IADIS International Journal on WWW/Internet
Vol. 23, No. 2, pp. 81-92
ISSN: 1645-7641

# FROM RECOGNITION TO RELIABILITY: A FRAMEWORK FOR TRUSTWORTHY MULTIMODAL AI IN CULTURAL HERITAGE (THE MUSEEPAL CASE)*

Ahmed M. H. Abdelfattah[1] & David George Atef[2]
[1]*Math. Dept., Faculty of Science, Ain Shams University, Egypt*
[2]*Faculty of Media Engineering & Technology, GUC – German University in Cairo, Egypt*

**ABSTRACT**

The integration of Artificial Intelligence (AI) into cultural heritage offers transformative potential for visitor engagement but faces a critical challenge: the "hallucination" of facts by Generative models in high-stakes domains. This paper presents `MuseePal`, an AI-powered tourist assistant designed to mitigate these risks through a formalized, constraint-based architecture. Unlike generic chatbots, `MuseePal` operates on a rigorous Grounding-Aware framework that synthesizes real-time vision with a strict retrieval-augmented generation (RAG) pipeline. We introduce a theoretical formalism to define a generalized Cultural Knowledge Mapping function, ensuring that generated outputs are structurally bound to a curated ontology and verifiably attributed to source documents. While our proof-of-concept implementation utilizes the YOLOv12 architecture and Gemini models, the proposed framework is model-agnostic, adaptable to various perception and reasoning engines. Evaluation within the Grand Egyptian Museum context demonstrates how these formal constraints bridge the gap between generative AI capability and reliability.

**KEYWORDS**

Trustworthy AI (TAI), Cultural Heritage, Multimodal RAG, Formal Knowledge Representation

*This work is an extended and significantly enhanced version of a preliminary study originally presented at the 24th International Conference on WWW/Internet (ICWI 2025), titled "MUSEEPAL: AN AI-POWERED CULTURAL GUIDE FOR MUSEUMS (THE CASE OF THE GRAND EGYPTIAN MUSEUM)" by the same authors (Abdelfattah & Atef, 2025).

# 1. INTRODUCTION

Museums are among the most visited cultural venues worldwide, offering a valuable opportunity to engage with history, art, and culture. They serve as the custodians of human history, yet the transmission of this knowledge to visitors remains bottlenecked by static plaques and the limited availability of human guides. Self-guided visits often feel limited, as most artifacts are accompanied only by brief, generic plaques that lack depth and personalization (Falk & Dierking, 2018). This challenge is magnified in vast institutions like the Grand Egyptian Museum (GEM), which houses alone over 100,000 artifacts from one civilization, making it impossible for visitors to receive detailed information on every piece. Guided tours offer more context but can be costly, rigid, or uncomfortable for visitors with language barriers or social anxiety. They are also prone to information overload and what is known as “museum fatigue,” a phenomenon where visitors become mentally and physically drained due to the volume and density of content, resulting in poor focus and minimal retention (Bown, 2020; Davey, 2005). Additionally, tour quality can vary significantly depending on the guide, with studies showing that inconsistent delivery and engagement often lead to underwhelming visitor experiences (Andelkovic et al., 2022). Guided tours may also overlook certain artifacts that individual visitors find meaningful.

While recent advancements in Large Language Models (LLMs) promise to democratize access to cultural knowledge by enabling the generation of real-time, contextual explanations, they introduce a significant "Trust Gap." Generative models, optimized for plausibility rather than truth, frequently "hallucinate" details when queried about specific, lesser-known artifacts due to their generic nature and lack of specialization in museum-specific content (Béchard & Ayala, 2024; Ji et al., 2023). This creates a gap in delivering accurate, accessible, and personalized museum experiences. In the domain of Cultural Heritage, where historical accuracy is paramount, such unreliability is unacceptable.

