A Tool-Augmented AI Agent for Grounded Access to European Cultural Heritage

Pierluca Landolfi[1 [0009−0008−2996−3835]], Fabio Porcelli[1 [0009−0008−5250−8662]], Simone Cioffi[1 [0009−0007−4054−7560]], Laura Di Cuffa[2 [0009−0004−5988−6245]], Maria Volpe[2 [0009−0000−5263−9373]], Emanuel Di Nardo[1 [0000−0002−6589−9323]], and Angelo Ciaramella[1 [0000−0001−5592−7995]]

[1] University of Naples Parthenope, Naples NA 80143, Italy
{emanuel.dinardo}@uniparthenope.it,
{pierluca.landolfi}@studenti.uniparthenope.it
[2] University of Naples L'Orientale, Naples NA, 80134 Italy

Abstract. Large Language Models (LLMs) have demonstrated impressive conversational abilities, yet they remain prone to *hallucination*, the generation of plausible but factually incorrect information. This limitation is particularly critical in the cultural heritage domain, where the accuracy and reliability of historical and artistic data are fundamental requirements. In this paper we present a tool-augmented AI agent designed to answer natural-language queries about European artists and artworks by grounding every response on verified metadata retrieved from the Europeana digital platform. The system integrates the CrewAI agent framework with a GPT-family language model and the Europeana REST API, constraining the model to a single external tool call per conversational turn to ensure deterministic and verifiable behaviour. A lightweight short-term memory mechanism enables coherent multi-turn interactions without context contamination. Empirical evaluation on a systematic benchmark of 150 queries spanning 50 European artists demonstrates perfect tool discipline (100.0%), near-universal source adherence (98.7%), and complete temporal coverage (100.0%), compared with 87.3% temporal coverage for an isolated LLM baseline that cannot cite verifiable sources. The Streamlit-based interface exposes technical provenance details to the user, fostering transparency and trust. Our results confirm the viability of tool-augmented generation as a mediator between structured cultural archives and lay audiences, contributing to the democratisation of digital heritage in the Digital Humanities field.

Keywords: Tool-augmented generation · AI agents · Digital humanities · Cultural heritage · Hallucination mitigation · Large Language Models

1 Introduction

The rapid proliferation of Large Language Models (LLMs) has radically reshaped the landscape of human-computer interaction, enabling systems that understand

and generate natural language with unprecedented fluency [3]. Conversational assistants built on such foundations are increasingly deployed in information-intensive domains, from medicine to legal research, where the reliability of generated content is non-negotiable. Among these domains, *cultural heritage* stands out for its dual nature: it is at once a repository of authoritative knowledge, codified in curated digital archives, and a subject of broad public interest that demands accessible, engaging presentation.

The European cultural heritage is particularly rich in this regard. Platforms such as Europeana aggregate more than fifty million digitised objects, paintings, manuscripts, photographs, and artefacts, contributed by thousands of cultural institutions across Europe [7]. Despite their considerable scope, these archives are largely underutilised by non-specialist audiences owing to the technical complexity of their query interfaces and metadata schemas [10].

Contemporary LLMs could, in principle, bridge this gap by offering natural-language access to cultural knowledge. In practice, however, general-purpose models suffer from a well-documented failure mode known as *hallucination*: the generation of confident but factually incorrect statements [11]. When applied to cultural heritage queries, this phenomenon is particularly harmful, since erroneous attributions of authorship, dating, or provenance may mislead users and undermine the authority of underlying archival sources. Bender et al. [2] have highlighted the broader risks posed by statistical text generators operating without grounding mechanisms, an observation that motivates the present work.

Tool-augmented generation [17] addresses this limitation by equipping language models with access to external knowledge sources at inference time, effectively constraining the generation process to information retrieved from trusted databases. The ReAct paradigm [26] formalises this idea by interleaving reasoning traces with tool-invocation actions, producing outputs that are both interpretable and grounded in verifiable evidence.

In this paper we present a tool-augmented AI agent specifically designed for the cultural heritage domain.[3] The proposed system integrates a GPT-family language model with the Europeana REST API through the CrewAI orchestration framework, enabling users to query artists and artworks in natural language while receiving responses anchored to verified, citable metadata from Europeana's curated archives. The principal contributions of this work are:

1. A domain-specific architecture coupling a GPT-family LLM with the Europeana API under a strict single-call-per-turn constraint, reducing hallucination while preserving conversational fluency.
2. A cascading retrieval strategy that maximises recall from Europeana's heterogeneous metadata by progressively relaxing query filters.
3. A lightweight conversational memory module that maintains cross-turn coherence without accumulating irrelevant context.
4. An empirical evaluation comparing the agent against a standalone LLM baseline on a culturally diverse query set, with qualitative analysis of failure modes.

