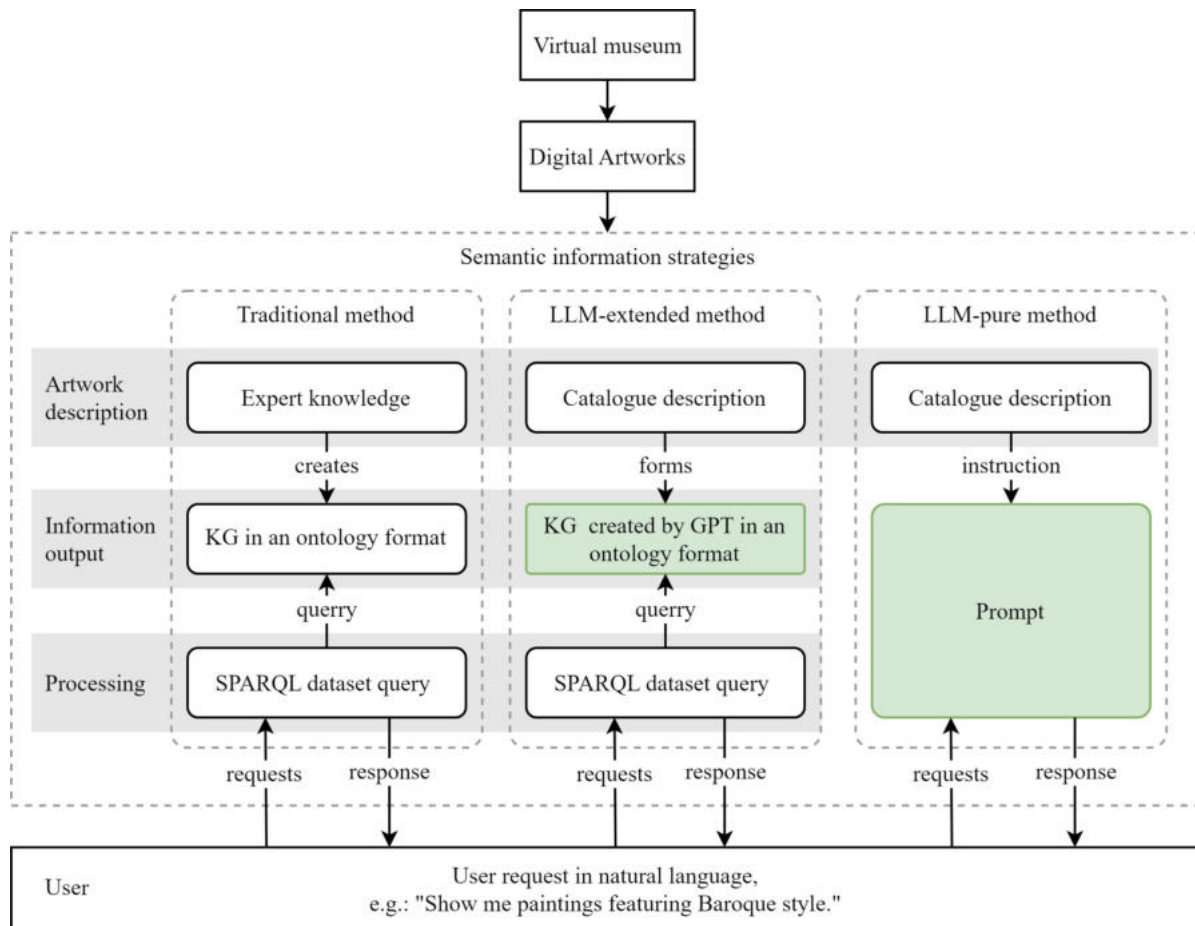


Fig. 2. General workflows for each method: Ontology-based approach on the left, a combination of ontology and LLMs in the middle, and an LLM-pure approach on the right. Colored boxes indicate the usage of LLMs.

for understanding in which ways LLM technologies can enhance or challenge semantic annotation in museum workflows. It also offers insights that may shape future approaches to CH data management and interpretation.

The remainder of the article is organized as follows. In Section 2, we briefly introduce the main concepts, including KGs based on a selected ontology, and LLMs. Section 3 describes related work in the semantic information strategies framework and various LLM applications. Section 4 describes the specific component of each of the three identified methods, as aligned with our research goal. Experiments and validations focusing on automation and LLM-based strategies are included in Section 5. The comparison of the three methods is presented in Section 6, followed by a conclusion and potential ideas for future works in Section 7.

## 2 Foundations

This section introduces the key concepts for our study. Our three methods involve traditional KG creation using ontology and/or recently developed generative artificial intelligence tasks. LLMs have recently been introduced and considered “black boxes.” Ontology within KGs, on the contrary, belongs to a logic-based framework that provides clear structure and reasoning capabilities for representing knowledge. While these areas overlap, we

emphasize the importance of establishing their main concepts as a foundation, recognizing that their combined use is still an evolving field. Therefore, we explain KGs and their role in structuring semantic knowledge for specific domains, particularly in the context of CH and CIDOC-CRM ontology. Consequently, a brief explanation of the emerging LLM technology is presented.

*KGs and CIDOC-CRM Ontology.* An ontology is an explicit representation of the specific domains [25]. In the context of KGs, such semantic knowledge is encoded using the RDF framework, which enables structured and machine-readable data exchange on the web. However, sole usage of RDF is not enough in all cases and that is where RDF vocabulary description language (RDFS) is useful. RDFS is defined as a semantic extension of RDF, defining classes, properties, and their interconnections. Ontologies can be used to define and reason about semantics used for the KG creation, i.e., to define nodes and edges which are the main structures that represent entities and their relationships within KGs [27]. In this work, we revert to the CIDOC-CRM as a commonly used object-oriented ontology<sup>2</sup> for the semantic artwork annotations. It is an international standard (ISO 21127:2014) and elaborated for the CH domain, making it a widely recognized and adopted framework for organizing and describing CH information. It involved semantic content reorganization from various databases of information collected from museums, archives, and libraries [17]. In the context of our research, we use the RDF data [12] corresponding to version 7.1.2 of the CIDOC-CRM.<sup>3</sup> The resulting KGs leverage both the ontological rigor of CIDOC-CRM and the flexibility of graph-based representations to support machine-interpretable semantic reasoning over artwork descriptions.

*LLMs.* Since the release of ChatGPT [39], numerous commercial and open source LLMs have become available. A particular advancement seen in 2023 is the capability of LLMs to process also visual content, both in terms of generating as well as interpreting it [13]. The core of GPT-4 [41], which we leverage in this research, is a Transformer-style model [54] relying solely on a self-attention mechanism instead of using a Convolutional Neural Networks or **Recurrent Neural Networks (RNNs)**. The advent of this approach notably improved NLP, becoming widely used for tasks such as translation, question answering, code generation, and text summarizing. Given the outstanding capabilities of the GPT models, we chose the GPT-4 and the OpenAI API for the KG creation from natural language descriptions and ranking paintings based on user requirements and predefined catalog descriptions. The extensive knowledge used for training GPT results in its ability to generate a variety of formal CIDOC-CRM ontology statements, each intended for describing a specific object, event, placement, or symbolism of a painting.

### 3 Related Work

In this section, we present the relevant work with respect to the processes within the three identified methods. Therefore, we divide this paragraph into three parts that are interconnected in the general overview of the semantic information strategy. First, we investigate the semantic annotation of CH artifacts that has been usually performed manually using domain ontologies. Next, we introduce the literature, including an automated semantic annotation. Finally, LLM-pure tasks are described within the works of various backgrounds, drawing on some of the use cases identified by Pan et al. [42].

#### 3.1 Manual Semantic Annotation Using Ontologies

The manual annotation process mostly encompasses representing and processing knowledge. Knowledge representation is modeled using existing ontologies and aligning or extending them to correspond to the specific requirements. The subsequent step is representing this knowledge to the end users by querying the constructed database. Related works usually incorporate annotation tools as the final result of this process.

<sup>2</sup><https://cidoccrm.org/>, last accessed: 23 January 2025.

<sup>3</sup>[https://cidoc-crm.org/html/cidoc\\_crm\\_v7.1.2.html](https://cidoc-crm.org/html/cidoc_crm_v7.1.2.html), last accessed: 22 January 2025.

Digital artworks, represented in the 2D format can be annotated using different tools in the same way as any kind of image. Schreiber et al. [48] developed an annotation tool for semantic annotation and searching the photographs. The ontology in RDFS is constructed manually in a software application and used in the developed annotation tool for annotation and storing in an RDF file. The query method is also based on the created ontology, i.e., it queries the database and finds the values that match these in the query. Schreiber et al. [47] present semantic annotation of CH artworks and a facility for searching the resulting graph through the MultimediaN E-Culture demonstrator using keywords. To facilitate the search process and initially construct a semantic graph, the semantic annotation involved translating vocabularies, aligning metadata schema, enriching metadata, and aligning vocabulary(ies). SPARQL and graph-based APIs are developed for the background search process. A web-based annotation tool for documenting relationships between 3D digital CH artifacts is developed by Yu and Hunter [58]. The project is intended both for curators and scholars. While this work does not address the issue of delivering relevant content to users during the virtual tour, it serves as an initial step within the framework of the traditional method we have outlined. “Museum Finland” [9] is a semantic portal developed for the Finish museums for publishing rich semantic content within the Semantic Web framework. Seven domain ontologies are employed for the semantic annotation and the content is sourced from different databases. In the case of “Museum of the Person” introduced by Araújo et al. [1], ontology is created from the **eXtensible Markup Language (XML)** documents and transformed into a tailored OntoMP derived from CIDOC-CRM, Friend of a Friend, and DBpedia. Semantic technologies have also been investigated by Vlachidis et al. [55] in the case of the Archaeological Museum of Tripoli following FAIR principles and using CIDOC-CRM ontology. Unstructured catalog data underwent mapping to CIDOC-CRM format, manually extracting relevant information given the small museum’ scale.

### 3.2 Automated Semantic Information Strategies

In an ontology, automatic annotation associates low-level features in an image (such as pixel colors, shapes, texture, and so on) with semantic meanings through the process of feature extraction [30]. This study introduced an automatic framework for image annotations using ontologies derived from multiple information sources. “Low-level” image feature bridged with the “high-level” semantic annotations has been described by Chai et al. [7] for the automatic content-based image retrieval. Sezen et al. [49] demonstrate natural language description generation combining neural networks to detect objects and ontology to refine and specialize them.

In the context of CH and automated semantic information strategies, knowledge extraction is usually tackled as a predominant field of study. Felicetti et al. [20] presented a tool that generates metadata in a machine-understandable language such as RDF, from the natural-language content. The syntax and semantics are based on CIDOC-CRM ontology. The parts of the provided input text are highlighted with the corresponding CIDOC-CRM classes. Only some of them are taken into consideration instead of the entire set of classes that constitute the CIDOC-CRM ontology. In the study of Dou et al. [19], Chinese intangible CH was represented as a KG using knowledge extraction techniques to describe patterns and characteristics for knowledge management. Manual tagging of the text was performed in the first step while the entity and relationships extraction based on CIDOC-CRM was executed with deep learning methods. Metilli et al. [36] developed a tool for extracting formal narratives from the text format. The two requirements for accomplishing this task are elaborated in this study. Event detection and classification from the text are performed using software based on RNNs. Schleider et al. [46] proposed an image-based retrieval tool that leverages a KG describing the silk textile production as well as rules formulated by experts in this domain. The method developed by Navigli et al. contributes to automatic ontology enrichment using CIDOC-CRM as a domain core ontology [37]. The natural language descriptions of the Art and Architecture Thesaurus glossary of art are automatically extracted and converted into formal (OWL) definitions. Annotation is also explored with the video formats. The semi-automatic annotation web-based Traditional Dance Annotation Tool was presented by Lagrue et al. [32]. Integrating the dance ontology, it illustrates the process of annotating intangible CH dance videos to help dance experts annotate videos of CH dances.

