Was Michelangelo actually a Turtle?

Semantic Interpretation of Vasari's *Le vite* through Large Language Models

Gauri Bhagwat[1][0000−0003−2633−5995], Ricardo Carvalho[2][0009−0001−6605−7603], Balázs Mosolygó[3][0000−0003−2166−4255], Kristian Noullet[4][0000−0002−4916−9443], Ruben Peeters[5][0000−0002−0905−7033], Lucrezia Pograri[6][0009−0005−0129−9627], and Harald Sack[7][0000−0001−7069−9804]

[1] DHLab, KNAW Humanities Cluster
gauri.bhagwat@dh.huc.knaw.nl
[2] LASIGE, Faculdade de Ciências da Universidade de Lisboa
rmscarvalho@fc.ul.pt
[3] University of Bergen, Norway
balazs.mosolygo@uib.no
[4] AIFB, Karlsruhe Institute for Technology, Karlsruhe, Germany
{firstname.surname}@kit.edu
[5] KU Leuven, Leuven, Belgium
ruben.peeters@kuleuven.be
[6] Digital Humanities and Digital Knowledge, University of Bologna
lucrezia.pograri@studio.unibo.it

Abstract. This study addresses the underexplored efficacy of Large Language Models (LLMs) in extracting structured information from historical, multilingual, and culturally dense texts. Specifically, we focus on Giorgio Vasari's 1568 *Le vite*, a challenging corpus due to its 16th-century Florentine Italian and interpretive English translations. We evaluate whether prompt-based methods can enable historically informed, semantically meaningful data extraction without fine-tuning or large training corpora. We also examine the role of structured ontologies, such as CIDOC CRM, in enhancing LLM performance. This work provides insights into LLM-driven KG construction for cultural heritage, presenting our prompt-based NER and EL approach, datasets, models, and results.

Keywords: Knowledge Extraction · Large Language Models · Giorgio Vasari · Cultural Heritage · Entity Linking · Named Entity Recognition

1 Introduction

Recent advances in large language models (LLMs) have demonstrated their potential in extracting structured information from unstructured text, particularly in tasks such as Named Entity Recognition (NER) and Entity Linking (EL). However, their performance in historical, multilingual, and culturally dense corpora remains underexplored. In this context, Giorgio Vasari's *Le vite de' più*

eccellenti pittori, scultori e architettori (1568) [17], hereafter referred to as *Le vite*, presents a particularly demanding challenge: The use of 16th-century Florentine Italian, the absence of gold-standard annotated datasets , and the interpretative nature of many English translations, which may diverge significantly from the philological truthfulness of the original text, all complicate information extraction (IE) and semantic interpretation. This study investigates whether LLMs can be reliably used to automatically construct knowledge graphs (KGs) from Vasari's *Le vite*, and more broadly, whether prompt-based approaches can support historically informed, semantically meaningful data extraction in the absence of fine-tuning or annotated corpora. We also explore the role of structured ontologies, as e.g., CIDOC-CRM, in scaffolding LLM performance.

We address the following research questions:
RQ1: How accurately and consistently can LLMs perform NER and EL on multilingual, historical documents?
RQ2: Can reliable ground truths be created without being fluent in a given source language?
RQ3: Can ontologies enhance the evidence- finding process and the reliability of LLM-extracted data?

Through this study, we have provided insights into the viability and reliability of LLM-driven knowledge graph construction from complex historical text, particularly for applications in cultural heritage research. In the following sections, we outline our approach to prompt-based NER and EL, describe the datasets and models employed, and present our results and their implications for extracting historically informed knowledge.

2 Related work

Vasari's *Vite*, presents a rich but challenging source for computational analysis due to its combination of subjective narration, anecdotal storytelling, and historical documentation. As Steptoe [13] notes, Vasari's personal tone and artistic biases complicate efforts to extract verifiable data, such as provenance, chronology, or attribution, but these traits make the text compelling for semantic interpretation. To explore the feasibility of automated IE from historical texts, we leveraged LLMs for automatic named entity recognition (NER) and KG construction (KGC) from Vasari's *Vite*.

2.1 Automatic Knowledge Graph Construction

Automatic Knowledge Graph Construction (KGC) traditionally consists of three main processes: Knowledge Acquisition, Knowledge Refinement, and Knowledge Evolution [20]. For this work, we are most interested in knowledge acquisition, where a text input is transformed into a raw KG. To achieve this, two steps are involved: NER and EL. Optional, further coreference resolution (CO) and relation extraction (RE) can be introduced. **NER** is the process of identifying and categorising key pieces of information in text into predefined categories, as

e.g., the names of people, organisations, locations, dates, and other specific types of data. **EL** is the task of connecting entity mentions in a text to corresponding unique identifiers in a knowledge base including the disambiguation of same name entities referring to different real-world objects. This work assesses the reliability of LLMs for KGC, specifically evaluating their ability to perform NER and EL on a historical text in 16th-century Florentine Italian. Santini et al. [12] introduced structured knowledge extraction techniques to formalise references to artists, artworks, places, and patrons, while subsequent work [9] modeled the implicit artistic and social relationships via semantic graphs, capturing notions like influence and rivalry. Building on this, Santini et al. [10] explored the use of LLMs for NER and EL in historical Italian texts, such as Leopardi's *Zibaldone*, offering methodological insights applicable to current prompting-based experiments on Vasari's narrative. Unlike these works, our research exploits prompting with LLMs as a standalone approach to extract entities from Vasari's historical prose, focusing on evaluating the reliability of LLM outputs and developing guidelines for their critical use in art-historical knowledge extraction.