[3] Source code available at `https://github.com/CI-SSLab/tool-augm-europeana`.

2 Related Work

Hallucination in Large Language Models, the tendency to produce factually incorrect yet convincing outputs, has been surveyed by Ji et al. [11] and critically analysed by Bender et al. [2], who highlight the need for external validation mechanisms. Within the cultural heritage domain these risks are amplified by the specificity and verifiability of historical facts.

Retrieval-Augmented Generation (RAG) [12, 1] addresses this limitation by augmenting LLMs with a retrieval component, significantly improving factual accuracy on knowledge-intensive tasks. Schick et al. [17] extended this idea to heterogeneous tool use, training models to invoke APIs and calculators autonomously, while the ReAct framework [26] combines chain-of-thought reasoning with action sequences. The proposed approach applies tool-augmented generation to a curated, domain-specific API, trading breadth for provenance guarantees.

LLM-based agents exploit foundation models as a cognitive core, augmenting them with memory, planning, and tool-use modules [25, 22]. The agent reasons about how to decompose a task, decides when and which tools to invoke, and synthesises retrieved information into a coherent response. Multi-agent frameworks such as CrewAI [5] formalise this architecture through specialised agents with distinct roles, goals, and toolsets orchestrated by a shared execution pipeline.

The expressiveness and broad knowledge encoded in LLMs make them natural candidates for domain-specific applications, including cultural heritage. In such contexts, however, generative fluency alone is insufficient: responses must be traceable to authoritative sources, since errors in attribution or dating can mislead users and erode institutional trust. The Europeana Data Model (EDM) [10] provides the semantic backbone for Europe's primary cultural aggregation platform, enabling structured queries over millions of digitised objects. While EDM ensures interoperability across institutions, the complexity of its query syntax limits direct access for non-expert users, creating a gap that a natural-language agent is well positioned to fill. Transparency in how AI systems handle such sensitive knowledge is equally important: the UNESCO Recommendation on the Ethics of Artificial Intelligence [21] underscores the need for trustworthy and accountable AI in domains of high cultural relevance, a principle that directly informs the provenance-disclosure interface of the proposed system.

A dedicated line of research has specifically addressed the challenges of natural-language access to cultural heritage repositories. Sporleder [18] provides an early survey of NLP applications to museum and archival data, highlighting difficulties specific to the domain: sparse and multilingual catalogue records, historical language variation, and the absence of domain-adapted evaluation benchmarks. Hyvönen [9] systematically addresses the publication and consumption of cultural heritage Linked Open Data on the Semantic Web, establishing the semantic infrastructure on which aggregation platforms such as Europeana are grounded. More recently, conversational and generative approaches have gained traction: Trichopoulos et al. [20] deploy and evaluate LLM-driven chatbots across three museum installations, a digital art exhibition, a painting

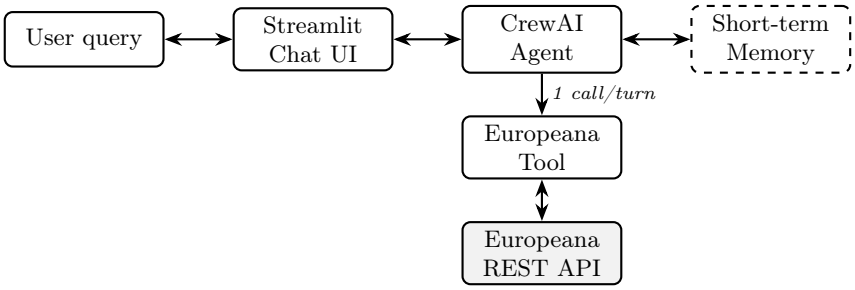


Fig. 1. System architecture overview. A user query flows through the Streamlit chat interface to the CrewAI agent, which is constrained to issue *exactly one* call per conversational turn to the Europeana tool. The tool applies a three-stage cascading retrieval strategy against the Europeana REST API and returns normalised JSON metadata. The short-term memory module (dashed) maintains artist context across turns.

gallery, and an archaeological museum, reporting positive impact on visitor engagement while identifying persistent limitations in contextual grounding and factual precision. Wang et al. [23] propose a conversational AI system for querying digitised natural-history collections containing nearly 1.7 million specimen records, and identify hallucination control and query reformulation as key open challenges. The present work contributes to this landscape by grounding generation exclusively in Europeana-retrieved, citable metadata, directly confronting the hallucination and provenance challenges identified in these prior studies.