In the realm of paintings, we also point out a European project “Saint George On a Bike”<sup>4</sup> that investigated the meaning of paintings. It yielded several publications on discovering the semantics of the paintings using artificial intelligence by creating metadata in an interoperable format. As a result, Wu et al. [57] explore the potentials and limitations of machine learning in automated metadata annotation, while Marinescu et al. [35] combine deep learning approaches and semantic metadata to enhance object detection in paintings, focusing on the role of time contexts.

### 3.3 LLMs Applications

Both manual and automatic semantic annotation tasks using ontology could be significantly enhanced by incorporating LLMs into the process. For example, the LLMs4OL approach by Babaei Giglou et al. [3] investigates the use of LLMs for Ontology Learning through NLP tasks, demonstrating potential improvements in handling complex language patterns for ontology construction. In addition to enhancing annotation processes, LLMs also play a crucial role in extracting information from tabular data [16, 50] and can aid in KG construction, which is beneficial during the semantic interpretation task. Chen and Bertozzi [8] introduce AutoKG methods which leverage LLMs to generate KGs directly from blocks of text.

With the advancement of the tools designed to understand and generate content such as ChatGPT<sup>5</sup> and Gemini (formerly BARD),<sup>6</sup> several applications are worth mentioning. Fill et al. [22] demonstrated the creation and understanding of conceptual models using ChatGPT, which was later extended in a framework for using conceptual models as a general interface to LLMs [21]. Mai et al. [34] explored LLMs for the enhancement of teaching and research in education. Drew Thomas [29] used GPT-4o to generate a short description of the Bible illustrations as part of the Ornamento project. Lewis and White [33] exploit smaller models learned from ChatGPT for a virtual museum tour guide incorporating an avatar that gives information about the museum and exhibits. The authors utilize ChatGPT indirectly by employing it to generate training data. This data is then used to train two distinct models, with their performances subsequently compared against that of ChatGPT.

## 4 Our Approach

In this section, we introduce the three methods for semantic information strategies by highlighting the design of their specific processes. Recall that the traditional method is primarily analyzed theoretically, based on insights from the literature and accumulated experience. To support this analysis, we developed a web-based annotation tool to illustrate the functionalities aligned with the expert-driven annotation process. It serves as a conceptual framework rather than a fully tested solution and it is intended for manual annotation by experts, which is beyond the scope of this study. LLM-based tasks are evaluated in the later sections through experimentation. For a brief overview of the characteristics of different methods—such as the use of manual annotation, expert validation, implementation tools, study inclusion criteria, and evaluation basis—see Table 1.

To complement Figure 2, which provides a high-level overview of the methods, we introduce Figure 3 to show a detailed overview of how the integral processes are designed and implemented in our study. It further serves as a conceptual complement to the detailed descriptions of all methods provided in the subsequent sections.

### 4.1 Traditional Method

Usually, an entire piece of artwork is described by the expert and the content is structured according to different standards including General International Standard Archival Description [28], Getty [24], Iconclass [11], and museum institution management schemes. On the other hand, there are initiatives to bring this workflow to a higher level. For example, LinkedArt, a community of museums and CH professionals, developed a data model for

<sup>4</sup><https://saintgeorgeonabike.eu/>, last accessed: 22 January 2025.

<sup>5</sup><https://chat.openai.com/>, last accessed: 22 January 2025.

<sup>6</sup><https://gemini.google.com/app>, last accessed: 22 January 2025.

Table 1. Characteristics of Different Methods

Characteristic	Traditional	LLM-Extended	LLM-Pure
Manual Annotation	Yes	No	No
Expert Validation	No	No	No
Implementation	Web-Based Annotation Tool	OpenAI API	OpenAI API
Study Inclusion	Annotation by Modeling KG	Annotation by Generating KGs	Ranking Based on Annotation Statements
Evaluation Basis	Annotation Tool and Literature	KG Accuracy Using CQs	User Study and Variance Measure
Environment	Civic Art Gallery of Ascoli–Virtual Tour	Getty Collection Online Repository <sup>a</sup>	Getty Collection Online Repository

CQs, Competency Questions. <sup>a</sup> <https://www.getty.edu/art/collection/>, last accessed: 22 January 2025.

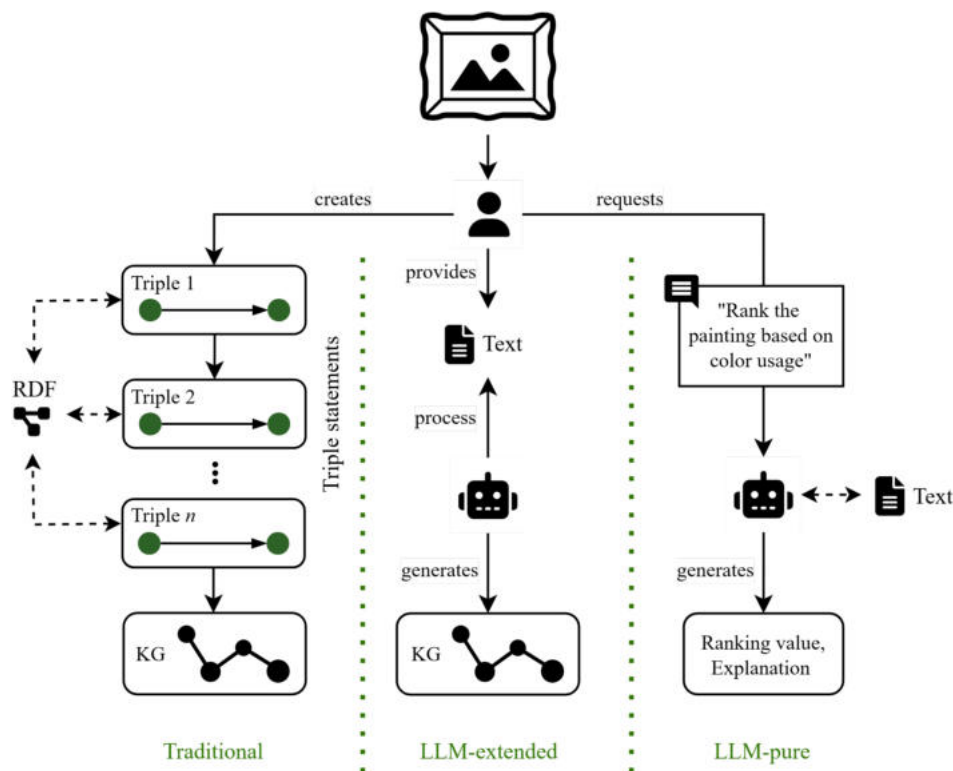


Fig. 3. Overview of the workflow design and key processes for each method.

describing CH and interacting with it using CIDOC-CRM as a core ontology and Linked Open Data techniques [15]. We developed a simple annotation prototype that serves as a foundational tool for exploring the application of our research concepts in practice. Although we approach this method theoretically, the tool allows the selection of the CIDOC-CRM entities and assigning them to the selected regions in the panoramic view. Based on content selected by a user, the annotation is typically accomplished using RDF-based ontologies, which codify knowledge

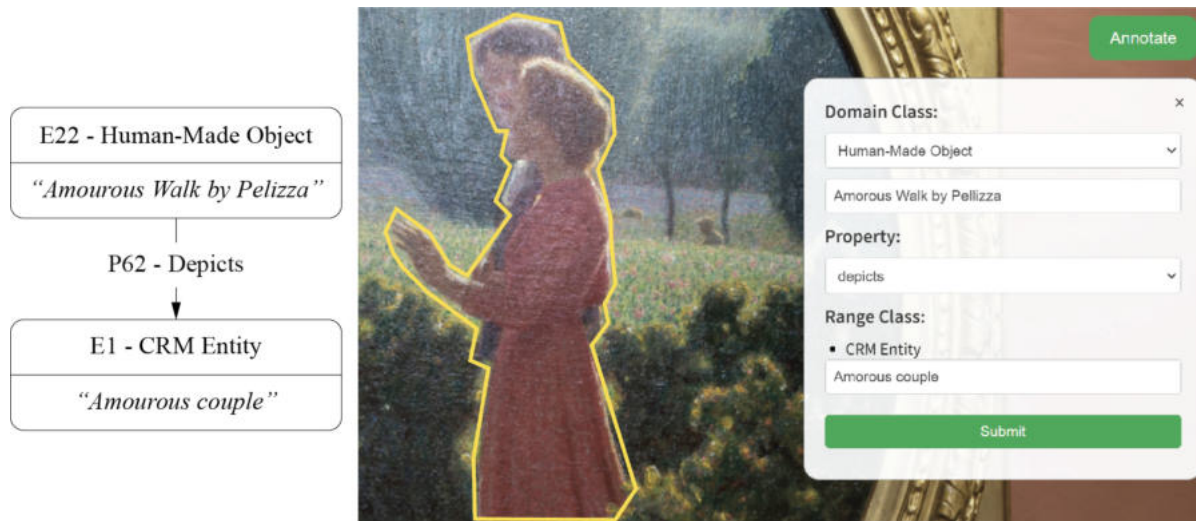


Fig. 4. A prototypical user interface with the annotation panel containing a CIDOC-CRM example that describes a couple in the painting “Amorous Walk” by Giuseppe Pellizza da Volpedo. Its formal representation is shown on the left-hand side.

in subject-predicate-object triples. The interface of the virtual tour and the panel for the annotation is depicted in Figure 4. On the left side is a conceptual formal representation obtained upon the submission button. Panoramic images represent the Civic Art Gallery of Ascoli in Italy.

The annotation tool is developed using JavaScript programming language, with the integration of the Three.js<sup>7</sup> library within a Node.js<sup>8</sup> environment. It is foremost intended for conceptualizing and approximating the process used in traditional approaches. In our tool, the user assigns a triple using CIDOC-CRM ontology directly on the panoramas. Such input is then connected with the image and 2D coordinates of the selected polygonal selection. This information is obtained using the method from [53]. This is explained conceptually in Figure 5, depicting the flow of RDF querying and selecting entities in CIDOC-CRM RDF data. The RDF data is based on the following schema:

```
<rdfs:Class rdf:about="">
  <rdfs:label xml:lang="">""</rdfs:label>
  <rdfs:comment>""</rdfs:comment>
  <rdfs:subClassOf rdf:resource="" />
</rdfs:Class>
<rdf:Property rdf:about="">
  <rdfs:label xml:lang="">""</rdfs:label>
  <rdfs:comment>""</rdfs:comment>
  <rdfs:domain rdf:resource="" />
  <rdfs:range rdf:resource="" />
  <rdfs:subPropertyOf rdf:resource="" />
  <owl:inverseOf rdf:resource="" />
```

<sup>7</sup><https://threejs.org/>, last accessed: 22 January 2025.

<sup>8</sup><https://nodejs.org/en>, last accessed: 22 January 2025.

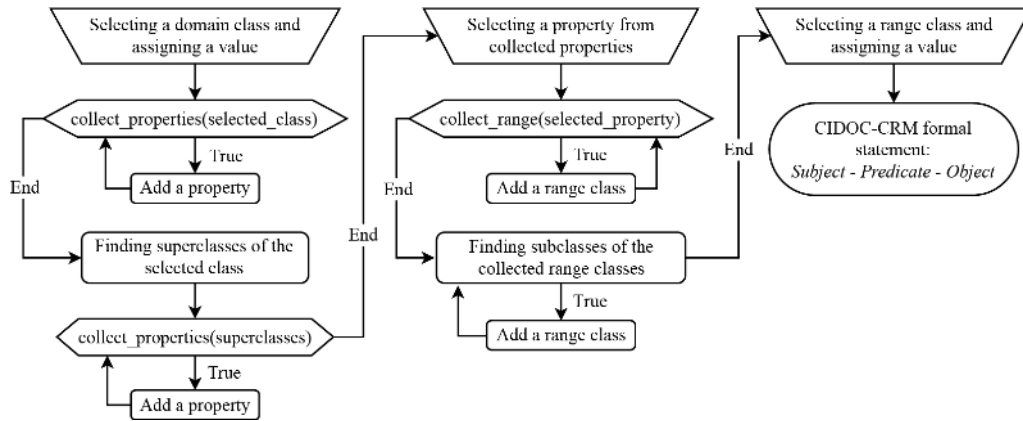


Fig. 5. The process of CIDOC-CRM formal statements construction within the traditional (ontology-based) annotation method through RDF data querying and collecting.