2.2 LLM for historical NLP

Santini et al. [11] introduced the ExtrART guidelines to identify persons, locations, and artworks in Vasari's biographies. We extended this work by embedding selected definitions and CIDOC-CRM [1] categories into LLM prompts to test whether semantic scaffolding improves model understanding. Wang et al. [18], in their GPT-NER study, showed that carefully designed zero- and few-shot prompts can enable GPT models to achieve competitive NER performance on standard datasets, without the need for fine-tuning. Inspired by this, we examine whether similar prompt-only methods can function effectively in the more unpredictable terrain of Renaissance art texts, where entities are highly domain-specific and the language is stylistically idiosyncratic. In a similar style, Xiao et al. [19] presented LLM-DER, applying prompt-based NER to Chinese coal chemical literature. While their approach successfully adapts prompting to a structured industrial domain, our work diverges by confronting the challenges posed by a limited non-literally translated, and culturally rich corpus. Finally, De Toni et al. [14] inform about the historical NLP challenges and recommends fine-tuning. In our approach, we explore whether prompt engineering combined with ontologies and expert guidelines can enable LLMs to effectively perform historical NER. Thereby, we aim to understand strengths and limitations of LLMs challenging the boundaries of linguistic and cultural heritage data processing.

2.3 Multi-lingual performance of LLMs

Lai et al. [5] comprehensively evaluate ChatGPT on seven knowledge extraction tasks (NER, RE, etc.) across thirty-seven languages, including Italian. The authors compare ChatGPT performance along two axes: cross-linguistic comparison on identical tasks, and English versus language-specific prompts. For NER, ChatGPT shows a 7.3-point F_1 score gap between English (37.2) and other

languages (29.9 average) when using English prompts, though Italian was not explicitly assessed. For RE, ChatGPT outperforms the English baseline by 12.5 points (61.9 vs. 74.4) on Italian with English prompts. Comparing English versus language-specific prompts reveals minimal differences: NER favors English prompts by 1.9 points, while RE favors language-specific prompts by 0.8 points. Chollampat et al. [2] provide a second multilingual LLM performance comparison. The authors propose two metrics: XE using Sentence Embeddings (XESE) and XE by Translation Quality (XETQ), measuring similarity between expected and actual responses and response quality, respectively. They compare XESE and XETQ performance between English and target language prompts. For Italian, all evaluated LLMs, such as BLOOM[7], Llama-2[8], and Gemma[9], reported improved performance when utilised with a prompt in Italian, compared to the English prompt. There are two exceptions to this, namely, both OpenAI GPT-3.5 (`gpt-3.5-turbo-1106`) and GPT-4 (`gpt-4-1106-preview`) were reported to have improved performance with an English prompt, compared to an Italian prompt. These studies [2, 5] provide strong rationale for evaluating our approach on both English translations and original Italian text. They demonstrate that comparing English versus Italian prompts across multiple LLM families (GPT, Gemma, Llama) is essential for comprehensive reliability assessment.

3 Resources

This section describes the key resources employed in our work to evaluate LLM performance on historical NER. We utilise Vasari's seminal work in both its original Italian and English translation, complemented by expert-curated ground truth data and multiple state-of-the-art language models. Additionally, we incorporate the CIDOC-CRM ontology to provide structured semantic guidance for cultural heritage entity recognition.

3.1 Vasari's work and its translations

To evaluate the consistency of LLMs we utilised not only the English but the original 1568 version of Vasari's *Le Vite* [16] via their multilingual open Wikisource[10] transcriptions.

3.2 Ground Truth Dataset

We used a manually validated dataset sampled from the English language translation of Vasari's *Le Vite* curated by art historians focusing on the on NER and EL parts.[11] The dataset includes annotated entity start and end positions for

[7] `https://huggingface.co/bigscience/bloom`
[8] `https://www.llama.com/llama2/`
[9] `https://deepmind.google/models/gemma/`
[10] `https://it.wikisource.org/`
[11] `https://github.com/ISE-FIZKarlsruhe/vasari\_nlp/`

names of people and locations, along with their corresponding text segments. Where available, each entity is linked to its Wikidata identifier. The dataset contains entities extracted from 53 sections across 27 chapters of Vasari's work, with each section containing multiple sentences. Each section's source chapter is indicated by its English title.

3.3 Large Language Models

During the development of our prompting strategies multiple closed source LLMs were utilized. These were OpenAI's GPT-4o, Google's Gemini Pro and Gemini 2.5 Flash, DeepSeek-R1 from Hangzhou and Llama 3.3 70B. These models were selected for their efficiency and accessibility, which facilitated rapid prototyping and iteration. All models except Llama 3.3 were accessed via their respective graphical user interface and their outputs compiled manually. Llama 3.3 was accessed through HuggingFace chat[12].