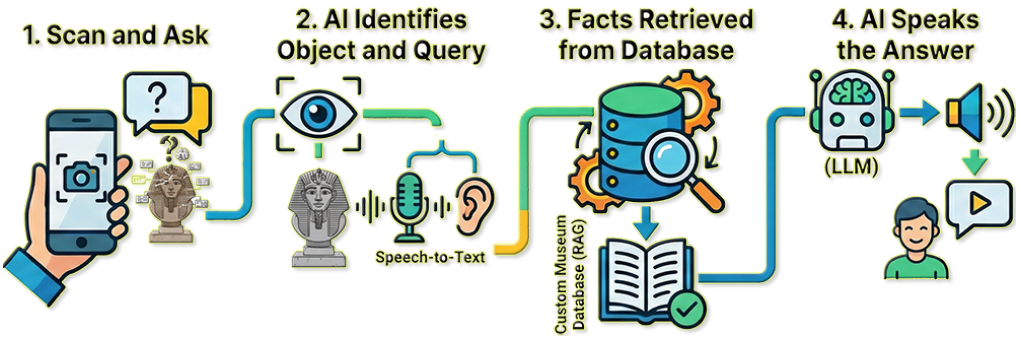


Figure 1. A conceptual illustration of the core design and functionality of our system architecture

This paper addresses this core challenge by proposing `MuseePal`, a multimodal system that acts not merely as a tour guide, but as a Grounding-Aware Retrieval System. Rather than replacing LLMs, we enhance them to deliver reliable responses tailored to underrepresented content (Figure 1). In fact, our contribution is twofold: (1) we establish a formal theoretical framework for Trustworthy AI (TAI) in cultural contexts, and (2) we substantiate this theory through a rigorous experimental proof-of-concept, detailing the engineering of `MuseePal`. Consequently, the remainder of this paper is structured as follows: After related work is

summarized in the next section, Section 3 establishes the generalized mathematical formalism for schema-guided retrieval; Section 4 validates this theory through the specific `MuseePal` implementation at the GEM; and Section 5 discusses future extensions.

We posit that reliability in Generative AI cannot be achieved through model scaling alone but requires an architectural approach, perhaps a Neuro-Symbolic one, that involves:

- **Neural Perception:** Utilizing computer vision to act as a deterministic trigger for information retrieval.
- **Symbolic Grounding:** Constraining the generative output using a formal "Cultural Schema" –an ontological structure defined in recent work on Smart Tourism (Zakzouk & Abdelfattah, 2025)—to ensure comprehensive coverage of historical attributes.
- **Verifiable Attribution:** Implementing a strict citation mechanism, as proposed in the "Studymate" framework (Gebrail & Abdelfattah, 2025), to enforce source tracking.

By formalizing the interaction between the physical artifact, the digital knowledge base, and the generative model, `MuseePal` transforms the museum visit from a passive observation into a rigorous, interactive, and trustworthy educational dialogue.

# 2. RELATED WORK

## 2.1 AI in Cultural Heritage

Early integrations of AI in museums focused on Visual Question Answering (vQA) (Bongini et al., 2020) and virtual tours. The *VISCOUNTH* dataset (Becattini et al., 2023) expanded this direction, introducing a multilingual dataset of over 500K cultural assets and 6.5M QA pairs to improve robustness and inclusivity in vQA for heritage contexts. This provided the scale needed to train robust models but often emphasize breadth across cultural assets rather than deep, site-specific coverage. Research has also begun exploring AI systems that function as virtual human guides in real-world cultural tourism: Zhou and Zhang (2025) discuss the integration of digital twins and virtual humans in a tourism metaverse, emphasizing multimodal interaction and intelligent dialogue systems for cultural product development. While effective in simulated environments, these systems often lacked real-time adaptability to physical artifacts and didn't ensure precision in real-world deployment. They remain limited either to simulated environments or generic datasets. More recent works like *VirtuWander* (Wang et al., 2024) have explored multimodal interaction, and highlighted the importance of personalization and conversational adaptability in heritage contexts. Xie et al. (2025) proposed *MetaDecorator*, a framework that augments 360° panoramas with multimodal content, text prompts, and synthesized imagery to generate immersive virtual tours. However, these works often rely on open-ended LLMs, leaving the "hallucination problem" unaddressed.

`MuseePal` contributes a novel synthesis: combining real-time artifact detection, **Retrieval-Augmented Generation** (RAG) grounded in museum content, and a voice-to-voice pipeline. It shifts from static image-text matching toward real-time detection, and advances the abovementioned related work by shifting from open-ended generation to RAG (Karpukhin et al., 2020; Lewis et al., 2020), grounding responses in a curated, museum-specific database. This integration addresses gaps in factual reliability, personalization, and accessibility, offering a scalable model for cultural AI assistants in physical museum contexts.