</rdf:Property>

A step-by-step algorithm of our tool is depicted in Figure 5, illustrating the example of the process that should be performed when the LLM or other artificial intelligence-powered technology is not utilized. This is a general demonstration of how the expert would model KG without artificial intelligence assistance, and how the machine would, in response, query the corresponding properties and range classes.

## 4.2 LLM-Extended Method

The main idea of this method is to construct a KG of CIDOC-CRM classes and properties from the Getty Collection catalog descriptions using the GPT model. As can be seen in Figure 2, the *Processing* task ideally involves a SPARQL query to retrieve data from the constructed KG. However, this is not presented in the article as we focus primarily on the KG creation. Our algorithm functions within a structured and simple user-system interaction framework. The initial message to the system we call *Instruction*, outlines the model’s role as an expert in CH and its familiarity with the CIDOC-CRM ontology. This message specifies the task of creating a list of formal triple statements derived solely from the provided CIDOC-CRM classes and properties. The model should generate a complete KG based on the painting description supplied from the user\_input. Note that we prompt the model to always generate 20 statements to prevent an excessive number resulting from the large set of paintings. The KG is represented in a structured JSON format of the triple statements with values that may include classes rather than literals. The JSON output type is chosen because it ensures a straightforward way for KG validation by querying afterward, due to its augmented accuracy compared to the natural language output. Note that this query process is not associated with the *Processing* step detailed in the LLM-extended method.

The application starts with the prompt question to the user: “Please enter the description of the painting:”. The JSON schema is supplied to the GPT model to ensure the structured output. Once the painting description is submitted as a user\_input, the model creates a corresponding KG, adding it to the array of all KGs. Assume  $t_i$  is the title of the  $i$ -th painting from the set  $T = \{t_1, \dots, t_n\}$ , where  $n$  is the total number of paintings. Let  $G = \{G_1, \dots, G_n\}$  be the set of all KGs, where  $G_i$  represents the KG for the  $i$ -th painting. We denote the output of the corresponding painting as a tuple  $PT_i = (t_i, G_i)$  where each KG,  $G_i$  is a set of triples defined as:

$$G_i = \{(S_{ij}, V_{S_{ij}}), PR_{ij}, (O_{ij}, V_{O_{ij}})\}, \quad j = \{1, \dots, m\}, \quad (1)$$

where in general,  $S_j$  is a set of CIDOC-CRM subject classes,  $V_{S_j}$  is a set of subject values (instances),  $PR_j$  is a set of CIDOC-CRM properties representing the predicates,  $O_j$  is a set of CIDOC-CRM object classes, and  $V_{O_j}$  is a set of

object values (instances). Here,  $m$  denotes the total number of triples in the  $i$ -th KG. The construction of each  $G_i$  generally can be represented as a result of the mapping of user input to a set of triples:  $f_{\text{construct}} : \text{user\_input} \rightarrow G$ . The resulting output—CIDOC-CRM triple statements in text format—and the validation are presented in Section 5.1. Our algorithm operates locally, using the Python programming language and the GPT-4 model. We do not extend the capability to continuous conversation since the desired output is a KG based on a single description. Therefore, the output is not adjustable using additional prompting. The schematic prompting mechanism for painting “Entrance to the Jardin Turc” is shown in Prompt Workflow 1 (ellipsis indicates text omitted for brevity). This process works on one painting at a time.

### Prompt Workflow 1:

#### ☰ Instructions:

You are an expert in CH and CIDOC-CRM ontology. Your task is to construct triple formal statements using only CIDOC-CRM classes and properties, ensuring the input painting is always considered in creating the KG. Your response should be in the form: Domain class (value)—property—Range class (value). Values can also be the classes but not the literals. Focus on capturing the most important information about the painting in no more than 20 statements. Use CIDOC-CRM 7.1.2 version. Provide a complete and concise KG based on the given painting description.

Painting description: “**user\_input**.”

#### ⚙️ Message to User:

Please enter the description of the painting:

#### 👤 user\_input:

... The Entrance to the Turkish Garden Cafe (Alternate Title), Department: Paintings, Classification: Painting, Object Type: Painting, Crisply painted and teeming with detail, Boilly’s picture transports us to Napoleonic Paris, where we stand on a shady boulevard outside the Jardin Turc (Turkish Garden Café),... [truncated]

#### 📄 Expected Output (JSON Schema):

```
{ "painting_name": "Entrance to the Jardin Turc",
  "triples": [ {
    "domain_class": "E12_Production",
    "domain_instance": "Production_of_Entrance_to_the_Jardin_Turc",
    "property": "P4_has_time-span",
    "range_class": "E52_Time-Span",
    "range_instance": "1812" }, ...
  ]
}
```

### 4.3 LLM-Pure Method

The goal of employing only LLMs eliminates some steps crucial to the methods that involve ontology, such as forming KG and SPARQL queries. It is particularly interesting because the *Information output* and *Processing* steps within the general workflow in Figure 2 happen within the functionality of GPT. We conducted an experiment

based on the study by Vasic et al. [51], which incorporates LLM to provide users with the most suitable painting that matches their expectations in an interactive virtual panoramic tour. This broader concept was refined in this study as ranking paintings according to annotation statements describing each artwork.

The same sample of Getty Collection paintings used in the LLM-extended method was employed here, with a subset of twenty artworks selected for evaluation. The *Instruction* message to GPT provides guidelines for the ranking process. While our tool automatically iterates through all the painting descriptions, the OpenAI model receives and processes only one painting at a time. This approach is necessary due to the substantial length of the text, which may exceed processing limits. The user is prompted to enter the statement for evaluation, which we call **User Request (UR)** or descriptive statements. It is important to note that the set of URs is general and that the same set is used for all paintings. This allows for testing the model's responses even when the UR is not explicitly contained in the painting description, which closely reflects real-life scenarios involving URs in a museum setting.

The paintings should always be ranked on a scale from 1 to 5 Likert scale, where 1 indicates the least alignment with the statement and 5 indicates the most alignment with the statement. The lowest ranking should also be assigned to the painting whose description or metadata lacks the data related to the UR. JSON schema is used for the output to ensure data accuracy and consistency, making it easily accessible for the subsequent validation process. The *temperature* value of the model has been set to 1 to ensure a balance between randomness and determinism in the generated outputs. The schematic prompting mechanism is shown in Prompt Workflow 2.

### Prompt Workflow 2:

#### ☰ Instructions:

You are a painting evaluator. You will be presented with content that consists only of the textual descriptions of one painting and 10 form statements. Your task is to rank how well the painting's textual description aligns with the statements (e.g., "The description includes references to nature, such as landscapes.").

Your task: Rank each painting from 1 to 5 based on how well it meets the form statements:

- (1) Does not meet the criteria or relevant information is missing from the description;
- (2) Meets very few aspects;
- (3) Meets some aspects;
- (4) Meets most aspects;
- (5) Meets entirely the criteria.

If the information needed to answer a question is missing from the painting's metadata or description (e.g., description, place of creation, or other relevant details), assign a rank of 1.

#### 📄 Painting Description:

...Chinese porcelain and tea-drinking were the rage of fashionable Europe when Jean-Étienne Liotard was born. He began painting tea and coffee sets in the last two decades of his life when age, changes in taste, and his political beliefs caused a decline in requests for the pastel portraits that were his specialty. He had,

however, been including fruit and porcelain still lifes in some of his portraits since about 1740. Only five of his paintings of tea and coffee sets are known today. [truncated]

#### ⚙️ Message to User:

Enter your evaluation question:

#### 👤 user\_input:

The painting depicts a single moment rather than a sequence of events.

#### 📄 Expected Output (JSON Schema):

```
[ {
  "painting_number": 1,
  "evaluation": 5,
  "explanation": "The painting captures the specific moment of the miracle when Jesus
  raises his hand in benediction, thus emphasizing a single moment
  rather than a sequence of events."
},...
]
```

## 5 Experiments and Validation

In the following section, we demonstrate experiments that we conducted considering methods that involve LLM and their related validations. Validating the LLM-extended method is not a trivial task because artwork can be formally described in different ways depending on the targeted subject and the approaches of experts. Similarly, evaluating painting rankings based on descriptive statements is challenging, as even humans may express differing opinions and interpretations. To evaluate the LLM-extended method—given that our study is theoretical and does not involve experts in the validation process—we choose to assess the semantic accuracy of the generated KG. To evaluate the rankings in the LLM-pure method, we conducted a user study on the same task to establish a baseline, which was then compared with the LLM’s responses. Additionally, for the last method, we measure the model’s consistency using a repetitive prompting technique. Recall that for the last method, we use a smaller painting sample to reduce the workload in the user study.

### 5.1 LLM-Extended Method

In this section, we demonstrate an example of LLM-generated KG based on CIDOC-CRM in the form of triple statements. The exact *Instruction* and *user\_input* are included in the code within the supplementary material in [52]. We tested the model on a sample of 50 paintings, which resulted in 50 CIDOC-CRM KGs each containing 20 formal statements, producing a total of 1,000 formal statements. The criteria for the painting selection are described in Appendix A. Output 1 lists the generated triple statements formally describing the painting “Entrance to the Jardin Turc” by Louis-Léopold Boilly, while the KGs for the rest of the paintings can be found in [52].

#### Output 1:

E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P102\_has\_title - E35\_Title (L’Entrée au Jardin Turc)

E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P102\_has\_title - E35\_Title (The Entrance to the Turkish Garden Cafe)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P108i\_was\_produced\_by  
 E12\_Production (Production\_of\_Entrance\_to\_the\_Jardin\_Turc)  
 E12\_Production (Production\_of\_Entrance\_to\_the\_Jardin\_Turc) - P14\_carried\_out\_by - E21\_Person (Louis-Léopold\_Boilly)  
 E12\_Production (Production\_of\_Entrance\_to\_the\_Jardin\_Turc) - P4\_has\_time-span - E52\_Time-Span (1812)  
 E12\_Production (Production\_of\_Entrance\_to\_the\_Jardin\_Turc) - P7\_took\_place\_at - E53\_Place (France)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P45\_consists\_of - E57\_Material (Oil\_on\_canvas)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P43\_has\_dimension - E54\_Dimension (Unframed\_Dimensions)  
 E54\_Dimension (Unframed\_Dimensions) - P90\_has\_value - E60\_Number (73.3\_x\_91.4\_cm)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P43\_has\_dimension - E54\_Dimension (Framed\_Dimensions)  
 E54\_Dimension (Framed\_Dimensions) - P90\_has\_value - E60\_Number (90.5\_x\_109.2\_x\_10.8\_cm)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P55\_has\_current\_location  
 E53\_Place (The\_J\_Paul\_Getty\_Museum)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P1\_is\_identified\_by - E42\_Identifier (Object\_Number\_2010.11)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P65\_shows\_visual\_item - E24\_Physical\_Man-Made\_Thing (Self-portrait\_of\_Boilly)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P128\_carries - E34\_Inscription (Signature\_Lower\_Right)  
 E34\_Inscription (Signature\_Lower\_Right) - P3\_has\_note - E62\_String (L\_Boilly\_1812)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P128\_carries - E34\_Inscription (Verso\_Labels)  
 E34\_Inscription (Verso\_Labels)  
 P3\_has\_note - E62\_String (Various\_labels\_including\_Christie's\_and\_Collection\_Georges\_Lutz)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P2\_has\_type - E55\_Type (Painting)  
 E22\_Man-Made\_Object (Entrance\_to\_the\_Jardin\_Turc) - P2\_has\_type - E55\_Type (French\_Culture)

**5.1.1 Validation of the LLM-Generated KGs.** The validation of our LLM-generated KG is performed employing **Competency Questions (CQs)** [26], somewhat replicating the approach from [45]. The authors employ CQs for the manual SPARQL query creation to test whether the LLM-generated ontology from ontology stories includes related responses. Additionally, we extend the task by using the same CQs to check whether the baseline KG of the Getty Collection provides the same answers as our KG. In [45], the total score of validation is calculated as an average value of ontology evaluations across all CQs:

$$score = \frac{\sum_{w=1}^q f(O, CQ_w)}{q}, \quad (2)$$

where  $q$  is the total number of CQs,  $O$  is the LLM-generated KG,  $CQ$  can be represented as a vector such as  $CQ = \{CQ_1, \dots, CQ_q\}$ , and  $f(O, CQ_w)$  is the function that evaluates  $O$  with respect to  $CQ_w$  and returns either 0 if the ontology does not contain the answer to the CQ, and 1 if it does.