3.4 CIDOC-CRM Ontology

The CIDOC Conceptual Reference Model (CIDOC-CRM) [1] is a high-level conceptual model ontology for the cultural heritage (CH) domain, formally recognised as an international standard (ISO 21127:2023). Its primary purpose is to act as a "semantic glue", enabling the integration and interoperability of data from diverse sources such as museums, archives, and libraries. As an official standard, the definitions of its classes and properties are meticulously crafted and globally recognised, providing a stable and authoritative vocabulary. Within the scope of this research, we leverage fundamental CRM classes, such as *E21_Person* for artists and collectors, *E5_Event* for historical occurrences, including exhibitions or sales, and *E53_Place* for significant locations.

4 Approach

This study develops a prompt-based IE pipeline for Giorgio Vasari's *Le vite* (1568), a foundational yet linguistically and structurally complex text in early modern art historiography. Our goal is to construct a semantically enriched KG grounded in the original Italian corpus, enabling structured access to entities such as artists, locations, artworks, and historical events. The pipeline comprises four main stages: (1) aligning a high-quality English ground truth with the Italian source text, (2) extracting CIDOC-aligned named entities using LLMs, (3) performing EL to external knowledge bases, and (4) generating an RDF-based KG to enable semantic querying and interoperability. Each component is fully implemented via prompt engineering and light-weight NLP methods, allowing for cross-lingual, culturally sensitive processing without the need for supervised retraining or domain-specific annotation.

[12] `https://huggingface.co/chat/`

4.1 Multilingual Dataset Creation

The construction of a multilingual ground truth from Giorgio V asari's *Le vite* [17] poses unique challenges due to the richness, length, and early modern Italian language of the text. Its original form presents significant barriers to computational processing and structured IE. To facilitate more accessible and systematic exploration of Vasari's work, we employed NLP methods – particularly LLMs – in an attempt to extract and structure biographical and artistic information into a machine-readable KG. We implemented generalised methods enabling cross-lingual mapping and dataset creation by leveraging a manually-annotated English ground truth corpus[13]. We implemented Maximised Substring Mapping, Fuzzy Mapping, LLM-Assisted Entity Projection, and Wikidata-Based Entity Matching (described in subsequent sections) to locate high-confidence correspondences between English ground truth entities and their Italian equivalents. Upon successful alignment, we would construct a KG using standard IE techniques, transforming plaintext into an interconnected network of detected `person`, `location`, `organisation`, and `miscellaneous` entities. This KG, grounded in established ontological frameworks (cf. Section 3.4), supports semantic querying and reasoning across biographical and art historical data, enhancing accessibility and analyses for researchers studying Vasari's work and digital humanities more broadly.

Maximised Substring Mapping Our *Maximised Substring Mapping* method is a deterministic approach identifying the longest lexical overlap between annotated English documents and Vasari's original Italian treatise. We translate annotated English documents into Italian using an LLM (e.g., DeepSeek), then apply a substring expansion algorithm over the corresponding Italian chapter text. This method searches for exact matches, maximizing character-level continuity between translated and source texts. This technique proves effective for named entities like person and place names, which remain relatively stable across translations. The maximised substring heuristic minimizes ambiguity and enhances reproducibility by reducing context-sensitive interpretation dependency.

Fuzzy Mapping To complement exact matching, we implement fuzzy matching using similarity metrics (e.g., Levenshtein distance). It handles cases where morphological variation, paraphrasing, or stylistic differences between English annotations and Italian sources prevent exact substring alignment in our noisy source document.Fuzzy Mapping applies after filtering candidate spans by lexical and semantic constraints, with alignment confidence quantified through similarity score thresholding. For descriptive phrases, indirect references, or historical entity name variants, it significantly improves recall while maintaining precision.

LLM-Assisted Entity Projection LLM-Assisted Entity Projection combines candidate span selection with LLM inference and validation. Italian text spans

[13] `https://github.com/ISE-FIZKarlsruhe/vasari\_nlp/`

are selected using previously described string similarity methods, then English ground truth entities are provided to a language model prompted to locate these entities within candidate spans. Output is verified through direct string matching: entities with exact or near-exact matches in Italian text are retained. Only confirmed entities are included in the resulting Italian ground truth set.

Wikidata-Based Entity Matching Wikidata-Based Entity Matching uses multilingual alias data from Wikidata for knowledge-based cross-lingual alignment. For each annotated entity in English sections, Italian and English alternative names are retrieved via the Wikidata API. An equal number of Italian sentences are selected, then fuzzy string matching identifies tokens closely corresponding to retrieved aliases. Matches exceeding a predefined similarity threshold are retained. This process repeats across multiple Italian candidate sections until all entities are located or potential matches exhausted. The section with the highest number of matched entities becomes the aligned counterpart, with its identified entities forming the projected ground truth.

4.2 Named Entity Recognition

While traditional NER methods rely on annotated corpora and supervised learning, our approach uses the contextual and multilingual capabilities of LLMs via prompt engineering. This allows us to address the semantic complexity of Vasari's text, marked by indirect references, multilingual usage (e.g., early modern Italian, original Italian, English), and culturally embedded knowledge, without the need for fine-tuning or domain-specific training data. Prompting provides a lightweight yet effective way to align general-purpose LLMs with the interpretive goals of digital humanities research.

LLM-Based NER Approach Prompt-based NER addresses Vasari corpus's unique challenges, including referential ambiguity and historically situated descriptions of people, places, artworks, and events. Using structured prompts informed by cultural heritage ontologies, we achieve multilingual, context-sensitive entity recognition and produce explainable outputs for scholarly analysis.