## 2.2 Agentic AI and Cultural Schemas

To prevent AI from generating vague or generic descriptions, outputs must be structurally constrained. Zakzouk & Abdelfattah (2025) proposed a "**Cultural Schema**" for Smart Tourism (Zakzouk & Abdelfattah, 2025)—a structured ontology defining specific attributes (e.g., "Historical Significance," "Religious Beliefs," "Artistic Style") that an AI agent must populate. `MuseePal` integrates this schema into its retrieval logic, ensuring that the visual detection of an artifact triggers a structured, multi-dimensional narrative rather than a generic summary.

## 2.3 Trustworthy AI and Source Attribution

In high-stakes domains such as education and history, the veracity of AI output is critical. Gebrail & Abdelfattah (2025) introduced the *Studymate* framework (Gebrail & Abdelfattah, 2025), which established that "Source Attribution" is a necessary condition for trust in educational AI. Their work demonstrated that constraining an LLM to a closed set of instructor-curated materials, with mandatory citation, significantly reduces fabrication. `MuseePal` adapts this paradigm to the cultural sector, treating museum artifacts as "learning objects" that require the same rigorous attribution standards.

While the works of Zakzouk & Abdelfattah (2025) and Gebrail & Abdelfattah (2025) provide the necessary components for reliability (schemas and attribution, respectively), a unified formalization for combining these constraints within a multimodal environment is currently absent. Existing literature largely treats the engineering of museum AI as an architectural challenge rather than a formal knowledge retrieval problem. To bridge this gap, we must move beyond ad-hoc system descriptions. Before detailing the specific implementation of `MuseePal`, it is necessary to first establish a generalized theoretical framework. This framework defines the logical conditions required for “Trustworthiness” in a cultural setting, treating the system not merely as software, but as a deterministic mapping between physical artifacts, curated knowledge, and generative logic.

# 3. THEORETICAL FRAMEWORK: THE SCHEMA-GUIDED GROUNDING MODEL

To rigorously define the operation of `MuseePal`, we first move beyond software architecture to formalize the problem of *Trustworthy Multimodal Retrieval*, defining the necessary logical conditions for a reliable system. In this section, we abstract away the specific tools (such as YOLO or Gemini) used in our proof-of-concept to define a model-agnostic formalism for Trustworthy Multimodal AI. By defining the interaction between perception and retrieval using set theory and logic, we establish a blueprint that ensures reliability is a mathematical constraint of the system design, rather than a byproduct of model scale. This formalization is outlined in Figure 2, explained below, and constitutes the core scientific contribution of this work, of which the `MuseePal` application is a specific instantiation (as detailed in Section 4).

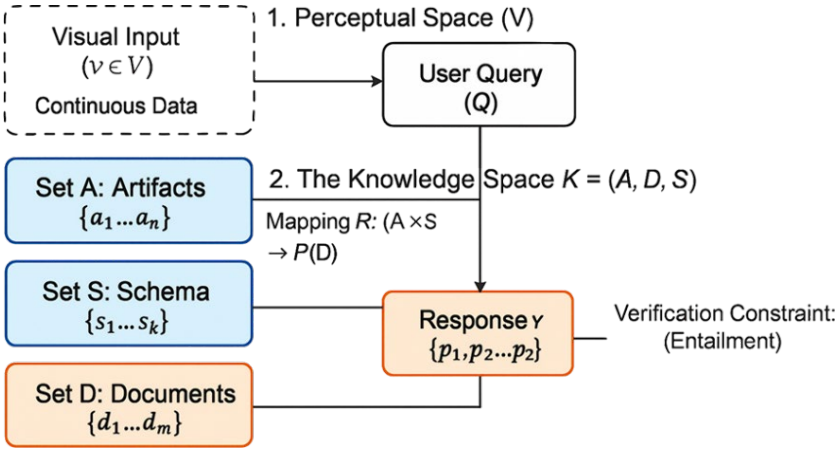


Figure 2. A high-level diagram showing the theoretical framework structure, consisting of the knowledge space and the abstract mapping functions, which are explained in section 3.1

## 3.1 The Knowledge Space

We begin by defining the domain in which the AI operates. Unlike open-domain chatbots, our system is constrained to a specific museum environment. Let the museum environment be defined as a **Knowledge Space** tuple, $\mathcal{K}=(\mathcal{A}, \mathcal{D}, S)$, where**:**

- $\mathcal{A} = \{a_1, \ldots, a_n\}$ is the finite set of distinct physical **artifacts** housed in the museum (e.g., the Statue of Ramesses II).
- $\mathcal{D} = \{d_1, \ldots, d_m\}$ is the set of curated **document chunks** stored in the knowledge base. These chunks serve as the atomic units of truth (the ground truth in a vector space).
- $S = \{s_1, \ldots, s_k\}$ is the **cultural schema** (ontology), where each $s_i$ represents a specific semantic dimension of inquiry (e.g., *Dynastic Period*, *Material Composition*, *Spiritual Function*).