As we aim to compare the results of our LLM-generated output and Getty Collection KG, we have to modify the previous score formula. First, we need to obtain the responses to the following CQs from our KGs:

- (1) (Title) What is the title of the painting?

Table 2. Semantic Queries for Painting Descriptions

Query	Domain Class	Property	Range Class
CQ1–Title	E22_Man-Made_Object	P102_has_title	E35_Title
CQ2–Date of Creation	E12_Production, E22_Human-Made_Object	P4_has_time-span	E52_Time-Span
CQ3–Material	E22_Man-Made_Object, E22_Human-Made_Object	P45_consists_of	E57_Material
CQ4–Type	E22_Man-Made_Object	P2_has_type	E55_Type
CQ5–Place	E12_Production	P7_took_place_at	E53_Place
CQ6–Artist	E12_Production	P14_carried_out_by	E21_Person, E39_Actor
CQ7–Object Identifier	E22_Man-Made_Object	P48_has_preferred_identifier, P1_is_identified_by	E42_Identifier

- (2) (Date of Creation) What is the creation time of the painting?
- (3) (Material) Which material is used for the painting?
- (4) (Type) How is the painting categorized by type?
- (5) (Place) What is the place where the painting was created?
- (6) (Artist) Who is the painter of the artwork?
- (7) (Object Identifier) What is the object identifier of the painting?

In Equation (1) (Section 4.2), we have introduced classes, properties, and the corresponding values of  $G_i$ . Each CQ will constitute the query into each  $G_i$  so we introduce queries as a set  $Q = \{Q_1, \dots, Q_q\}$  where  $q$  is the total number of queries (relates to the number of CQs). In our case specifically, seven queries have been proposed (see Table 2).

As shown in the query, it is sometimes necessary to query multiple classes and properties, depending on the specific usage in each KG (see the case of querying creation date, for example). The query function takes a painting title  $t_i$  from a painting  $PT_i$ ,  $Q_w$  from  $Q$ , and  $G_i$  from  $G$  and creates a set  $R_i$  of results. Specifically, for each element from  $Q$ , and  $G_i$ , the function  $f_{\text{query}}$  is defined as follows:

$$f_{\text{query}}(t_i, Q, G_i) = R_i, \quad (3)$$

where  $R_i$  consists of the painting *title*, queried element such as E35\_Title, and query value. Note that the *title* used here refers to the label or identifier of the painting within our JSON structure and not to the ontological class E35\_Title. It is used to locate and link the correct painting instance in the KG during query execution. This output is important in a subsequent step for comparing values  $v_{\text{title}}, v_{\text{date}}, \dots, v_{\text{id}}$  with those of the Getty Collection KG for a corresponding painting. Getty Collection KG is queried with SPARQL and LLM-generated KG with a custom Python script given the output in JSON format. To maintain the flow of the functions and annotations, we denote the entire Getty Collection database as  $G'$  and the results from query  $R'_i$ . Note that we use the same index  $i$  here as that of our KG,  $G_i$ , because we will obtain the same number of results. The function that queries Getty Collection becomes as follows:

$$g_{\text{query}}(G_{\text{query}_i}, G') = R'_i, \quad (4)$$

where  $G_{\text{query}_i}$  is the  $i$ -th SPARQL query of the entire Getty Collection KG. Query results consist of a similar structure as  $R_i$  such as  $R'_i = \{t_i, \text{"title,"} v'_{\text{title}_i}, \text{"date,"} v'_{\text{date}_i}, \dots, v'_{\text{id}_i}\}$ . Finally, we can compare values of results from

Table 3. Mean Scores for Each Query Element across All Paintings

Query	Mean Score
Title	0.940
Date of Creation	0.960
Material	0.980
Type	0.980
Place	0.920
Object Identifier	0.860
Artist	0.960
Overall Score	0.943

Equations (3) and (4) and find the score of the modified *score* function from Equation (2):

$$\text{score}' = \frac{1}{n} \sum_{i=1}^n \left( \frac{1}{q} \sum_{k=1}^q I(V_{ik} = V'_{ik}) \right), \quad (5)$$

where  $V_{ik}$  is a set of values  $v_{\text{title}_{ik}}, v_{\text{date}_{ik}}, \dots, v_{\text{id}_{ik}}$  from  $R_i$  corresponding to the  $k$ th query ( $q_k$ ),  $V'_{ik}$  is a set of values  $v'_{\text{title}_{ik}}, v'_{\text{date}_{ik}}, \dots, v'_{\text{id}_{ik}}$  from  $R'_i$  corresponding to the  $k$ th query, and  $I$  is an indicator that returns 1 if the values are semantically equal and 0 otherwise. This function returns the average similarity score between our annotations and the Getty Collection KG, resulting in a high similarity of 94.3%. In Table 3, we present the validation scores, which include the mean value for each query across all paintings. The overall score (*score'*) naturally corresponds to the average of the scores across all paintings.

The validation remains challenging because the two KGs being compared sometimes contain values that share the same semantic meaning but are represented in different formats. Second, when querying the Getty Collection, the place where the artwork is created is not always associated with the same entity, or it is nonexistent. In some cases, the queried values can be easily inferred from the painting description, but our goal was to obtain all the information using the conventional SPARQL query language method [43] for robustness.

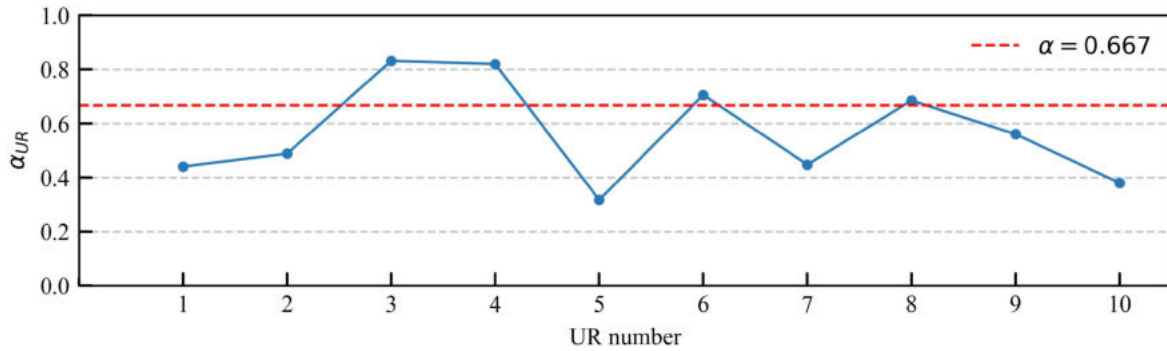
## 5.2 LLM-Pure Method

This experiment demonstrates the LLM's capability to analyze and rank paintings according to the URs. Recall that we use 10 uniform URs applied uniformly across all paintings (see Table B1 in Appendix B). Table 4 shows the rankings of 20 paintings based on all 10 URs.

**5.2.1 User Study-Based Baseline Creation.** To establish a reference point for evaluating the performance of our model, we conducted a user study, aiming to capture human judgments on the same ranking tasks. We thus created a web-based custom form presenting the 10 evaluation questions (URs) previously defined, applied to 20 selected paintings. It is available online at: <https://ivavasic.github.io/painting-survey/>. Given that the URs are general for all paintings, we assume that responses will be inconsistent in some cases. We acknowledge that aligning URs specifically with each painting could improve performance, but would distort the real-life scenario, artificially introducing tuning, which we intentionally avoid. Alongside participants' responses, we also collect data on age, gender, education level, country, profession, experience in art, IP address (when available), and response time. To mimic the task we gave to the LLM, we asked participants to rate each painting using a 5-point Likert scale. They were recruited through internal group channels, combining individuals from our network and

Table 4. Ranking Values of 20 Selected Paintings Based on Each UR

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Q1	5	5	3	5	3	5	1	4	1	4	1	5	2	5	5	5	5	5	5	5
Q2	1	1	5	1	3	1	1	1	1	1	5	1	1	1	2	1	1	5	1	1
Q3	5	1	1	1	1	5	1	1	3	1	5	1	1	1	1	1	1	1	1	1
Q4	5	5	5	4	5	5	1	1	1	1	5	1	1	5	1	1	1	4	1	1
Q5	5	2	4	2	5	5	4	2	3	1	5	4	1	3	4	1	2	5	1	1
Q6	5	1	4	1	4	2	5	1	1	2	5	2	1	1	5	1	3	5	1	1
Q7	5	1	5	4	1	1	1	1	1	1	4	4	1	4	1	1	1	5	1	3
Q8	1	1	1	1	1	3	4	5	1	1	2	1	1	1	3	5	5	2	5	1
Q9	1	1	1	1	4	2	1	5	1	1	3	1	1	1	4	1	1	1	1	1
Q10	2	3	4	1	5	4	2	4	4	1	4	1	1	1	1	2	4	5	1	1

Fig. 6. Krippendorff's alpha ( $\alpha_{UR}$ ) for each UR.

others who joined more spontaneously. More details on the survey architecture are described in Appendix C.1. All the collected data are available in the supplementary material in [52].

In total, 20 users participated in the study. While not fully randomized, the sample presents a group of varying demographic information. The summary of data is provided in Appendix C.2. We treated responses as interval data and computed Krippendorff's alpha to assess inter-rater agreement across all users, implemented using the Python library [6]. To better understand the user agreement patterns, we computed Krippendorff's alpha per painting (across all users and URs) and per UR (across all users and paintings). It varies both across paintings and URs, with several values falling below the commonly accepted threshold of 0.667 for tentative conclusions [31], indicating that inter-rater agreement is not consistently high among humans. This is not surprising since the diverse demographic data of participants may lead to some URs being ambiguous, too general, or interpreted differently by different users. Figure 6 shows the trend of Krippendorff's alpha ( $\alpha_{UR}$ ). Therefore, only URs with sufficiently reliable agreement ( $\alpha_{UR} \geq 0.667$ ) were selected for establishing the evaluation baseline. Figure 7 shows a comparison of Krippendorff's alpha per painting, computed using all URs and only the selected URs. Although some values have decreased, the filtering resulted in a more consistent ranking where alpha increased, making it easier to establish a reliable consensus for evaluation.

Finally, we use only URs with sufficient agreement to compute the mean of user ratings for each painting and UR. These averaged human responses serve as the evaluation baseline for comparison with GPT-generated ratings. For completeness, we also show the full evaluation table in Appendix C.2, including the mean, standard deviation, and variance of user responses for all URs (including those with low inter-rater agreement).