Prompt Design and Evolution Prompt development underwent iterative refinements to enhance precision, interpretability, and reliability:

1. Minimal Prompt (Zero-Shot) Initial prompts extracted named entities but lacked semantic filters. Recognition was inconsistent, especially for collective categories like artworks and events, and failed to resolve indirect references (e.g., *frate* for *Fra Giovanni Agnolo*).

2. CIDOC-Aligned Prompt (Few-Shot) To increase conceptual clarity, CIDOC-CRM classes were introduced to guide the models:

1. Persons (`E21`), including mythological, religious, and historical figures

2. Locations (`E53`) with geographical or institutional significance
3. Artworks (`E22`) such as paintings, buildings, and sculptural works
4. Events (`E5`) framed as commissions or other temporally bounded actions

To address the frequent indirect or anaphoric references in Vasari's text, **coreference resolution** was incorporated into the few-shot approach. This mechanism aimed to improve entity continuity, linking expressions like "the friar," "the master," or "it" to their canonical antecedents, and to enhance **explainability** by requiring contextual justifications for each resolution.

3. Full-Contextual Prompt (Few-Shot) The final version of the prompt integrated a fully structured, semantically grounded approach, incorporating three key enhancements aimed at improving accuracy, reliability, and explainability across all four CIDOC-aligned entity categories:

1. **Structured definitions**; for each CIDOC class (`E21 Person`, `E53 Place`, `E22 Artworks`, `E5 Event`), including explicit inclusion and exclusion criteria, and examples tailored to the Vasari corpus.
2. **Coreference resolution**; applied systematically across all categories, enabling the model to resolve pronouns, descriptive titles, and indirect references (e.g., "*the chapel*" → Brancacci Chapel). This step is aimed to improve continuity and minimize fragmentation by linking entities across long narrative spans. Furthermore, it enhances transparency through the inclusion of justification strings, explaining how each reference was resolved.
3. **Reliability scores** (ranging from 0 to 1) assigned to each identified entity, reflecting the model's contextual certainty based on clarity of reference and narrative context. This feature allows for post-hoc filtering of low-confidence results and add a layer of interpretability crucial for scholarly applications, where ambiguous identifications must be either verified or discarded.

The output of this prompt was returned in a structured JSON format, providing a flexible and explainable structure suitable for integration into knowledge graphs or digital editions of Vasari's text.

4.3 Entity Linking

Building on our structured NER output, we apply prompt-based Entity Linking (EL) to entities categorised as `E21 Person`. This step extends our fully prompt-driven pipeline and investigates the theoretical potential of LLMs to resolve historically and culturally complex references using minimal input.

LLM-Based EL The goal is to determine whether modern LLMs can reliably link figures from the Vasari corpus to their corresponding Wikidata Q-IDs and Wikipedia URLs, based solely on canonical name forms and contextual cues. Beyond evaluating model behaviour, this process plays a foundational role in KGC by aligning textual entities with globally recognised identifiers, enabling structured integration, semantic querying, and linkage to external cultural heritage datasets.

Prompt Structure For each set of `E21 Person` entities, we submitted the following prompt to four contemporary LLMs—GPT-4o, LLaMA 3.3, DeepSeek V3, and Gemini 2.5 Flash: *"Provide the Wikipedia URL's and Wikidata Q-ID's for the following entities: '[Enter entities in JSON format here]'"* This input derives directly from final-stage NER output, including semantically scoped, contextually grounded names from the Vasari corpus. No additional metadata or qualifiers are provided, allowing assessment of the model's disambiguation and entity resolution abilities based solely on internal representations and general knowledge. We examine whether LLMs can reliably identify historically grounded entities, their performance across knowledge bases (Wikidata vs. Wikipedia), and whether architectural differences affect handling of multilingual and early modern references.

4.4 Knowledge Graph Construction

To promote interoperability and long-term reuse, we converted the manually annotated Vasari-NLP English dataset into a KG using the NLP Interchange Format (NIF)[14]. This format shift enables structured linguistic and semantic annotations to be represented in RDF, facilitating integration with semantic web tools and workflows. Each document is modeled as a `nif:Context`, with entity spans as `nif:Phrase` resources linked to standard vocabularies (e.g., DBpedia Ontology, schema.org). This transformation preserves the original annotation semantics while enabling SPARQL querying, benchmarking, and linkage to external knowledge bases. The resulting NIF KG is published[15] under the MIT License to encourage reuse and adaptation in both academic and applied research contexts. It serves as a foundation for benchmarking EL and IE systems, and as a semantic layer for exploring Vasari's *Vite* within digital humanities.