3 The Proposed System

3.1 Overall Architecture

The system is structured as a single-agent pipeline whose components interact sequentially to transform a natural-language user question into a response grounded on Europeana metadata. The architecture comprises five principal modules: (i) a *Streamlit* conversational interface [19], (ii) an *intelligent agent* built with the CrewAI framework [5], (iii) an *open-weight GPT-style language model* (gpt-oss-20b) [14] served via the Ollama Cloud hosted service [13], (iv) a dedicated *Europeana tool*, and (v) a *short-term memory* module. Figure 1 illustrates the interaction between these components.

The central design decision distinguishing the proposed system from unconstrained tool-augmented systems is the **single-call constraint**: the agent is instructed, and prompt-engineered, to invoke the Europeana tool exactly once per conversational turn. This yields three concrete benefits: (1) it prevents runaway tool-call loops that arise when underconstrained models attempt iterative refinement; (2) it bounds API latency and usage; and (3) it makes system behaviour predictable and auditable, a crucial property in a provenance-sensitive domain. The `max_iter` parameter of the CrewAI agent is set to 4 as a defensive

cap, the nominal behaviour is one tool invocation followed by the final answer generation

3.2 Agent Design and Prompt Engineering

The CrewAI agent is configured with a role (*art history expert*), an operational goal, and a detailed backstory that collectively orient the language model toward the cultural heritage domain. Role assignment has been shown to reduce off-topic generation in domain-specific settings [24]. The backstory instructs the model to use its general knowledge only for contextual framing (historical period, artistic style, cultural significance), while relying exclusively on Europeana-retrieved data for specific claims such as titles, dates, and provenance URLs.

The task prompt enforces several constraints:

- **Tool invocation discipline.** The agent must call `europeana_search` exactly once; after receiving the tool's JSON response, even if empty, it must proceed immediately to final answer generation.
- **Anti-hallucination rule.** The agent must not invent titles, years, or URLs; all specific data must originate from the tool output.
- **Structured output.** The final answer must be enclosed between the delimiters `<<<ANSWER>>>` and `<<<END>>>`, followed by a mandatory `SOURCES:` block listing Europeana URLs for source attribution.
- **Graceful degradation.** When the tool returns an empty result set, the agent must explicitly acknowledge the data gap and flag the response as based solely on general knowledge.

The model temperature is set to 0.1 to favour stable, near-deterministic outputs. For the reasoning-capable model variant (`gpt-oss`), the `reasoning_effort` parameter is set to `low` to prevent multi-step reasoning from silently bypassing the single-call constraint.

3.3 Europeana Tool

The Europeana tool encapsulates all interaction with the Europeana REST API [7]. Its design addresses three challenges: query formulation, result normalisation, and error resilience.

Query pre-processing. Natural-language questions contain elements, interrogative prefixes, politeness markers, punctuation, that degrade API retrieval performance. A lightweight regex-based pre-processor strips these elements before issuing the API call, transforming inputs such as *"Who is Caravaggio?"* into the bare artist name *"Caravaggio"*, which aligns with the `who:` filter supported by the Europeana search endpoint. The regex covers common Italian and English interrogative patterns, ensuring robustness to the multilingual queries typical of a broad audience.

Cascading retrieval strategy. Europeana's metadata is heterogeneous: some records lack creator attribution; others are catalogued under variant name spellings. To maximise recall without sacrificing precision, the tool implements a three-stage cascade:

1. `who` filter + `TYPE:IMAGE`: retrieves visual works explicitly attributed to the artist via the creator field;
2. `who` filter only: relaxes the media-type constraint to capture non-image records (manuscripts, catalogues);
3. Free-text query + `TYPE:IMAGE`: last-resort fallback for artists not catalogued under the `who` field.

The cascade halts as soon as at least two validated results are returned. A creator-matching heuristic additionally filters records in which the queried artist appears as a *subject* rather than an *author*, a frequent source of spurious results in aggregated cultural databases.

Result normalisation and error handling. Raw Europeana records are normalised into a uniform JSON structure comprising four fields: `title`, `creator`, `year`, and `url`. Results are sorted deterministically (ascending year, then alphabetical title) and at most five records are passed to the language model. In-memory caching avoids redundant API calls within the process lifetime. HTTP 5xx errors trigger an exponential-backoff retry (up to three attempts). If all attempts fail, the tool returns a structured `error` field, enabling graceful degradation rather than a runtime exception.