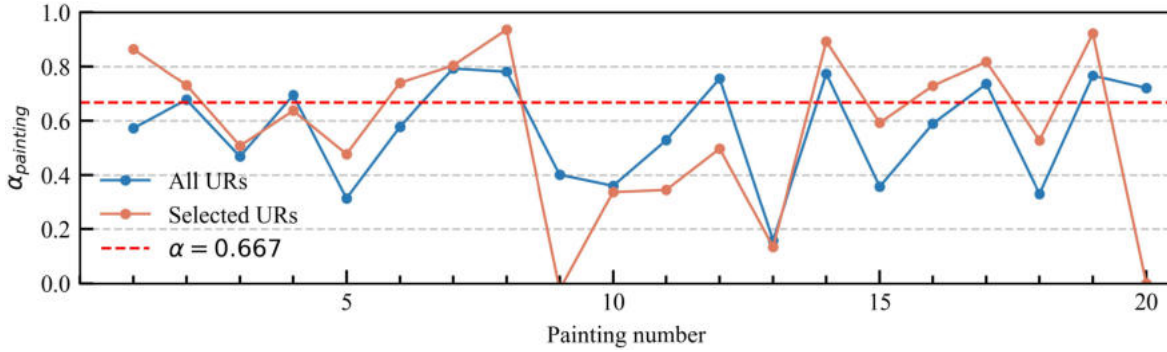


Fig. 7. Comparison of Krippendorff's alpha ( $\alpha_{painting}$ ) for each painting using all URs and the selected URs.

Table 5. Absolute Differences between Human and GPT Ratings per Painting and Question

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Q3	<b>0.20</b>	<b>0.15</b>	1.05	<b>0.10</b>	<b>0.15</b>	<b>0.20</b>	<b>0.10</b>	<b>0.10</b>	1.75	<b>0.00</b>	<b>0.30</b>	<b>0.15</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.00</b>	<b>0.05</b>	<b>0.00</b>	<b>0.00</b>
Q4	<b>0.15</b>	<b>0.50</b>	<b>0.45</b>	<b>0.25</b>	0.90	<b>0.20</b>	<b>0.10</b>	<b>0.15</b>	<b>0.40</b>	<b>0.10</b>	<b>0.15</b>	<b>0.10</b>	<b>0.10</b>	<b>0.50</b>	0.75	<b>0.20</b>	<b>0.05</b>	<b>0.10</b>	<b>0.10</b>	<b>0.10</b>
Q6	0.55	<b>0.20</b>	0.65	<b>0.25</b>	0.75	<b>0.50</b>	1.30	<b>0.05</b>	<b>0.20</b>	<b>0.50</b>	<b>0.20</b>	0.95	<b>0.15</b>	<b>0.20</b>	0.55	<b>0.00</b>	0.85	0.75	<b>0.15</b>	<b>0.00</b>
Q8	<b>0.20</b>	<b>0.40</b>	1.30	<b>0.45</b>	0.95	1.55	0.80	<b>0.15</b>	<b>0.20</b>	<b>0.05</b>	1.15	<b>0.05</b>	0.80	<b>0.10</b>	<b>0.25</b>	0.80	<b>0.25</b>	1.15	<b>0.40</b>	<b>0.00</b>
<b>Overall Mean = 0.37</b>																				

5.2.2 *Humans vs. LLM Evaluation.* To evaluate the GPT-generated ratings against the baseline, we computed the absolute differences for each painting and question. This allows us to quantify how far the GPT predictions deviate from the human-annotated means, regardless of direction. Results are shown in Table 5.

The overall mean deviation suggests that, on average, GPT predictions differ from human ratings by less than half a point on a 1–5 Likert scale, indicating generally good alignment. Values highlighted in bold indicate the deviation less than or equal to half a point on the Likert scale, which we deliberately chose as a threshold, albeit a full point would also be a reasonable consideration. Notably, some deviations are zeros, indicating a complete alignment between the model and a human judgment. Apart from the deviation, the model performed the rankings in an overall mean time of 49.24 seconds, which is approximately 89.78 times faster than humans did in the user study. The overall average for humans was 4,420.72 seconds (a bit more than 1 hour), calculated as the mean total time spent per user during the ranking task. We executed the LLM model on a machine with an AMD Ryzen 5 PRO 5675U CPU, AMD Radeon™ Graphics, and 15.2 GB RAM.

5.2.3 *Variance-Based Consistency Assessment.* The second validation is performed by calculating the variance measure of the obtained rankings. More precisely, we prompted the model five times consecutively with the same UR for each painting. Assume the set of  $c$ th ranking values for  $i$ -th painting and  $a$ th UR is the obtained set of values  $URV_{iac} = \{x_{ia1}, \dots, x_{iad}\}$ , where  $d$  is the number of times the prompt is repeated. The overall variance is then calculated as the average of all variances  $\sigma_{ia} = \frac{1}{d} \sum_{c=1}^d (x_{iac} - \mu)^2$  across all paintings and URs:

$$\sigma_{overall}^2 = \frac{1}{m} \sum_{i=1}^m \frac{1}{b} \sum_{a=1}^b \frac{1}{d} \sum_{c=1}^d (x_{iac} - \mu)^2, \quad (6)$$

where  $m$  is the number of paintings,  $b$  is the number of URs and  $\mu$  is the mean value of all  $x_{iac}$  in the corresponding set. Ideally, results are consistent, meaning the variance is close or equal to zero and the ranking values are repetitive. The results are presented in Figure 8, using the temperature of 0 for our model. Most of the values

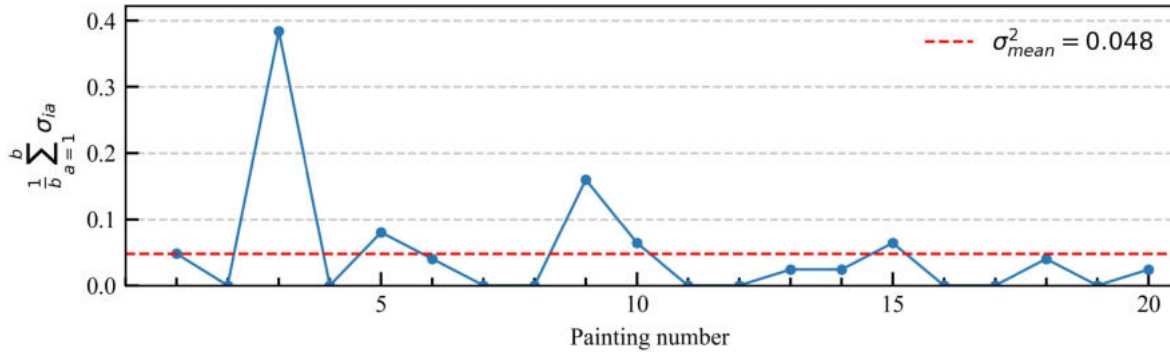


Fig. 8. The overall variance for each painting across all 10 URs.

are near zero, which indicates that all ranking values are identical. The obtained overall variance suggests small differences in the LLM model response. The maximum obtained variance is 3.840 for painting “The Dream of Pope Sergius,” based on the UR3: “*The painting depicts a biblical event*,” with values [5, 1, 1, 5, 5]. The total overall variance for all paintings across all URs is 0.048.

Validating the results obtained from ranking can sometimes be subjective. For example, if a UR involves ranking a “beautiful” painting, the result may not be satisfactory, as beauty is a subjective concept. Similarly, if the statement depicts the “dark” moment, it could be interpreted as a painting with dark colors or one depicting dark emotions. In addition, although we instruct the model to assign a rank of 1 to paintings without information related to the UR, it might not always adhere to this rule consistently, which can result in variability in rankings and an increase in variance.

## 6 Discussion and Comparison Results

Textual input or narratives have long been widely used by CH institutions and experts—a practice that continues today due to its natural expressiveness, like in the case of the Civic Art Gallery of Ascoli described in Section 1.1. In contrast, working with ontologies imposes certain constraints, with predefined classes and properties linked together in a rigid structure, requiring users to adhere to a specific ontological order. It is important to note that working with KGs and ontologies requires not only CH expertise but also explicit knowledge of the complex schema behind them as demonstrated in our developed tool. CH professionals may not always be familiar with such representations. However, the Traditional method plays an important role in data management, as it utilizes algorithms that accurately and consistently capture information from experts, in contrast to the variability often observed in LLM-based methods.

Conducted experiments showed the great potential of LLMs for interpreting the semantic content of the paintings. Together with additional natural language knowledge, e.g., from museum catalogs, LLMs can also derive suggestions for similar images. Such functionality could be directly integrated into virtual museum platforms. LLM-based processing of semantic information is easier to perform than traditional ontology-based semantic annotation because it does not require understanding complex systems or following specific guidelines as required when working with ontologies.

On the downside, LLMs like GPT can produce inaccurate responses because they are trained on broad datasets and lack access to real-time information or specific sources. Moreover, it is hard to determine whether GPT incorporates more advanced extensions of ontologies. This is particularly problematic in specialized domains such as art description, where detailed, accurate knowledge is essential.

The comparison of the three methods is summarized in Table 6, based on the selected parameters identified throughout the research process. Although some of our experiments are demonstrated in practice, the comparison

Table 6. Comparison of Three Methods

Parameter	Description	Traditional Method	LLM-Extended Method	LLM-Pure Method
User Experience	The annotation tool should be designed to be simple and user-friendly, allowing experts to annotate paintings in a few steps.	Low	Medium	High
Efficiency	Refers to the time needed for task execution.	Low	High	High
Accuracy	Given that the ontology implies that the user sticks to the rules of the triple statements and respects class inheritance, any false or ambiguous input becomes detectable.	High	Medium	Medium
Data Richness	The extent to which the methods handle different amounts of data in input.	High	Medium	Low
Input Precision	The ability of the system to detect errors in input, both in natural language format or formal statements. It also represents the syntactic accuracy of the user input.	Low	High	High

remains mostly on the conceptual level. A detailed discussion related to each evaluation is provided in subsequent paragraphs.

*User Experience.* In the context of traditional-based tasks with KG and ontology, *user experience* is often less satisfactory compared to using LLMs. However, user experience can vary amongst different individuals, influenced by factors such as their technical expertise, experience, and personal preferences, amongst others. This is due to the need for specialized knowledge in ontology selection. In contrast, LLMs, when integrated into platforms such as virtual museums, offer ease of use, partly because they can be effectively incorporated into chatbot systems and support natural language input. LLM-based annotation is easier to perform because it does not require an understanding of some complex systems or using specific guidelines.

*Efficiency.* Our experiments show that LLM can easily interpret CIDOC-CRM ontology statements and convert them into a narrative, while the accuracy may be questionable. Thus the *efficacy* is higher in the approaches with LLMs, compared to the traditional due to the time for manually constructing each formal expression in a KG, especially for large-scale KGs. Working with ontologies, specifically with the CIDOC-CRM, requires not only expertise in CH but also explicit knowledge of this complex system. Although the CIDOC-CRM ontology has been accepted and extended on several occasions by the scientific community, from our knowledge, CH experts may not always be familiar with it.

*Accuracy.* Traditional methods perform better and are more robust than those relying solely on LLM usage when it comes to *accuracy*. LLMs like ChatGPT are also prone to hallucinations which may lead to inaccurate responses as they are trained on broad datasets and lack access to real-time information or specific sources. This is especially problematic in specialized fields like art description, where detailed, accurate knowledge is essential. Our observation from the experiments is also that our model did not adhere to the specified CIDOC-CRM schema version in the *Instruction*. For example, it used an older version of class naming such as Man-Made\_Object instead of Human-Made\_Object for the class E22. Therefore, a combination of ontology and LLM presents the optimal

approach, as it provides the accuracy of structured knowledge while also offering a tool that interprets these elements into natural language and vice versa.

*Data Richness.* In museology, content can be exhaustively described by different experts and stored across various sources. Cataloging sites and platforms like Europeana and Google Arts & Culture<sup>9</sup>, along with museum institutions, display only a curated selection of content. This approach facilitates comprehensive and balanced exploration, allowing visitors to have more enriching and balanced experiences during their visits. We are interested in how the three selected methods manage different amounts of data and how this affects the ultimate outcome, i.e., in choosing artworks aligned with visitor preferences. Based on the results from experiments, it is clear that the input within the methods involving LLMs supports a large quantity of data but must be segmented at certain points to remain within the maximum token limits. This is reflected in our experiment design which processes paintings individually. However, for the ranking approach, it would be beneficial to enable direct comparisons between paintings to enhance the accuracy and relevance of the results. This process, however, introduces challenges for the model, as handling all descriptions simultaneously in natural language requires a different approach compared to processing individual inputs. Ontology annotation does not restrict the quantity of information but requires more effort in its formulation.