**The Neural Perception Mapping:** The first stage of the pipeline is **Visual Perception**. The goal of this stage is to convert continuous visual data (pixels) into a discrete symbolic identifier that matches an element in our artifact set $\mathcal{A}$.

- Let $\mathcal{V}$ be the (infinite) space of all possible visual inputs. We define the **perception model** not merely as a classifier, but as a mapping function $\Phi: \mathcal{V} \rightarrow \mathcal{A} \cup \{\emptyset\}$ that maps the visual input space $\mathcal{V}$ (e.g., camera frames) to the artifact set $\mathcal{A}$, where $\Phi(v)=a_i$ if an artifact is detected with confidence $c > \tau$ (threshold), and the empty set otherwise. This function effectively discretizes the continuous visual world into symbolic tokens (artifact IDs). This definition holds regardless of whether the underlying implementation uses YOLO, Faster R-CNN, or future vision transformers, provided the confidence constraint is met. $\Phi$ takes an image $v \in \mathcal{V}$ and maps it to a specific artifact $a_i \in \mathcal{A}$ if and only if the model detects it with a confidence score $c$ exceeding a strict threshold $\tau$. If no artifact meets this threshold, the function maps to the empty set $\emptyset$. This discretizes the noisy visual world into a clean symbolic token usable by the retrieval system.

**The Grounded Retrieval Function:** RAG systems often fail because they retrieve documents based solely on generic similarity to a user query (Wang et al., 2023). Unlike standard RAG, `MuseePal` enforces a structure by defining Schema-Guided Retrieval: a retrieval process that is guided by a Cultural Schema $S$.

- We define an embedding function E that maps text or symbolic tokens into a high-dimensional vector space, E: $(\mathcal{A} \cup S \cup \mathcal{D}) \rightarrow \mathbb{R}^d$. We define **the retrieval function** as a mapping $\mathcal{R}: (\mathcal{A} \times S) \rightarrow \mathcal{P}(\mathcal{D})$ that takes the detected artifact $a_i$ and a specific schema attribute $s_j$ (e.g., "*History*" of "*Ramesses II*") derived from the user's query or the agent's narrative plan, and outputs a subset of (relevant) documents $\mathcal{D}_{\text{rel}} \subseteq \mathcal{D}$ of the document store, where $\mathcal{P}(\mathcal{D})$ is the power set of documents. This means the function $\mathcal{R}$ returns a specific collection of document chunks relevant to the intersection of the artifact and the attribute. Formally, a retrieved subset $\mathcal{D}_{\text{rel}} \subseteq \mathcal{D}$ is defined as $\{d \in D | \text{sim}\left(\text{E}(d), \text{E}(a_i \oplus s_j)\right) \geq \epsilon\}$, where $\oplus$ denotes semantic concatenation (combining the artifact name with the attribute description), and the function sim computes cosine similarity between the document vector E($d$) and the query vector. A retrieved subset $\mathcal{D}_{\text{rel}} = \mathcal{R}(a_i, s_j)$ is considered **valid** if and only if $\forall d \in \mathcal{D}_{\text{rel}}$, sim (E($d$), E($a_i \oplus s_j$)) $\geq \epsilon$. Thus, the vector similarity threshold (condition $\geq \epsilon$) ensures that the retrieval is not just semantically close enough to the query but topologically bound to the intersection of the specific artifact and the specific cultural attribute.