3.4 Short-Term Conversational Memory

Multi-turn interactions require the agent to resolve co-references such as *"When did he die?"* following an earlier artist query. The system implements a minimal memory module retaining a summary of the *most recent* user question and assistant response (truncated to 400 characters, cut at sentence boundaries). The artist identifier maintained alongside this memory is injected into the current task prompt as the *"current artist in thread"* field, enabling co-reference resolution for follow-up questions.

The deliberate limitation to a single prior turn avoids context accumulation effects. When the user switches to a different artist, the memory is reset, preventing information from prior exchanges from contaminating the Europeana query. This *controlled isolation* strategy, confirmed in multi-turn tests, successfully handles artist-switching scenarios while maintaining coherence within a topic.

3.5 User Interface and Provenance Transparency

The Streamlit-based interface presents a chat-style dialogue history. A key design choice is an optional expandable *execution detail panel* accompanying each assistant response (Figs. 2 and 3), which exposes three inspection tabs: (i) the original user query, the query submitted to Europeana, and the current artist in thread; (ii) the raw JSON returned by Europeana; and (iii) the unprocessed CrewAI output before marker extraction. By making the full retrieval chain visible on demand, the system promotes an epistemological shift from passive consumption to active verification, an important property for educational and scholarly use of cultural archives.

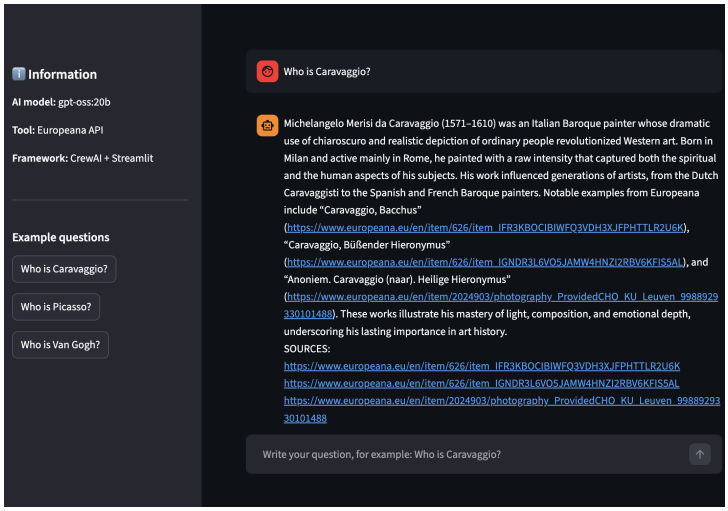


Fig. 2. Art Expert chat interface. A Caravaggio query produces a response enriched with three verified Europeana URLs, citing works retrieved in a single tool call. The sidebar shows the active model, tool, and current artist context.

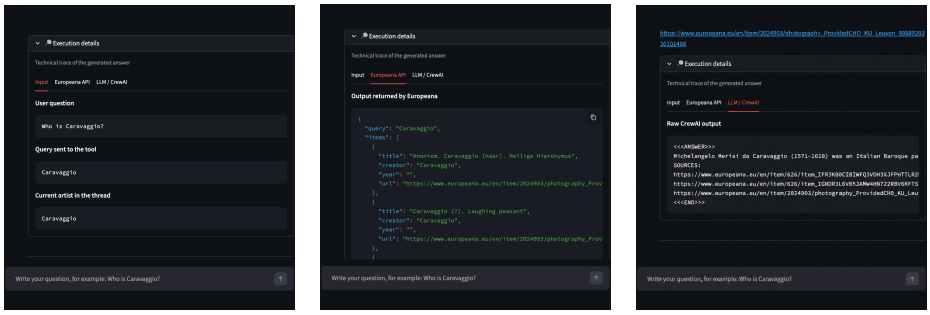
4 Experimental Evaluation

4.1 Experimental Setup

To the best of the authors' knowledge, no publicly available evaluation benchmark exists that is specifically designed for assessing conversational AI agents querying cultural heritage APIs. Comparable systems such as Wang et al. [23] uses a different type of dataset and cultural-heritage-related datasets available on platforms such as HuggingFace address fundamentally different tasks, primarily visual object detection in historical paintings. The test suite adopted in this work is therefore self-constructed.

We designed a systematic evaluation benchmark comprising 150 queries applied to 50 representative European artists spanning multiple periods (Renaissance through Modernism), nationalities, and Europeana coverage densities. For each artist, three question templates were applied: (Q1) a biographical query (*"Who is {artist}?"*); (Q2) an artwork enumeration query (*"Name some works by {artist}."*); and (Q3) a temporal query (*"When did {artist} live?"*). Each query was processed under two conditions: (i) the proposed tool-augmented agent and (ii) an isolated LLM baseline (same model, identical prompt structure, tool access disabled). Short-term memory was simulated identically in both conditions using the same session-state logic as the production application, ensuring full architectural equivalence.