*Input Precision.* This is another vital variable in our research and refers to syntactic accuracy. Each of the three approaches—whether based solely on ontology selection, a combination of ontology and LLMs, or LLMs alone—requires NLP for every kind of input, whether for lengthy descriptions or single words that correspond to values in the CIDOC-CRM ontology classes and properties. The processed input must always be accurate to ensure error-free content. This is especially important for the high level of recognition and effective querying of the entire input by the system. Thus, we identify a combination of ontology input with the support of LLM as the most suitable option because LLMs can identify textual errors and convert them into the correct form.

## 7 Conclusion and Future Works

The main contribution of this study is the evaluation of traditional, LLM-extended, and LLM-pure methods based on various variables and empirical evaluation through several experiments on CH content. Specifically, we investigated the semantic information strategies of the artworks to better understand the usefulness of KGs based on ontology and LLMs usage in the CH field. The traditional approach is still the most robust and accurate, while the usage of LLMs offers various benefits when it is used both independently and when it is combined with ontologies. The LLM performs considerably faster than humans on our tasks, which has been demonstrated in previous sections of this study in the ranking process. Our experiments with LLMs have shown the potential to support artwork annotation, which can be further used in use cases such as guiding museum visitors to artifacts that match their interests, assisting museum curators with documentation, or helping CH researchers identify similarities in artifact collections.

The artwork annotation process is a complex task within our methods, requiring precise descriptions from verified sources and constrained web querying to ensure accuracy. A potential direction when using LLMs as support is the development of a system dedicated exclusively to utilizing descriptions from experts or other verified sources to minimize the risk of misinformation and inaccuracies. The solution may be fine-tuning the existing LLMs and/or developing small-size models trained on the specific domain knowledge. Additionally, the integration of ontologies could enhance the robustness of the output information by ensuring interconnected and structured data representation. The experiments on painting ranking show the capability of understanding the meaning of descriptions in a natural language, and ranking each of them according to the URs. Moreover, the results of the conducted user study show a close alignment between the LLM-generated rankings and human judgment.

<sup>9</sup><https://artsandculture.google.com/>, last accessed 23 January 2025.

Future work encompasses putting the experiments with LLMs into practice and developing a museum virtual tour prototype as a compact tool for semantic annotation, integrated with a virtual guide. One of the promising directions is leveraging novel technologies like LangChain for building an end-to-end application. As a framework that utilizes LLMs for accessing external data sources, and operating with queries, amongst others, it would be a strong choice in collaboration with museum institutions, while also improving accuracy in the implementation of LLM-based methods. For potential large-scale projects concerning an LLM-extended method, this would help the tedious process of manual SPARQL query creation. Similarly, incorporating expert-in-the-loop for LLM-generated annotations could enhance the quality of KGs. Additionally, employing different LLM models and evaluating their performance would be beneficial. Further experiments may incorporate LLMs with vision capabilities, which will be particularly beneficial for the spatial information of the paintings in addition to the textual descriptions already explored in the study.

## Acknowledgments

The work on this article was partially carried out at the Smart Living Lab (<https://www.smartlivinglab.ch/en/>), a joint project funded by the University of Fribourg, EPFL, and HEIA-FR.

A grateful thank to Ascoli Musei and Ascoli municipality, in particular to the Director of the collections, Stefano Papetti, and Matteo Felicetti.

The authors would also like to acknowledge Paolo Clini, Emanuele Frontoni, and Eva S. Malinverni for their support of the present research, as well as Simon Curty for his technical assistance with the web application at the beginning of the research. We are also grateful to all users who participated in our online survey, helping establish a baseline for evaluating the LLM-pure method.

## References

- [1] Cristiana Araújo, Ricardo G. Martini, Pedro Rangel Henriques, and José João Almeida. 2018. Annotated documents and expanded CIDOC-CRM ontology in the automatic construction of a virtual museum. In *Developments and Advances in Intelligent Systems and Applications*. Álvaro Rocha and Luís Paulo Reis (Eds.), Springer International Publishing, Cham, 91–110. DOI: [https://doi.org/10.1007/978-3-319-58965-7\\_7](https://doi.org/10.1007/978-3-319-58965-7_7)
- [2] Fabien Arnaud, Cécile Pignol, Pierre Stéphane, Anne-Lise Develle, P. Sabatier, Olivier Evrard, Brice Mourier, Maxime Debret, Cécile Grobois, Laurent Millet, et al. 2017. From core referencing to data re-use: Two French national initiatives to reinforce paleodata stewardship (National Cyber Core Repository and LTER France Retro-Observatory). In *Proceedings of the 5th PAGES Open Science Meeting*. HAL. Retrieved from <https://hal.archives-ouvertes.fr/hal-01616109>
- [3] Hamed Babaei Giglou, Jennifer D’Souza, and Sören Auer. 2023. LLMs4OL: Large language models for ontology learning. In *The Semantic Web—ISWC 2023*. Terry R. Payne, Valentina Presutti, Guilin Qi, María Poveda-Villalón, Giorgos Stoilos, Laura Hollink, Zoi Kaoudi, Gong Cheng, and Juanzi Li (Eds.), Springer Nature, Cham, 408–427.
- [4] Winfried Bergmeyer, Paul Klimpel, Anna Benedek, Ismo Malinen, Mareen Maaß, Ariadna Matas, Liv Rasmkjær, and Sylvia Willkomm. 2022. *Digital Basic Cataloguing: 10 Principles* (1st ed.). Network of European Museum Organisations and German Museums Association.
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Proceedings of the Advances in Neural Information Processing Systems*. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33, Curran Associates, Inc., 1877–1901. Retrieved January 23, 2025 from [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bf8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bf8ac142f64a-Paper.pdf)
- [6] Santiago Castro. 2017. Fast Krippendorff: Fast Computation of Krippendorff’s Alpha Agreement Measure. Retrieved from <https://github.com/pln-fing-udelar/fast-krippendorff>
- [7] Yanmei Chai, Xiaoyan Zhu, Sen Zhou, Yiting Bian, Fan Bu, Wei Li, and Jing Zhu. 2009. Ontology-based digital photo annotation using multi-source information. In *Proceedings of the 2009 IEEE International Conference on Computational Intelligence for Measurement Systems and Applications*. IEEE, 38–41. DOI: <https://doi.org/10.1109/CIMSA.2009.5069914>
- [8] Bohan Chen and Andrea L. Bertozzi. 2023. AutoKG: Efficient automated knowledge graph generation for language models. In *Proceedings of the 2023 IEEE International Conference on Big Data (BigData)*. IEEE, 3117–3126. DOI: <https://doi.org/10.1109/BigData59044.2023.10386454>
- [9] Naem Chowdhury, Abdoulsalam Awais Osman, and Sheuly Aktar. 2023. Improving Customer Care with ChatGPT: A Case Study. Retrieved January 23, 2025 from <https://zenodo.org/records/7699658>

- [10] Paolo Clini, Iva Vasic, Ramona Quattrini, Roberto Pierdicca, Eva S Malinverni, Stefano Papetti, and Matteo Felicetti. 2024. Trasformazione digitale delle Collezioni civiche di Ascoli Piceno: Un progetto innovativo tra università e musei. *Archeomatica* 15, 2 (2024), 38–47.
- [11] L.D. Couprie. 1983. Iconclass: An iconographic classification system. *Art Libraries Journal* 8, 2 (1983), 32–49. DOI: <https://doi.org/10.1017/S0307472200003436>
- [12] CIDOC CRM. 2021. CIDOC CRM RDF Data. Retrieved December 11, 2023 from [https://cidoc-crm.org/rdfs/7.1.2/CIDOC\\_CRM\\_v7.1.2.rdf](https://cidoc-crm.org/rdfs/7.1.2/CIDOC_CRM_v7.1.2.rdf)
- [13] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, 958–979. DOI: <https://doi.org/10.1109/WACVW60836.2024.00106>
- [14] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2nd Asian Conference on Asian Semantic Web Conference*, Busan, Korea. Springer-Verlag, Berlin, 722–735. DOI: 10.5555/1785162.1785216
- [15] Emmanuelle Delmas-Glass and Robert Sanderson. 2020. Fostering a community of PHAROS scholars through the adoption of open standards. *Art Libraries Journal* 45, 1 (2020), 19–23. DOI: <https://doi.org/10.1017/alj.2019.32>
- [16] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. 2022. TURL: Table understanding through representation learning. *SIGMOD Record* 51, 1 (June 2022), 33–40. DOI: <https://doi.org/10.1145/3542700.3542709>
- [17] Martin Doerr. 2009. Ontologies for cultural heritage. In *Handbook on Ontologies*. Steffen Staab and Rudi Studer (Eds.), Springer, Berlin, 463–486. DOI: [https://doi.org/10.1007/978-3-540-92673-3\\_21](https://doi.org/10.1007/978-3-540-92673-3_21)
- [18] Martin Doerr, Richard Light, and Gerald Hiebel. 2020. Implementing the CIDOC Conceptual Reference Model in RDF. Retrieved July 18, 2025 from <https://cidoc-crm.org/sites/default/files/issue%20443%20-%20Implementing%20CIDOC%20CRM%20in%20RDF%20v1.1.pdf>
- [19] Jinhua Dou, Jingyan Qin, Zanzia Jin, and Zhuang Li. 2018. Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. *Journal of Visual Languages & Computing* 48 (2018), 19–28.
- [20] Achille Felicetti, Daniel Williams, Ilenia Galluccio, Douglas Tudhope, and Franco Niccolucci. 2018. NLP tools for knowledge extraction from Italian archaeological free text. In *Proceedings of the 2018 3rd Digital Heritage International Congress (DigitalHERITAGE) Held Jointly with 2018 24th International Conference on Virtual Systems and Multimedia (VSMM '18)*. IEEE, 1–8. DOI: <https://doi.org/10.1109/DigitalHeritage.2018.8810001>
- [21] Hans-Georg Fill, Felix Härer, Iva Vasic, Daniel Borcard, Benedikt Reitemeyer, Fabian Muff, Simon Curty, and Marcel Bühlmann. 2024. CMAG: A framework for conceptual model augmented generative artificial intelligence. In *Companion Proceedings of the 43rd International Conference on Conceptual Modeling: ER Forum, Special Topics, Posters and Demos Co-located with ER 2024*. Enrico Gallinucci, Hasan Yasar, Sotirios Liaskos, Patrick Marcel, Peter P. Chen, Sergio de Cesare, and Frederik Gailly (Eds.), CEUR Workshop Proceedings, Vol. 3849, CEUR-WS.org, 56–69. Retrieved from <https://ceur-ws.org/Vol-3849/forum5.pdf>
- [22] Hans-Georg Fill, Peter Fettke, and Julius Köpke. 2023. Conceptual modeling and large language models: Impressions from first experiments with ChatGPT. *Enterprise Modelling and Information Systems Architectures* 18, 3 (2023), 1–15. DOI: <https://doi.org/10.18417/emisa.18.3>
- [23] Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovì. 2017. Semantic web machine reading with FRED. *Semantic Web* 8, 6 (2017), 873–893.
- [24] Getty Research Institute. 2023. Getty Vocabularies: AAT, ULAN, TGN. Retrieved January 9, 2024 from <https://www.getty.edu/research/tools/vocabularies/index.html>
- [25] Thomas R Gruber. 1995. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies* 43, 5–6 (1995), 907–928.
- [26] Michael Grüninger and Mark Fox. 1995. Methodology for the design and evaluation of ontologies. In *The 1995 International Joint Conference on AI*, Montreal, Quebec, Canada. Retrieved from <https://www.ijcai.org/past/ijcai-95/>
- [27] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d’Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, et al. 2021. *Knowledge Graphs*. Springer, Cham. DOI: <https://doi.org/10.2200/S01125ED1V01Y202109DSK022>
- [28] International Council on Archives (ICA). 2000. *International Standard Archival Description (General)* (2nd ed.). Retrieved January 9, 2024 from <https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition>
- [29] Dawn Kanter. 2024. Harnessing AI to Classify Early Modern Bible Illustrations. Features Research by Drew Thomas as Part of the Ornamento Project. Retrieved January 22, 2025 from [https://research.reading.ac.uk/digitalhumanities/harnessing-ai-to-classify-early-modern-bible-illustrations/#\\_ftn2](https://research.reading.ac.uk/digitalhumanities/harnessing-ai-to-classify-early-modern-bible-illustrations/#_ftn2)
- [30] Pyrros Koletsis and Euripides G. M. Petrakis. 2010. SIA: Semantic image annotation using ontologies and image content analysis. In *Image Analysis and Recognition*. Aurélio Campilho, and Mohamed Kamel (Eds.), Springer, Berlin, 374–383.
- [31] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research* 30, 3 (2004), 411–433.
- [32] Sylvain Lagrue, Nathalie Chetcuti-Sperandio, Fabien Delorme, Ma Thi Chau, Duyen Ngo Thi, Karim Tabia, and Salem Benferhat. 2019. An ontology web application-based annotation tool for intangible culture heritage dance videos. In *Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents*. ACM, New York, NY, 75–81.