5 Results and Discussion

Although we were not able to conduct a complete quantitative evaluation of the proposed pipeline, the development process yielded several insights relevant to the application of IE and KGC in the context of long-tail, domain-specific entities in historical texts. Addressing RQ1, we transformed an existing English-language ground truth into a NIF-compliant KG and benchmark our method on it. We demonstrated the feasibility of repurposing curated data for semantic web applications. The resulting resource is interoperable, queryable, and suitable for benchmarking in future EL and IE tasks. In Table 1, we evaluated our LLM-based NER method against traditional annotators for an excerpt of Vasari's *Vite*. We note that Deepseek V3 (70.97%) performs best out of our selection of LLMs, yet vastly underperforms in comparison to traditional approaches, such as DBpedia Spotlight (52.05%), Babelfy (74.36%) and top-performing annotator WAT

[14] `https://persistence.uni-leipzig.org/nlp2rdf/`
[15] `https://github.com/kmdn/isws_vulkan`

(92.68%). All LLMs reach perfect precision scores, but display underperforming recall values, leading to subpar F_1 scores. As such, we observe that LLMs in our experiments were reluctantly, but correctly recommending named entities. We also provide results to the EL task in our appendix (see Fig. 1, 2, 3, 4 and 5).

Second, all alignment methods in this paper proved insufficient in some manner. Consequently, we could not provide positive results for RQ2. Maximized substring matching and fuzzy matching struggled to find large and meaningful spans, although the latter showed promise as it managed to identify a few complete sentences. LLM-Assisted Entity Projection proved to be unreliable as LLMs were unable to consistently identify entities even across multiple runs over the same text. Wikidata-based Entity Matching, while promising, is prohibitively computationally expensive and was therefore not applicable in the project's context. Regarding RQ3, we build on the wide use of the CIDOC-CRM ontology, providing the LLM with a domain-relevant definition of concepts, reflecting the manual process of annotating named entities by an expert in the domain (RQ3). Moreover, the integration of LLMs facilitated the handling of such complexity by enabling context-aware cross-lingual mapping, albeit with variable reliability depending on entity type and phrasing. This underscores both the promise and the current limitations of LLM-based methods in structured data extraction from historical humanities corpora. Finally, the absence of an evaluation step limits the conclusions draw regarding accuracy and generalisability of our approach. However, the modular nature of our pipeline and the public availability of the NIF KG provide a solid foundation for future work. In particular, the resource may support the development of Italian-language gold standards, comparative system benchmarking, and deeper semantic enrichment of early modern art-historical texts.

We evaluated traditional annotators as well as our LLM-based NER method on a small-scale benchmark (Vasari EN (Mini)) and report results in Table 1. While our prompt-based method achieved slightly lower F_1 scores compared to token-level baselines, this reduction reflects a deliberate shift in emphasis from surface-level string matching to semantically motivated reference resolution. In particular, our model's ability to produce coreference rationales introduces a novel analytical dimension: it enables the explicit tracing of indirect or anaphoric references (e.g., pronouns such as "*he*" or descriptors like "*the master*") back to canonical entities. This feature offers significant utility for art historians and digital humanists, as many apparent evaluation mismatches stem not from model failure, but from correct identification of referents made explicit earlier in the text. By providing these resolution chains with justifications, the model enables new engagement with historical narrative structure, allowing tracking of how Vasari refers to individuals across long discursive spans. While traditional metrics may underrepresent interpretive performance, the outputs enable more nuanced, transparent, and historically informed entity analysis.

We also experimented with end-to-end KGC using LLMs. Although the output was incomplete and contextually limited, it demonstrated the dangers of using such tools without critical oversight. This highlights the need for robust,

generalisable guidelines; without them, LLM-generated results risk compromising academic integrity and misrepresenting historical truth. We evaluated four LLMs on EL for extracted person entities, measuring their ability to associate names with correct Wikipedia URLs and Wikidata IDs. Gemini 2.5 Flash performed best, producing the most accurate links for both. ChatGPT-4o followed, with reasonable Wikipedia linking but inconsistent and occasionally fabricated Wikidata IDs. DeepSeek-V3 showed similar issues, while LLAMA 3.3 70B failed to return any correct IDs. Most LLMs retrieved relevant Wikipedia URLs but frequently hallucinated Wikidata IDs, likely due to Wikidata's less commonly seen entries compared to Wikipedia's narrative-rich, widely referenced content. Notably, Gemini 2.5 Flash demonstrated useful disambiguation in complex cases like "Pope Clement," providing explanations listing multiple possibilities (e.g., Clement VII, Clement VIII) with corresponding Wikidata Q-IDs. This contextual awareness supports its reliability for historical entity linking.

6 Conclusions and Future work

In this work, we present a proof-of-concept for applying IE techniques to KGC involving long-tail entities within a major art-historical text from Giorgio Vasari. We developed a functional pipeline and transformed an existing manually annotated ground truth into a NIF-based KG to support structured semantic representation. However, due to time and scope constraints, we were not able to conduct a quantitative evaluation of the extraction or alignment results.

7 AI Disclaimer

Portions of this manuscript were refined using LLMs to improve clarity, coherence, and style. All conceptual development, experimental design, data analysis, and conclusions were conducted and verified by the authors.