A second benchmark evaluates the short-term conversational memory module through 42 multi-turn sessions comprising three scenario types, applied to a 14-artist subset drawn from the main evaluation pool. *Scenario A* (co-reference



(a) Input tab (b) Europeana API tab (c) LLM / CrewAI tab

Fig. 3. Execution detail panel (expandable, shown open). Each response exposes three inspection tabs: the user question, Europeana query, and current artist (a); the raw JSON payload returned by the Europeana REST API (b); and the unprocessed CrewAI output before answer-marker extraction (c).

chain, 3 turns) presents a biographical query followed by two implicit follow-up questions that omit the artist name, *"When did this artist live?"* and *"Name three famous works by this artist."*, requiring the memory module to maintain artist context without re-identification. *Scenario B* (artist switch, 2 turns) introduces a different artist mid-session via a new biographical query and verifies that the memory reset mechanism correctly isolates the new context. *Scenario C* (implicit follow-up, 2 turns) asks the agent to enumerate works and subsequently to characterise the earliest example, requiring the follow-up answer to be grounded in the preceding response. No isolated-LLM baseline is reported for the multi-turn protocol: the baseline operates without a memory module and is therefore not architecturally comparable on context-dependent tasks.

4.2 Evaluation Metrics

Because no established evaluation protocol exists for tool-augmented cultural heritage agents, the metrics were derived from the core properties the system is designed to guarantee: compliance with the *single-call constraint*, *provenance transparency* through cited sources, and *temporal enrichment* beyond what the base model provides from training data alone. All metrics are binary and automatable; human-judgment metrics (fluency, relevance) were deliberately excluded as they evaluate generic language quality rather than the retrieval properties that distinguish the proposed system from its baseline.

Four automated metrics are computed per query:

– **Tool Discipline** (*tool_called*): whether the agent invoked the Europeana tool exactly once, confirming compliance with the single-call constraint. The metric is conceptually related to tool-call success rate in tool-augmented LLM frameworks [16], though those target correctness of tool selection rather than enforcement of a calling constraint.

- **Source Adherence** (*source_ adherence*): whether the response includes at least one Europeana URL, operationalising verifiability of factual claims. This is closest to the attribution recall dimension of the ALCE benchmark [8] and the faithfulness metric in RAGAS [6]; the present operationalisation is simplified to a binary URL-presence signal.
- **Temporal Coverage** (*year_ present*): whether the response contains at least one four-digit year, indicating provision of temporal context; no established benchmark covers this property in the cultural heritage domain.
- **Temporal Grounding Violations** (*hallucination_ flag*): whether the response contains a year outside a ±2-year tolerance window of every year in the retrieved Europeana records. The tolerance accommodates catalogue dating conventions. Ground-truth years derive from the EDM `dc:date` field, a semantically polymorphic field that may encode creation, digitisation, or acquisition date, so this metric measures *temporal grounding to artifact records* rather than biographical correctness; this distinction is discussed further in Section 4.4.

The multi-turn benchmark requires two additional memory-specific metrics. Since the memory module's correctness is most directly observable in what the agent *queries*, not what it *answers*, both metrics are computed by inspecting the Europeana query string, which is unaffected by paraphrase in the surface response:

- **Co-reference Resolution Rate (CRR)** (*artist_ resolved*): whether the query issued in an implicit follow-up turn (Scenario A, turns 2–3; Scenario C, turn 2) contains the correct artist name, verifying that the memory module propagated the referent established in the priming turn. The task is related to coreference resolution evaluation [15], though standard metrics operate on surface text annotations rather than on agent query behaviour, limiting direct comparability.
- **Switch Isolation Rate (SIR)** (*switch_ isolated*): whether the query in the artist-switching turn (Scenario B, turn 2) contains the new artist's name *and not* the previous one, verifying that the memory reset mechanism discarded the prior context without contamination. No standard metric exists for this property; the measure is informed by dialogue state tracking evaluation [4], where entity switching is a known challenge, but differs in evaluation target and method.

Tool Discipline and Source Adherence are additionally computed per turn across all 98 multi-turn turns to confirm that memory context does not interfere with the tool-calling constraints enforced by prompt engineering.

4.3 Evaluation Results

Table 1 reports aggregate metrics over all 150 single-turn queries; Table 2 presents the multi-turn memory evaluation over 42 sessions. Table 3 provides

Table 1. Automated evaluation over 150 queries (50 artists × 3 question types). †Source Adherence is structurally zero for the baseline, which has no tool access.