- [33] Ashley Lewis and Michael White. 2023. Mitigating harms of LLMs via knowledge distillation for a virtual museum tour guide. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!* Devamanyu Hazarika, Xiangru Robert Tang, and Di Jin (Eds.), Association for Computational Linguistics, Kerrville, TX, 31–45. Retrieved from <https://aclanthology.org/2023.tilm-1.4/>
- [34] Duong Thi Thuy Mai, Can Van Da, and Nguyen Van Hanh. 2024. The use of ChatGPT in teaching and learning: A systematic review through SWOT analysis approach. *Frontiers in Education* 9 (2024). DOI : <https://doi.org/10.3389/educ.2024.1328769>
- [35] Maria-Cristina Marinescu, Artem Reshetnikov, and Joaquim Moré López. 2020. Improving object detection in paintings based on time contexts. In *Proceedings of the 2020 International Conference on Data Mining Workshops (ICDMW)*, 926–932. DOI : <https://doi.org/10.1109/ICDMW51313.2020.00133>
- [36] Daniele Metilli, Valentina Bartalesi, and Carlo Meghini. 2019. Steps towards a system to extract formal narratives from text. In *Proceedings of Text2Story—2nd Workshop on Narrative Extraction From Texts, co-located with the 41st European Conference on Information Retrieval (Text2Story@ECIR '19)*. Alípio Mário Jorge, Ricardo Campos, Adam Jatowt, and Sumit Bhatia (Eds.), CEUR Workshop Proceedings, Vol. 2342, CEUR-WS.org, 53–61. Retrieved from <https://ceur-ws.org/Vol-2342/paper7.pdf>
- [37] Roberto Navigli and Paola Velardi. 2006. Ontology enrichment through automatic semantic annotation of on-line glossaries. In *Proceedings of the 15th International Conference on Managing Knowledge in a World of Networks (EKAW '06)*. Springer, 126–140.
- [38] Fabian Neuhaus. 2023. Ontologies in the era of large language models—A perspective. *Applied Ontology* 18, 4 (2023), 399–407. Retrieved from <https://content.iospress.com/articles/applied-ontology/ao230072>
- [39] OpenAI. 2023. ChatGPT. Retrieved January 23, 2023 from <https://chat.openai.com/>
- [40] OpenAI. 2023. GPT-4V(ision) System Card. OpenAI Research. Retrieved January 23, 2025 from <https://openai.com/research/gpt-4v-system-card>
- [41] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, et al. 2024. GPT-4 technical report. arXiv:2303.08774. Retrieved from <https://arxiv.org/abs/2303.08774>
- [42] Jeff Z. Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, et al. 2023. Large language models and knowledge graphs: Opportunities and challenges. arXiv:2308.06374. Retrieved from <https://arxiv.org/abs/2308.06374>
- [43] Axel Polleres. 2014. SPARQL. In *Encyclopedia of Social Network Analysis and Mining*. Reda Alhaji and Jon Rokne (Eds.), Springer, New York, NY, 1960–1966. DOI : [https://doi.org/10.1007/978-1-4614-6170-8\\_124](https://doi.org/10.1007/978-1-4614-6170-8_124)
- [44] Giuseppe Resta, Fabiana Dicuonzo, Evrim Karacan, and Domenico Pastore. 2021. The impact of virtual tours on museum exhibitions after the onset of Covid-19 restrictions: Visitor engagement and long-term perspectives. *SCIRES-IT—SCientific REsearch and Information Technology* 11, 1 (July 2021), 151–166. DOI : <https://doi.org/10.2423/122394303V11N1P151>
- [45] Mohammad Javad Saeedzade and Eva Blomqvist. 2024. Navigating Ontology Development with Large Language Models. In *Proceedings of the 21st International Conference on Semantic Web (ESWC '24)*, Part I. Albert Meroño-Peñuela, Anastasia Dimou, Raphaël Troncy, Olaf Hartig, Maribel Acosta, Mehwish Alam, Heiko Paulheim, and Pasquale Lisena (Eds.), Lecture Notes in Computer Science, Vol. 14664, Springer, 143–161. DOI : [https://doi.org/10.1007/978-3-031-60626-7\\_8](https://doi.org/10.1007/978-3-031-60626-7_8)
- [46] Thomas Schleider, Raphael Troncy, Thibault Ehrhart, Mareike Dorozynski, Franz Rottensteiner, Jorge Sebastián Lozano, and Georgia Lo Cicero. 2021. Searching silk fabrics by images leveraging on knowledge graph and domain expert rules. In *Proceedings of the 3rd Workshop on Structuring and Understanding of Multimedia HeritAge Contents (Virtual Event) (SUMAC '21)*. ACM, New York, NY, 41–49. DOI : <https://doi.org/10.1145/3475720.3484445>
- [47] Guus Schreiber, Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Borys Omelayenko, Jacco van Osenbruggen, Anna Tordai, et al. 2008. Semantic annotation and search of cultural-heritage collections: The MultimediaN E-Culture demonstrator. *Journal of Web Semantics* 6, 4 (2008), 243–249.
- [48] Guus Schreiber, Barbara Dubbeldam, Jan Wielemaker, and Bob Wielinga. 2001. Ontology-based photo annotation. *IEEE Intelligent Systems* 16, 3 (2001), 66–74. DOI : <https://doi.org/10.1109/5254.940028>
- [49] Arda Sezen, Cigdem Turhan, and Gokhan Sengul. 2021. A hybrid approach for semantic image annotation. *IEEE Access* 9 (2021). DOI : <https://doi.org/10.1109/ACCESS.2021.3114968>
- [50] Nan Tang, Ju Fan, Fangyi Li, Jianhong Tu, Xiaoyong Du, Guoliang Li, Sam Madden, and Mourad Ouzzani. 2021. RPT: Relational pre-trained transformer is almost all you need towards democratizing data preparation. *Proceedings of the VLDB Endowment* 14, 8 (Apr. 2021), 1254–1261. DOI : <https://doi.org/10.14778/3457390.3457391>
- [51] Iva Vasic, Hans-Georg Fill, Ramona Quattrini, and Roberto Pierdicca. 2024. LLM-aided museum guide: Personalized tours based on user preferences. In *Extended Reality*. Lucio Tommaso De Paolis, Pasquale Arpaia, and Marco Sacco (Eds.), Springer Nature, Cham, 249–262. DOI : [https://doi.org/10.1007/978-3-031-71710-9\\_18](https://doi.org/10.1007/978-3-031-71710-9_18)
- [52] Iva Vasic, Hans-Georg Fill, Ramona Quattrini, and Roberto Pierdicca. 2025. *kg-vs-llm-virtual-museums*. Zenodo. DOI : <https://doi.org/10.5281/zenodo.16212425>
- [53] Iva Vasic, Ramona Quattrini, Roberto Pierdicca, Adriano Mancini, and Bata Vasic. 2024. 3VR: Vice versa virtual reality algorithm to track and map user experience. *ACM Journal on Computing and Cultural Heritage* 17, 3 (2024), 1–19. DOI : <https://doi.org/10.1145/3656346>

- [54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS '17)*. Curran Associates Inc., Red Hook, NY, 6000–6010.
- [55] Andreas Vlachidis, Angeliki Antoniou, Antonis Bikakis, and Melissa Terras. 2021. Semantic metadata enrichment and data augmentation of small museum collections following the FAIR principles. In *Information and Knowledge Organisation in Digital Humanities*. Koraljka Golub and Ying-Hsang Liu (Eds.), Routledge, 106–129.
- [56] Rafal Wojciechowski, Krzysztof Walczak, Martin White, and Wojciech Cellary. 2004. Building Virtual and Augmented Reality museum exhibitions. In *Proceedings of the 9th International Conference on 3D Web Technology (Web3D '04)*. ACM, New York, NY, 135–144. DOI: <https://doi.org/10.1145/985040.985060>
- [57] Mingfang Wu, Hans Brandhorst, Maria-Cristina Marinescu, Joaquim More Lopez, Margorie Hlava, and Joseph Busch. 2023. Automated metadata annotation: What is and is not possible with machine learning. *Data Intelligence* 5, 1 (2023), 122–138. DOI: [https://doi.org/10.1162/dint\\_a\\_00162](https://doi.org/10.1162/dint_a_00162)
- [58] Chih-Hao Yu and Jane Hunter. 2013. Documenting and sharing comparative analyses of 3D digital museum artifacts through semantic web annotations. *Journal on Computing and Cultural Heritage* 6, 4 (2013), 1–20. DOI: <https://doi.org/10.1145/2532630.2532634>

## Appendices

### A Painting Selection Criteria for Experiments

In this section, we explain the criteria for selecting the paintings for the experiments. As our goal is a theoretical comparison of the three methods and identifying the benefits of LLMs in the workflow, our study involves only paintings as key studies. To avoid alteration in the research directions, we also limited queries to the Getty Collection as a recognized source of querying and its publicly available SPARQL API. The diversity of the choice has been taken into account. The dataset is based on the creation date, artist, and culture. Therefore, the first filter of the query is Object Type set to Painting, which finds 970 results in the database. However, this also includes other object types, such as Fresco and Print, making it necessary to double-check that the selected item is indeed a Painting object type. Next, we based our query on the four main historical periods from the European perspective, using the Getty Collection’s date limits from 400 BCE to 2000 CE: Antiquity (400<sup>10</sup> BCE–476 CE), Medieval Ages (476 CE–1492 CE), Modern Ages (1492 CE–1789 CE), and Contemporary Ages (1789 CE–2000<sup>11</sup> CE). Table A1 details 20 of the total of 50 paintings used in this study, including the title, object number, artist’s full name, creation date based on four historical periods, and cultures. This subset of paintings is used for the LLM-pure method, while the entire set of 50 has been used for the LLM-extended method. The full list of paintings is included in the supplementary material [52].

### B User Requests for Painting Rankings

In this section, we list all 10 URs used for painting rankings (see Table B1).

### C Survey Design, Data Collection, and Participant Analysis

#### C.1 Web-Based Survey and Data Collection Methodology

The survey was implemented as a multi-page web application developed in HTML, CSS, and JavaScript. It consists of 20 pages, each displaying a textual description of a painting and a corresponding evaluation form. For each painting, participants were asked to respond to 10 URs from Table B1 using a 5-point Likert scale (the same as for the GPT model in the LLM-pure method), resulting in a total of 200 responses per participant.

Participants first completed a consent form for data collection and an identification form capturing demographic and contextual metadata (age, gender, education, country, profession, experience level). A unique user code we distributed was also required to distinguish valid submissions. Note that some users introduced some minimal typos within the code, which we corrected later since the code was obvious. Upon completion of the survey, all

<sup>10</sup>The Getty Collection includes works dated from the year 400 BCE.