References

1. Version 7.1.3 | CIDOC CRM, https://cidoc-crm.org/Version/version-7.1.3
2. Chollampatt, S., Pham, M.Q., Indurthi, S.R., Turchi, M.: Cross-lingual Evaluation of Multilingual Text Generation. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 7766–7777. Association for Computational Linguistics, https://aclanthology.org/2025.coling-main.520/
3. Flati, T., Navigli, R.: Three Birds (in the LLOD Cloud) with One Stone: BabelNet, Babelfy and the Wikipedia Bitaxonomy. In: Proceedings of the Posters and Demos Track of 10th International Conference on Semantic Systems. SEMANTiCS'14, vol. 1224, pp. 10–13. CEUR-WS.org (2014)
4. van Hulst, J.M., Hasibi, F., Dercksen, K., Balog, K., de Vries, A.P.: Rel: An entity linker standing on the shoulders of giants. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '20, ACM (2020)

5. Lai, V.D., Ngo, N.T., Veyseh, A.P.B., Man, H., Dernoncourt, F., Bui, T., Nguyen, T.H.: Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning (2023), https://arxiv.org/abs/2304.05613
6. Mendes, P.N., Jakob, M., García-Silva, A., Bizer, C.: DBpedia Spotlight: Shedding Light on the Web of Documents. In: Proceedings the 7th International Conference on Semantic Systems. pp. 1–8. I-SEMANTICS'11, ACM (2011)
7. Piccinno, F., Ferragina, P.: From tagme to WAT: a new entity annotator. In: ERD'14, Proceedings of the First ACM International Workshop on Entity Recognition & Disambiguation, July 11, 2014, Gold Coast, Queensland, Australia. pp. 55–62. ACM (2014). https://doi.org/10.1145/2633211.2634350, https://doi.org/10.1145/2633211.2634350
8. Rizzo, G., et al.: NERD: a framework for unifying NERD extraction tools. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics
9. Santini, C., Garay, N., Posthumus, E., Sack, H.: The Art of Relations . https://doi.org/10.5281/ZENODO.10698245, https://zenodo.org/doi/10.5281/zenodo.10698245
10. Santini, C., Melosi, L., Frontoni, E.: Named entity recognition in historical italian: The case of giacomo leopardi's zibaldone (2025), https://arxiv.org/abs/2505.20113
11. Santini, C., Tan, M.A., Bruns, O., Tietz, T., Posthumus, E., Sack, H.: Guidelines for the Annotation of ExtrART: Evaluation Dataset for Entity Extraction from The Lives Of The Artists (1550). https://doi.org/10.5281/ZENODO.7642989, https://zenodo.org/record/7642989
12. Santini, C., Tan, M.A., Tietz, T., Bruns, O., Posthumus, E., Sack, H.: Knowledge extraction for art history: the case of vasari's the lives of the artists (1568). In: Qurator Conference 2022: Third Conference on Digital Curation Technologies ; 19–23 September 2022, Berlin, Germany (2022)
13. Steptoe, A.: Genius and the MindStudies of Creativity and Temperament. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198523734.001.0001, http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198523734.001.0001/acprof-9780198523734
14. Toni, F.D., Akiki, C., de la Rosa, J., Fourrier, C., Manjavacas, E., Schweter, S., van Strien, D.: Entities, dates, and languages: Zero-shot on historical texts with t0 (2022), https://arxiv.org/abs/2204.05211
15. Usbeck, R., Röder, M., Ngomo, A.N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., Wesemann, L.: GERBIL: General Entity Annotator Benchmarking Framework. In: Proceedings of the 24th International Conference on World Wide Web. pp. 1133–1143. WWW'15, ACM (2015)
16. Vasari, G.: Lives of the Most Eminent Painters, Sculptors and Architects: Tr. from the Italian of Giorgio Vasari, vol. 2. HG Bohn (1851)
17. Vasari, G., Frey, K., Frey, K.: Le vite de'più eccellenti pittori, scultori e architettori, vol. 3. Giunti Florence (1967)
18. Wang, S., Sun, X., Li, X., Ouyang, R., Wu, F., Zhang, T., Li, J., Wang, G.: Gptner: Named entity recognition via large language models (2023), https://arxiv.org/abs/2304.10428

19. Xiao, L., Xu, Y., Zhao, J.: Llm-der:a named entity recognition method based on large language models for chinese coal chemical domain (2024), https://arxiv.org/abs/2409.10077

20. Zhong, L., Wu, J., Li, Q., Peng, H., Wu, X.: A Comprehensive Survey on Automatic Knowledge Graph Construction **56**(4), 1–62. https://doi.org/10.1145/3618295, https://dl.acm.org/doi/10.1145/3618295

Appendix

RESULTS - Zero Shot - PERSON -ENTITIES - ENG

Person	DeepSeek - V3	Gemini 2.5 Flash	ChatGPT - 4o	LLAMA 3.37B
Pope Clement	✅	✅	✅	✅
Buonarroti	✅	✅	✅	❌
Michelagnolo	✅	✅	✅	✅
Fra Giovanni Agnolo PER	✅	✅	✅	✅
Antonio da San Gallo PER	✅	✅	✅	✅
Frate	⚠️	⚠️	⚠️	⚠️
Duke Lorenzo	✅	✅	✅	✅
Duke Giuliano	✅	✅	✅	✅
Cosimo	✅	❌	❌	❌
Damiano	✅	❌	❌	❌
Madonna	❌	❌	❌	❌
Raffaello da Montelupo	✅	✅	✅	✅
Pope	❌	⚠️	⚠️	❌

Fig. 1. Results for NER using Zero-Shot Prompt for English text

Minimal Prompt for 'Person'