Metric	Proposed Agent	LLM Baseline
Tool Discipline (↑)	**100.0%**	n/a
Source Adherence† (↑)	**98.7%**	0%
Temporal Coverage (↑)	**100.0%**	87.3%
Temporal Grounding Violations (↓)	78.0%	**69.3%**

six illustrative cases that exemplify the key interaction patterns identified in the quantitative results.

The agent achieved perfect tool discipline (100.0%), confirming that prompt engineering combined with CrewAI's framework-level constraints successfully enforced the single-call boundary across all 150 queries. Source adherence reached 98.7% (148/150): in two queries the agent produced a valid response without emitting a Europeana URL, consistent with the graceful-degradation path triggered when the tool returns an empty result set. Temporal coverage was 100.0% for the agent compared with 87.3% (131/150) for the baseline, indicating that Europeana retrieval systematically enriches responses with dated artifact records even when the user's question does not explicitly request a date.

The temporal grounding violations metric requires careful interpretation. The agent exhibits a higher violation rate (78.0%) than the baseline (69.3%), a result that is counterintuitive if the metric is read as a direct hallucination measure but consistent with the artifact described in Section 4.2: the agent cites more dates, combining general biographical knowledge with Europeana-retrieved artifact creation years, thereby multiplying the opportunities for a date to fall outside the narrow ±2-year artifact window. The baseline, constrained to its training distribution, produces fewer and less specific date references, yielding fewer metric triggers. Broken down by question type, baseline violations are notably lower for artwork enumeration queries (52.0%, 26/50), where the model tends to describe works without precise creation dates; for biographical (78.0%) and temporal (78.0%) question types, the two conditions are equivalent, reflecting the tendency of both systems to emit specific years when the question explicitly concerns time. We therefore treat the temporal grounding metric as a *precision indicator on artifact date matching* rather than a measure of factual accuracy.

Table 2 reports results for the multi-turn benchmark. Co-reference resolution was correct in all Scenario A implicit turns where the memory module had been primed by a prior biographical query. Artist-switch isolation succeeded in all Scenario B sessions: the memory reset mechanism correctly discarded the prior artist context when a new biographical query was detected, and the Europeana query for the second turn referenced the new artist exclusively. Per-turn tool discipline and source adherence remained consistent with the single-turn results,

Table 2. Multi-turn memory evaluation over 42 sessions (14 artists × 3 scenario types). CRR = Co-reference Resolution Rate (Scenario A, turns 2–3; Scenario C, turn 2); SIR = Switch Isolation Rate (Scenario B, turn 2).

Metric	Proposed Agent
Co-reference Resolution Rate , CRR (↑)	**100.0%**
Switch Isolation Rate , SIR (↑)	**100.0%**
Per-turn Tool Discipline (↑)	**100.0%**
Per-turn Source Adherence (↑)	**100.0%**

confirming that multi-turn context does not disrupt the tool-calling behaviour enforced by prompt engineering.

Table 3 presents a side-by-side comparison of six illustrative queries selected to exemplify the key interaction patterns identified above.

Among these six illustrative cases the agent produced a fully correct response in five instances (83.3%), with the single partial result arising from the open-ended artist-selection query discussed below. In all six cases the tool-call constraint was respected: the agent issued exactly one Europeana query per turn.

4.4 Analysis

Hallucination correction. The most significant result involves the *Fourth Estate* query. The isolated model produced a double hallucination, wrong author (Boccioni instead of Pellizza da Volpedo) and wrong decade (1910–15 instead of 1901), likely because statistical associations between early-20th-century Italian social painting and Futurism bias the model toward Umberto Boccioni. The proposed agent anchors the response to the Europeana metadata record, which unambiguously attributes the work to Pellizza da Volpedo and dates it to 1901. This case illustrates the system's role as a *reality anchor*: when internal model associations diverge from historical fact, the external source forces the correct grounding.

Metadata fidelity vs. historical knowledge. The *Gioconda* case reveals a structurally distinct class of error that differs from hallucination proper and is therefore important to characterise precisely. The isolated model correctly reports the historical creation date (1503–1506), a fact reliably encoded in its training corpus. The proposed agent, by contrast, reports 1977, the year associated with the specific digital record retrieved from Europeana, rather than the date of the painting itself.

This behaviour is architecturally correct: the agent faithfully propagates what the metadata contains. The root cause lies in the semantic flexibility of Europeana's `year` response field, which aggregates the EDM `dc:date` property. In the EDM schema, `dc:date` is intentionally broad and may denote the original creation date of the object, the date of a digitisation event, the date of a catalogue publication, or the date of an acquisition [10]. For the specific *Gioconda*

Table 3. Comparative evaluation: isolated LLM (gpt-oss) vs. the proposed agent. Each row reports the query, the characteristic response of each system, and the evaluation verdict.