<sup>11</sup>The Getty Collection includes works dated up to the year 2000 CE.

Table A1. Partial List of the Selected Paintings

N.	Title	Object Number	Artist	Creation Date	Cultures
1	The Wedding Feast at Cana	2022.28	Lavinia Fontana	Modern Age	Italian
2	Portrait of Isabella of Portugal	78.PB.3	Workshop of Rogier van der Weyden	Modern Age	Netherlandish
3	The Dream of Pope Sergius	72.PB.20	Workshop of Rogier van der Weyden	Medieval Age	Netherlandish
4	Portrait of J. Paul Getty	2003.94	Shinjiro Nakamura	Contemporary	Japanese
5	Fragmentary Coffin Lid and Fragments of Wood	75.AJ.20	Unknown Artist/Maker	Uncertain	Egyptian
6	Christ on the Cross	2000.40	El Greco (Domenico Theotokopoulos)	Modern Ages	Greek
7	The Wounded Foot	78.PA.68	Joaquín Sorolla y Bastida	Modern Ages	Spanish
8	A Hare in the Forest	2001.12	Hans Hoffmann	Modern Ages	German
9	Four Studies of a Male Head	71.PB.39	Workshop of Peter Paul Rubens	Modern Ages	Flemish
10	Wall Fragment with Two Women	96.AG.302	Unknown Artist/Maker	Antiquity	Roman
11	The Story of Joseph	70.PB.41	Biagio d'Antonio	Medieval Ages	Italian
12	The Milliners	2005.14	Edgar Degas	Contemporary	French
13	Untitled [Street Scene]	98.PA.197	Walker Evans	Contemporary	American
14	Portrait of J. Paul Getty (2)	2003.93	Armando Drechsler	Contemporary	German or Austrian
15	Entrance to the Jardin Turc	2010.11	Louis-Léopold Boilly	Contemporary	French
16	Irisés (in the Paper)	90.PA.20	Vincent van Gogh	Contemporary	Dutch
17	Starry Night (in the paper)	84.PA.681	Edvard Munch	Contemporary	Norwegian
18	The Glorification of the Union of the Houses of Hapsburg and Lorraine	84.PA.75	Franz Anton Maulbertsch	Modern Ages	Austrian
19	Corner of the Garden, Alcazar, Sevilla	79.PA.155	Joaquín Sorolla y Bastida	Contemporary	Spanish
20	Still Life: Tea Set	84.PA.57	Jean-Étienne Liotard	Modern Age	Swiss

Table B1. URs Used in the LLM-Pure Method for Painting Ranking

UR1	The painting depicts a single moment rather than a sequence of events.
UR2	The painting depicts a sequence of events rather than a single moment.
UR3	The painting depicts a biblical event.
UR4	The painting includes recognizable historical figures or well-known literary characters.
UR5	The description of the painting suggests a dramatic scene with emphasis on a specific person, action, or object.
UR6	The painting depicts interaction between people.
UR7	The description indicates that the scene takes place indoors or in an architectural setting.
UR8	The description includes references to nature, such as landscapes.
UR9	The painting includes animals.
UR10	The description suggests imaginative or symbolic elements rather than realistic ones.

data were transmitted to a private Google Sheets backend via HTTP GET requests using Google Apps Script endpoints. We included the code for the survey in the supplementary material [52], while the survey itself is available publicly online at: <https://ivavasic.github.io/painting-survey/>.

Collected data along with ratings included: identification code, user age, user gender, user education, user country, user profession, familiarity with art or museums (user experience), random code, user IP address, start time, and end time of the survey. Note that the random code, along with the IP address, was used to differentiate users. Ipify API<sup>12</sup> was used to retrieve public IP addresses. Basic validation ensured all required fields were completed before proceeding, ensuring that no URs are recorded as null. All collected data are provided in the supplementary material. The survey is also included in the supplementary material in a digitally printed form.

## C.2 Participant Demographics and Metadata Analysis

In this section, we present the demographic data of the participants.

The user sample in terms of familiarity with art or museums includes:

- Moderate: 60% (12 users); Expert: 20% (4 users); Beginner: 15% (3 users); No declared: 5% (1 user).

The following are geographical data:

- Italy: 45% (9 users); Switzerland: 25% (5 users); Serbia: 20% (4 users); Spain: 10% (2 users).

In terms of gender, the sample further includes:

- Male: 70% (14 users); Female: 30% (6 users)

In terms of user profession, the following roles were represented:

- Researcher: 50% (10 users); Student: 25% (5 users); Professor: 15% (3 users); Teacher: 5% (1 user); Chief communications officer: 5% (1 user).

In terms of user age, the following data was collected:

- 21–29: 45% (9 users); 30–39: 30% (6 users); 40–49: 20% (4 users); 50+: 5% (1 user).

To represent fully the user judgment and its analysis across all paintings and URs, we calculate Mean ( $M$ ), Standard deviation ( $SD$ ), and Variance ( $Var$ ) of all ratings. These are represented in Table C1, where the consistency of human answers varies per painting and UR. Overall variance and standard deviation are 2.88 and 1.70, respectively. Given  $M_{max} = 4$  and  $SD_{max} = 2$ , the results highlight the importance of excluding noisy URs that are obviously an obstacle for consistent human judgments.

<sup>12</sup><https://www.ipify.org/>, last accessed: 06 January 2025.

Table C1. Mean ( $M_{UR_b}$ ), Standard Deviation ( $SD_{UR_b}$ ), and Variance ( $Var_{UR_b}$ ) of Human Ratings per  $b$ th UR and Painting (P1–P20)

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
$M_1$	3.30	4.95	2.05	4.95	2.95	4.55	4.90	4.90	3.90	3.35	1.30	4.65	2.55	4.90	3.80	4.25	4.75	2.80	4.65	4.85
$SD_1$	1.59	0.22	1.39	0.22	1.54	1.23	0.45	0.31	1.52	1.79	0.98	0.75	1.70	0.31	1.64	1.45	0.55	1.70	0.99	0.49
$Var_1$	2.54	0.05	1.94	0.05	2.37	1.52	0.20	0.09	2.31	3.19	0.96	0.56	2.89	0.09	2.69	2.09	0.30	2.91	0.98	0.24
$M_2$	3.05	1.10	3.95	1.10	2.60	1.55	1.20	1.05	1.85	1.15	4.70	1.25	1.65	1.00	2.05	1.35	1.05	2.85	1.20	1.00
$SD_2$	1.70	0.45	1.23	0.45	1.47	1.28	0.62	0.22	1.27	0.49	0.98	0.72	1.39	0.00	1.57	1.09	0.22	1.66	0.89	0.00
$Var_2$	2.89	0.20	1.52	0.20	2.15	1.63	0.38	0.05	1.61	0.24	0.96	0.51	1.92	0.00	2.47	1.19	0.05	2.77	0.80	0.00
$M_3$	4.80	1.15	2.05	1.10	1.15	4.80	1.10	1.10	1.25	1.00	4.70	1.15	1.00	1.00	1.00	1.00	1.00	1.05	1.00	1.00
$SD_3$	0.70	0.67	1.67	0.45	0.49	0.70	0.45	0.45	0.72	0.00	0.98	0.67	0.00	0.00	0.00	0.00	0.00	0.22	0.00	0.00
$Var_3$	0.48	0.45	2.79	0.20	0.24	0.48	0.20	0.20	0.51	0.00	0.96	0.45	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00
$M_4$	4.85	4.50	4.55	4.25	4.10	4.80	1.10	1.15	1.40	1.10	4.85	1.10	1.10	4.50	1.75	1.20	1.05	3.90	1.10	1.10
$SD_4$	0.37	0.89	0.60	1.37	1.07	0.70	0.45	0.49	0.68	0.31	0.37	0.31	0.45	0.83	1.33	0.89	0.22	1.41	0.45	0.45
$Var_4$	0.13	0.79	0.37	1.88	1.15	0.48	0.20	0.24	0.46	0.09	0.13	0.09	0.20	0.68	1.78	0.80	0.05	1.99	0.20	0.20
$M_5$	3.50	2.65	3.80	3.05	2.55	4.80	3.40	2.00	2.70	1.50	4.00	3.05	1.30	2.00	1.75	2.00	2.05	2.45	1.35	2.65
$SD_5$	1.32	1.50	1.15	1.57	1.28	0.62	1.27	1.38	1.49	1.05	1.08	1.32	0.73	1.17	1.21	1.41	1.32	1.47	0.93	1.76
$Var_5$	1.74	2.24	1.33	2.47	1.63	0.38	1.62	1.89	2.22	1.11	1.16	1.73	0.54	1.37	1.46	2.00	1.73	2.16	0.87	3.08
$M_6$	4.45	1.20	4.65	1.25	3.25	1.50	3.70	1.05	1.20	2.50	4.80	2.95	1.15	1.20	4.45	1.00	2.15	4.25	1.15	1.00
$SD_6$	0.76	0.70	0.67	0.72	1.52	1.10	1.38	0.22	0.52	1.67	0.41	1.39	0.49	0.52	0.89	0.00	1.23	1.02	0.49	0.00
$Var_6$	0.58	0.48	0.45	0.51	2.30	1.21	1.91	0.05	0.27	2.79	0.17	1.94	0.24	0.27	0.79	0.00	1.50	1.04	0.24	0.00
$M_7$	4.10	2.70	4.10	4.25	2.55	1.60	1.20	1.15	1.70	1.30	2.40	4.60	1.35	2.75	2.10	1.55	1.20	1.90	1.75	3.30
$SD_7$	1.45	1.34	1.07	1.37	1.28	1.27	0.89	0.67	1.03	0.73	1.43	0.60	0.67	1.62	1.37	1.05	0.70	1.52	1.37	1.56
$Var_7$	2.09	1.80	1.15	1.88	1.63	1.62	0.80	0.45	1.06	0.54	2.04	0.36	0.45	2.62	1.88	1.10	0.48	2.31	1.88	2.43
$M_8$	1.20	1.40	2.30	1.45	1.95	4.55	4.80	4.85	1.20	1.05	3.15	1.05	1.80	1.10	2.75	4.20	4.75	3.15	4.60	1.00
$SD_8$	0.52	1.10	1.42	1.10	1.36	0.76	0.52	0.49	0.70	0.22	1.53	0.22	1.28	0.31	1.41	1.40	0.72	1.53	0.60	0.00
$Var_8$	0.27	1.20	2.01	1.21	1.84	0.58	0.27	0.24	0.48	0.05	2.34	0.05	1.64	0.09	1.99	1.96	0.51	2.34	0.36	0.00
$M_9$	1.00	1.15	1.20	1.05	3.80	2.70	1.20	4.60	1.00	1.90	1.00	1.00	1.00	1.00	2.35	1.00	1.00	1.25	1.10	1.00
$SD_9$	0.00	0.67	0.70	0.22	1.44	1.89	0.70	1.23	0.00	0.00	1.55	0.00	0.00	0.00	1.90	0.00	0.00	0.64	0.45	0.00
$Var_9$	0.00	0.45	0.48	0.05	2.06	3.59	0.48	1.52	0.00	0.00	2.41	0.00	0.00	0.00	3.61	0.00	0.00	0.41	0.20	0.00
$M_{10}$	2.65	1.75	3.40	1.40	4.20	3.55	1.15	2.15	1.50	1.40	2.80	1.35	1.35	1.40	1.35	1.40	2.50	3.20	1.20	1.25
$SD_{10}$	1.23	1.02	1.50	1.10	1.15	1.50	0.37	1.57	1.05	0.99	1.58	0.81	0.93	0.82	0.81	0.82	1.54	1.58	0.41	0.55
$Var_{10}$	1.50	1.04	2.25	1.20	1.33	2.26	0.13	2.45	1.11	0.99	2.48	0.66	0.87	0.67	0.66	0.67	2.37	2.48	0.17	0.30

Received 20 May 2024; revised 31 May 2025; accepted 9 July 2025