**The Trustworthiness Constraint:** Finally, we formalize the generation step. We define the generative model $\mathcal{G}$ (LLM) not as a creative engine, but as a **transformation function** that takes the user query $Q$ and the retrieved relevant documents $\mathcal{D}_{\text{rel}}$ to generate a textual response (an answer $Y$). To satisfy the principles of Trustworthiness AI (TAI), we impose a verification constraint derived from the framework introduced in Gebrail & Abdelfattah (2025). We treat the output $Y$ as a set of atomic propositions (factual claims) $\{p_1, \ldots, p_z\}$, where the generation is subject to a **verification constraint**: "$Y$ is Trustworthy $\Leftrightarrow \forall$ proposition $p \in Y$, $\exists d \in \mathcal{D}_{\text{rel}}$ such that $d \vDash p$",where $\vDash$ represents logical entailment and the constraint states that for a response to be considered trustworthy, every single factual claim $p$ generated by the AI must be logically supported (modeled, $\vDash$) by a document $d$ found in the retrieved set. In a fully idealized system, this constraint would be validated by an external natural language inference (NLI) module. `MuseePal` approximates this entailment by enforcing a strict system-prompting strategy that penalizes unsupported claims and mandates a citation format ($p$, source($d$)). This shifts the burden of verification from the "black box" neural network weights to the explicit source mapping, allowing for human-in-the-loop validation.

# 4. METHODOLOGY & IMPLEMENTATION

The theoretical framework described above provides the blueprint and allows for various instantiations. We now describe the specific engineering implementation of `MuseePal` used to validate this framework within the Grand Egyptian Mueseum (GEM) and providing, thus, a proof-of-concept that shows the framework's efficacy.

## 4.1 Dataset & Visual Perception (Φ)

To realize the mapping Φ, we instantiate it by selecting the YOLOv12 architecture due to its superior balance of inference speed and accuracy on mobile edge devices.

- **Data Collection:** One of the primary challenges in developing MuseePal is the absence of a publicly available image dataset for the GEM. Unlike many renowned museums with online archives, GEM's collection remains largely undocumented. To overcome this, a custom dataset was created to support artifact detection. We constructed a custom dataset for the GEM during multiple field visits, covering five carefully selected artifact classes that represented both prominent and lesser-known exhibits. The dataset consists of more than 500 original high-resolution photographs (to simulate real-world usage) covering diverse artifact classes (e.g., *Statue of Ramesses II*, *King Amenhotep III*). After filtering out blurry or duplicate images, the dataset was refined to about 300 photos.
- **Augmentation:** To approximate the robustness required by Φ under varying lighting and angles, the dataset was expanded to 1,500 labeled instances using noise injection, rotation, and shearing. Each image was manually annotated with bounding boxes using tools like Roboflow and CVAT, with labels mapped to the corresponding artifact classes. All images were then uniformly cropped and resized to a fixed resolution of 640×640 pixels, matching the model's input dimensions and ensuring consistency with the image format captured and processed by the mobile application's camera module. This standardized preprocessing step guarantees reliable performance across both live camera input and uploaded photos within the app interface.
- **Performance:** The annotated dataset was split into 70% training, 20% testing, and 10% validation sets. This instantiation achieved a mean Average Precision at 0.5 IoU (mAP@0.5) of 99.5%, with a precision of 98.9% and recall of 97.8% (see Table 1 and Table 2 in Section 4.5). We acknowledge that this near-perfect score reflects the closed-world nature of our specific GEM dataset (5 classes). This high precision ensures that the detected artifact $a \in \mathcal{A}$ (which serves as the primary input to the retrieval function $\mathcal{R}$) is identified accurately, thereby minimizing downstream error propagation. While this confirms the feasibility of the pipeline for specific tours, significantly larger and more diverse datasets will be required to generalize Φ for cross-museum deployment without overfitting.

## 4.2 RAG Pipeline and Retrieval ($\mathcal{R}$)

Transformers, the backbone of modern LLMs, are built on three key innovations. First, self-attention allows models to focus on relevant tokens in a sequence (e.g., identifying "it" in "The cat sat on the mat, and it was fluffy" as referring to "the cat"). Second, multi-head attention enables learning of different relationships by running multiple self-attention mechanisms in parallel. Third, the context window defines how many tokens can be processed simultaneously larger windows improve handling of long-range dependencies. Despite advances, LLMs remain limited to their pretraining data and cannot retrieve external facts unless extended with additional mechanisms. Consequently, the next critical step was to enable our system to generate accurate and context-aware historical descriptions for detected artifacts.