Query	LLM (isolated)	Proposed Agent	Verdict
Date of Monet's *Water Lilies Pond*	Reports 1919; no citation	Reports 1907 from Israel Museum record; URL provided	Correct
Details and link for *La Gioconda*	Dates 1503–1506; rich context; no digital link	Reports 1977 (year of the digital record); Europeana URL provided	Metadata faithful
School of Athens (Raphael)	Dates 1510–1511; broad iconographic analysis	Dates 1509–1510; archival detail on specific figures	Correct
Date of *The Night Watch* (Rembrandt)	Reports 1642; no source	Reports 1642; Dutch title identified; URL provided	Correct
Date and author of *The Fourth Estate* (Pellizza da Volpedo)	**Hallucination:** attributes to Boccioni, dates 1910–15	Reports 1901; correct attribution via certified metadata	Correct
Name a Renaissance artist	Selects Leonardo; biography from internal memory	Selects Piero della Francesca; certified works from Europeana	Partial

record retrieved, the field reflects the digitisation catalogue entry (1977) rather than the painting's creation (c. 1503–1506).

This case should be classified not as a *hallucination* but as a *metadata ambiguity error*: no information has been fabricated; the agent has accurately transcribed a semantically ambiguous field. The distinction matters because the two failure modes require different remedies. Hallucination is mitigated by grounding generation in retrieved facts; metadata ambiguity requires an additional layer of schema-level interpretation applied *after* retrieval.

Two complementary mitigations are envisaged: at the field level, preferring `dcterms:created` over `dc:date` and flagging implausibly recent `year` values as probable digitisation dates; at the agent level, equipping the system with a *schema knowledge base* encoding the semantic conventions of each EDM property, so that the agent can reason explicitly about whether a retrieved value represents a creation date, a digitisation event, or a catalogue entry before presenting it to the user.

Finally, the Renaissance-artist query confirms that query precision and response quality are tightly coupled: the agent selected Piero della Francesca correctly, but the generic query yielded sparser Europeana records than artist-directed cases. Multi-turn evaluation confirmed that the memory strategy successfully resolves co-references (e.g., *"When did he die?"* after a Caravaggio query) without cross-context contamination; the artist-reset mechanism triggered correctly in all switching scenarios.

5 Discussion

The results demonstrate that the tool-augmented generation paradigm, when applied to a curated domain-specific API, substantially reduces hallucination without sacrificing the conversational fluency that makes LLM-based assistants compelling for general audiences. From a Digital Humanities perspective, the proposed system embodies what may be termed *cognitive accessibility*: the capacity of an intelligent system to absorb the technical complexity of archival query interfaces and return content that respects the conceptual vocabulary of non-specialist users.

The transparency mechanism shifts the user's relationship with the system from passive consumption to active verification. Unlike chatbots that operate as opaque black boxes, the proposed system provides a verifiable audit trail from user question to archival source, positioning AI not as a replacement for archival authority but as a transparent mediator between the complexity of structured digital archives and the information needs of a broad audience, a model aligned with the responsible-AI principles outlined by UNESCO [21].

Several limitations must be acknowledged. Response quality is bounded by the completeness of Europeana's metadata; moreover, the EDM `dc:date` property is semantically polymorphic , the same field may encode creation date, digitisation date, catalogue date, or acquisition date depending on the contributing institution's cataloguing practice, giving rise to *metadata ambiguity errors* that are architecturally distinct from hallucinations and call for schema-level disambiguation rather than additional retrieval grounding (see Section 4.4). The single-call constraint precludes comparative queries (e.g., *Caravaggio vs. Rembrandt*) that would require sequential or parallel API calls. Additionally, the cascading retrieval strategy occasionally surfaces *school-of* attributions that the model may over-generalise when constructing a narrative response.

6 Conclusions and Future Work

We have presented a tool-augmented AI agent for reliable, grounded natural-language access to the Europeana cultural heritage platform. The system combines the CrewAI orchestration framework with a GPT-family language model, constraining generation to a single Europeana API call per conversational turn. Empirical evaluation on a 150-query benchmark (50 artists $\times$ 3 question types) demonstrates perfect tool discipline (100.0%), near-universal source adherence (98.7%), and complete temporal coverage (100.0%) compared with 87.3% for the isolated LLM baseline. Short-term conversational memory enables coherent multi-turn dialogues without cross-topic contamination, and a transparent UI promotes user trust through provenance disclosure.