Task:
You are an expert in Renaissance art history and your task is to perform Named Entity Recognition (NER) on a text by Giorgio Vasari. Your goal is to identify and extract all [TARGET] entities.
Output Format: Return a JSON object containing:
1. A list of unique [TARGET] entities
2. A list of contextual triples (optional)
Now, analyze the following text and provide the output in the specified JSON format: [VASRI's Text]

RESULTS - WITHOUT RELIABILITY - PERSON -ENTITIES - ENG

Person	DeepSeek - V3	Gemini 2.5 Flash	ChatGPT - 4o	LLAMA 3.37B
Pope Clement	✅	✅	✅	✅
Buonarroti	✅	✅	❌	✅
Michelagnolo	✅	✅	✅	✅
Fra Giovanni Agnolo PER	✅	✅	✅	✅
Antonio da San Gallo PER	✅	✅	✅	✅
Frate	⚠️	⚠️	⚠️	⚠️
Duke Lorenzo	✅	✅	✅	✅
Duke Giuliano	✅	✅	✅	✅
Cosimo	✅	✅	✅	❌
Damiano	✅	✅	✅	❌
Madonna	✅	✅	❌	❌
Raffaello da Montelupo	✅	✅	✅	✅
Pope	❌	⚠️	⚠️	❌

Fig. 2. Results for NER using Few-shot Prompt without Reliability Score for English text

CIDOC-Aligned Prompt for 'Person'

Target:
act all "Person" entities mentioned in the provided text, using a semantic framework inspired by the CIDOC Conceptual Reference Model (CIDOC CRM).
Definition of a "Person" Entity: A "Person" corresponds to class E21 Person in CIDOC CRM: "An individual human being who is documented in historical, artistic, or legendary contexts, including real, fictional, or mythological individuals, when referenced as discrete agents or subjects."
1. Core Identity: Include all individual human agents identifiable as distinct persons in the text, such as: Historical figures (artists, patrons), Mythological or religious figures treated as individuated beings (e.g., Apollo, St. Francis, Archangel Michael). CIDOC Reference: Instances of E21 Person can be related to artworks or events through properties such as: P14 carried out, P107 has current or former member, P131 is identified by.
2. Forms of Reference: Extract all textual references to E21 Person instances, including, proper names, nicknames (e.g., "Tintoretto") and specific, contextually individuated titles (e.g., "the Pope," "the Duke," "the master"), provided the reference is disambiguated within the context.
3. Exclusions: Do not extract generic collectives (e.g., "the painters of Siena," "his disciples"). Instead, add contextual triples using CIDOC-style modeling (e.g., (Duccio, P107i is member of, the painters of Siena)). Plural family names without clear individual referents (e.g., "the Gaddi"). Giorgio Vasari should be excluded.
Instructions:
1. Read the text carefully.
2. Identify all entity strings matching the above "Person" definition.
3. Coreference Resolution: If a pronoun (e.g., "he," "his"), generic title (e.g., "the painter," "the friar"), or nickname (e.g., "Il Sodoma") clearly refers to a previously introduced individual, resolve it and add the canonical form of the person's name. In an auxiliary table, add a brief rationale (5–10 words) for each resolved reference.
Output Format: Return a JSON object containing:
1. A list of unique "Person" entities
2. A list of contextual triples (optional)
3. A list of coreference rationales
Now, analyze the following text and provide the output in the specified JSON format: "[VASRI's Text]"

RESULTS - FINAL - PERSON -ENTITIES - ITA

Person	DeepSeek - V3	Gemini 2.5 Flash	ChatGPT - 4o	LLAMA 3.37B
Pope Clement	✅	✅	✅	✅
Buonarroti	✅	❌	✅	❌
Michelagnolo	✅	✅	✅	✅
Fra Giovanni Agnolo PER	✅	✅	✅	✅
Antonio da San Gallo PER	✅	✅	✅	✅
Frate	⚠️	⚠️	⚠️	⚠️
Duke Lorenzo	✅	✅	✅	✅
Duke Giuliano	✅	✅	✅	✅
Cosimo	✅	✅	✅	❌
Damiano	✅	✅	✅	❌
Madonna	✅	✅	❌	❌
Raffaello da Montelupo	✅	✅	✅	✅
Pope	❌	⚠️	⚠️	❌

Fig. 3. Results for NER using Few-Shot Prompt with Reliability Score for English text

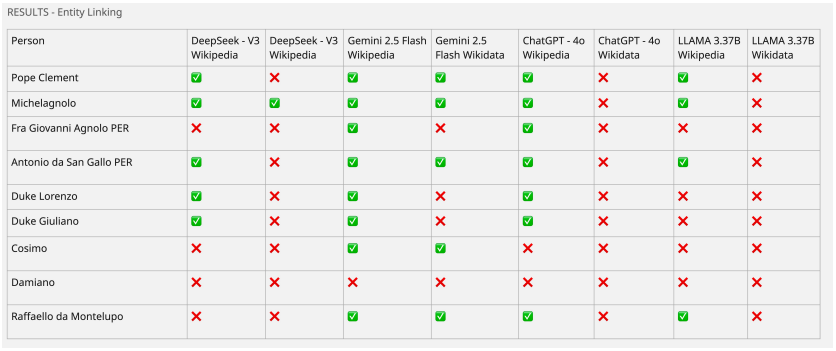
RESULTS - FINAL - PERSON -ENTITIES - ENG

Person	DeepSeek - V3	Gemini 2.5 Flash	ChatGPT - 4o	LLAMA 3.37B
Pope Clement	✅	✅	✅	✅
Buonarroti	✅	✅	✅	❌
Michelagnolo	✅	✅	✅	✅
Fra Giovanni Agnolo PER	✅	✅	✅	✅
Antonio da San Gallo PER	✅	✅	✅	✅
Frate	⚠️	⚠️	⚠️	⚠️
Duke Lorenzo	✅	✅	✅	✅
Duke Giuliano	✅	✅	✅	✅
Cosimo	✅	✅	✅	✅
Damiano	✅	✅	✅	✅
Madonna	✅	❌	❌	❌
Raffaello da Montelupo	✅	✅	✅	✅
Pope	❌	⚠️	⚠️	⚠️

Fig. 4. Results for NER using Few-Shot Prompt with Reliability Score for Italian Translated text

Full-Contextual Prompt for 'Person'

Task:
You are an expert in Renaissance art history and your task is to perform Named Entity Recognition (NER) on a text by Giorgio Vasari. Your goal is to identify and extract all [TARGET] entities mentioned in the provided text, using a semantic framework inspired by the CIDOC Conceptual Reference Model (CIDOC CRM).
Definition of a [TARGET] Entity:
A "Person" corresponds to class E21 Person in CIDOC CRM: "An individual human being who is documented in historical, artistic, or legendary contexts, including real, fictional, or mythological individuals, when referenced as discrete agents or subjects."
1. Core Identity: Include all individual human agents identifiable as distinct persons in the text, such as: Historical figures (artists, patrons), Mythological or religious figures treated as individuated beings (e.g., Apollo, St. Francis, Archangel Michael). CIDOC Reference: Instances of E21 Person can be related to artworks or events through properties such as: P14 carried out, P107 has current or former member, P131 is identified by.
2. Forms of Reference: Extract all textual references to E21 Person instances, including, proper names, nicknames (e.g., "Tintoretto") and specific, contextually individuated titles (e.g., "the Pope," "the Duke," "the master"), provided the reference is disambiguated within the context.
3. Exclusions: Do not extract generic collectives (e.g., "the painters of Siena," "his disciples"). Instead, add contextual triples using CIDOC-style modeling (e.g., (Duccio, P107i is member of, the painters of Siena)). Plural family names without clear individual referents (e.g., "the Gaddi"). Giorgio Vasari should be excluded.
Instructions:
1. Read the text carefully.
2. Identify all entity strings matching the above "Person" definition.
3. Assign a reliability score to each extracted entity between 0 (low certainty) and 1 (high certainty) based on clarity of reference and context.
4. Coreference Resolution: If a pronoun (e.g., "he," "his"), generic title (e.g., "the painter," "the friar"), or nickname (e.g., "Il Sodoma") clearly refers to a previously introduced individual, resolve it and add the canonical form of the person's name. In an auxiliary table, add a brief rationale (5–10 words) for each resolved reference.
Output Format:
Return a JSON object containing:
1. A list of unique "Person" entities with reliability scores
2. A list of contextual triples (optional)
3. A list of coreference rationales
Now, analyze the following text and provide the output in the specified JSON format: [VASRI's Text]



RESULTS - Entity Linking

Person	DeepSeek - V3 Wikipedia	DeepSeek - V3 Wikipedia	Gemini 2.5 Flash Wikipedia	Gemini 2.5 Flash Wikidata	ChatGPT - 4o Wikipedia	ChatGPT - 4o Wikidata	LLAMA 3.37B Wikipedia	LLAMA 3.37B Wikidata
Pope Clement	✅	❌	✅	✅	✅	❌	✅	❌
Michelagnolo	✅	✅	✅	✅	✅	❌	✅	❌
Fra Giovanni Agnolo PER	❌	❌	✅	❌	✅	❌	❌	❌
Antonio da San Gallo PER	✅	❌	✅	✅	✅	❌	✅	❌
Duke Lorenzo	✅	❌	✅	❌	✅	❌	❌	❌
Duke Giuliano	✅	❌	✅	❌	✅	❌	❌	❌
Cosimo	❌	❌	✅	✅	❌	❌	❌	❌
Damiano	❌	❌	❌	❌	❌	❌	❌	❌
Raffaello da Montelupo	❌	❌	✅	✅	✅	❌	✅	❌

Fig. 5. Results for Entity Linking using LLM

Table 1. NER Evaluation: Micro-F_1 results on Vasari EN (Mini) from GERBIL [15]. We include *Classic* annotators (e.g., Babelfy), as well as *LLM-based* ones.

Classic		**LLM-based**	
System	F_1	System	F_1
Babelfy [3]	0.7436	Gemini 2.5	0.6667
DBpedia Spotlight [6]	0.5205	Deepseek V3	**0.7097**
WAT [7]	**0.9268**	ChatGPT-4o	0.6667
NERD-ML [8]	0.0000	Llama 3.3	0.6207

Table 2. Entity Recognition: Micro-F1 results on Vasari EN. LLM-based approaches not included due to execution times.

System	F1
Babelfy [3]	0.5860
DBpedia Spotlight [6]	0.5850
WAT [7]	**0.8839**
NERD-ML [8]	0.0000
REL [4]	0.6393