Future work will pursue several directions. Decomposing the system into specialised *multi-agent crews* for retrieval, contextualisation, and fact-checking would enable richer, multi-source responses. Integrating vision-language models would support *multi-modal queries*, allowing artwork images as input. *Source*

fusion with complementary knowledge bases such as Wikidata and DBpedia would improve cross-source validation and coverage for under-represented artists. *Metadata disambiguation* strategies could resolve the semantic polymorphism of EDM date fields identified in this work. Finally, pre-indexing Europeana records in a local vector store [12] would reduce API latency and broaden retrieval coverage.

Acknowledgments. The authors acknowledge the Europeana Foundation for providing open access to its REST API and metadata corpus.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agliata, A., Pilato, A., Mariacarmen, S., Bottiglieri, S., Di Nardo, E., Ciaramella, A.: Generative ai and emotional health: Innovations with haystack. In: 2024 IEEE Symposium on Computers and Communications (ISCC). pp. 1–4. IEEE (2024)
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). pp. 610–623. ACM, New York (2021). https://doi.org/10.1145/3442188.3445922
3. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
4. Budzianowski, P., Wen, T.H., Tseng, B.H., Casanueva, I., Ultes, S., Ramadan, O., Gasic, M.: Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 5016–5026 (2018)
5. CrewAI: CrewAI framework documentation. `https://docs.crewai.com` (2024), last accessed 2025/05/01
6. Es, S., James, J., Anke, L.E., Schockaert, S.: Ragas: Automated evaluation of retrieval augmented generation. In: Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations. pp. 150–158 (2024)
7. Europeana Foundation: Europeana REST API documentation. `https://pro.europeana.eu/page/apis` (2024), last accessed 2025/05/01
8. Gao, T., Yen, H., Yu, J., Chen, D.: Enabling large language models to generate text with citations. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6465–6488 (2023)
9. Hyvönen, E.: Publishing and using cultural heritage linked data on the semantic web, vol. 3. Morgan & Claypool Publishers (2012)
10. Isaac, A., Haslhofer, B.: Europeana linked open data — data.europeana.eu. Semantic Web **4**(3), 291–297 (2013). https://doi.org/10.3233/SW-120092
11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM computing surveys **55**(12), 1–38 (2023)

12. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems **33**, 9459–9474 (2020)
13. Ollama: Ollama — local LLM deployment. `https://ollama.com` (2024), last accessed 2025/05/01
14. OpenAI: gpt-oss-120b & gpt-oss-20b model card. arXiv preprint arXiv:2508.10925 (2025), `https://arxiv.org/abs/2508.10925`
15. Pradhan, S., Moschitti, A., Xue, N., Uryupina, O., Zhang, Y.: Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In: Joint conference on EMNLP and CoNLL-shared task. pp. 1–40 (2012)
16. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., et al.: Toolllm: Facilitating large language models to master 16000+ real-world apis. In: International Conference on Learning Representations. vol. 2024, pp. 9695–9717 (2024)
17. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., Scialom, T.: Toolformer: Language models can teach themselves to use tools. Advances in neural information processing systems **36**, 68539–68551 (2023)
18. Sporleder, C.: Natural language processing for cultural heritage domains. Language and Linguistics Compass **4**(9), 750–768 (2010)
19. Streamlit: Streamlit documentation. `https://docs.streamlit.io` (2024), last accessed 2025/05/01
20. Trichopoulos, G., Ordoumpozanis, K., Caridakis, G.: An evaluation of llm-based chatbots for enhancing the visitor's user experience at cultural exhibits. ACM Journal on Computing and Cultural Heritage **19**(1), 1–25 (2026)
21. UNESCO: Recommendation on the ethics of artificial intelligence. Tech. rep., UNESCO, Paris (2021), `https://unesdoc.unesco.org/ark:/48223/pf0000381137`
22. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., et al.: A survey on large language model based autonomous agents. Frontiers of Computer Science **18**(6), 186345 (2024)
23. Wang, Y., Johnston, A., Sadokierski, Z., Stephens, R., Ahyong, S.T.: Conversational ai-enhanced exploration system to query large-scale digitised collections of natural history museums. arXiv preprint arXiv:2603.10285 (2026)
24. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.: A prompt pattern catalog to enhance prompt engineering with chatgpt. arXiv preprint arXiv:2302.11382 (2023)
25. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., et al.: The rise and potential of large language model based agents: A survey. Science China Information Sciences **68**(2), 121101 (2025)
26. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., Cao, Y.: React: Synergizing reasoning and acting in language models. arXiv preprint arXiv:2210.03629 (2022)