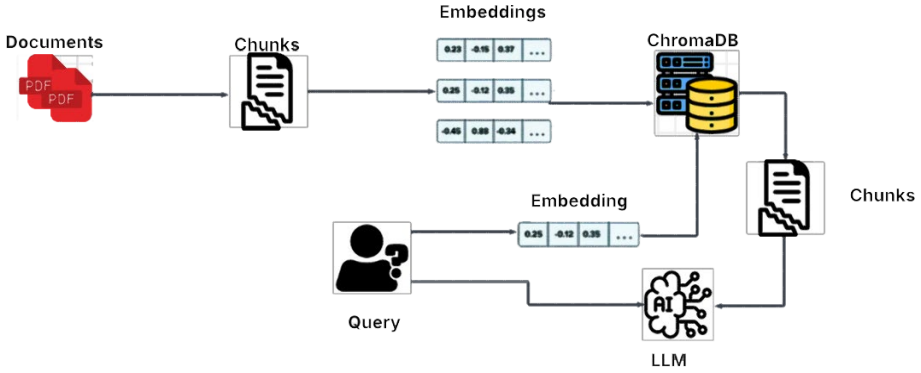


Figure 3. The `MuseePal` RAG pipeline as shown in Abdelfattah & Atef (2025). Documents are chunked, embedded, and stored in ChromaDB. Queries are matched to relevant chunks and passed to the LLM for response generation

To implement the retrieval function $\mathcal{R}$, we utilized ChromaDB–a high-performance vector database optimized for similarity search– as the vector store backend. `MuseePal` integrates a RAG pipeline (Figure 3), which enriches the LLM's output by incorporating structured external knowledge at runtime. In `MuseePal`, this process begins with document chunking, where textual descriptions of museum artifacts (sourced from field notes, scanned exhibit labels, and verified web references) are segmented into smaller, semantically meaningful passages. These chunks are then passed through an embedding model to produce high-dimensional vector representations. Each chunk is embedded into a 1024-dimensional space and stored in a local instance of ChromaDB. This hybrid architecture allows `MuseePal` to produce factually accurate explanations even for topics not covered in the LLM's pretraining distribution. Additionally, the use of a local database confers performance and privacy benefits, eliminating the need for external API calls for every user interaction.

- **Chunking Strategy:** Curated historical texts were segmented into semantic chunks. Crucially, these chunks were tagged according to the Cultural Schema $S$ defined in Section 3.1 (e.g., chunks regarding "*Dynastic History*" were indexed separately from "*Artistic Style*").
- **Vector Space:** The embedding function E was realized using a high-dimensional embedding model (1024 dimensions), converting text into vector representations.
- **Schema Integration:** Following the methodology in Zakzouk & Abdelfattah (2025), queries are not treated as simple keyword matches. The system injects context from the detected artifact class into the query vector, effectively performing the operation $E(a_i \oplus s_j)$.
- **Query-to-Schema Parsing:** To map a free-form user query $Q$ to a specific schema attribute $s_j \in \mathcal{S}$, we utilize an intermediate classification step. The LLM is prompted to classify $Q$ into one of the pre-defined schema slots (e.g., mapping "*How was this made?*" to *Material Composition*). If a query spans multiple slots, the system decomposes it into sub-queries, retrieving relevant chunks for each attribute before synthesis.

## 4.3 Conversational Agent ($\mathcal{G}$)

The backend orchestration was built using FastAPI, managing the data flow between perception, retrieval, and generation.

- **Input Processing:** Rather than relying on typed queries and text responses, users can speak naturally and receive spoken explanations, effectively simulating a conversation with a knowledgeable human guide. User voice input is transcribed via Whisper (an open-source automatic speech recognition (ASR) model developed by OpenAI), converting speech to text $Q$ and supporting multiple languages to offer strong transcription accuracy even in noisy environments, which handles a variety of accents with minimal degradation. These qualities make it particularly well-suited for use in museum settings, where background noise and multilingual visitors are common. Once the user speaks a query, Whisper converts the audio into text. This transcription is then forwarded to the RAG pipeline (Figure 4), where it is embedded and matched against artifact-related content stored in ChromaDB. The most relevant context is retrieved and sent to the reasoning engine (LLM).
- **Reasoning Engine:** Google's Gemini model serves as the generative function $\mathcal{G}$. It receives the retrieved context $\mathcal{D}_{rel}$ and the user query and generates a grounded and informative response, which is converted into speech for playback.
- **Attribution Enforcement:** The system prompt explicitly forbids the use of external knowledge for factual claims. It strictly adheres to the verification constraint ($d \nvDash p$) by requiring the model to append citations to every claim.
- **Output Synthesis:** Responses are synthesized into natural speech using Kokoro, enabling a hands-free, immersive museum experience.

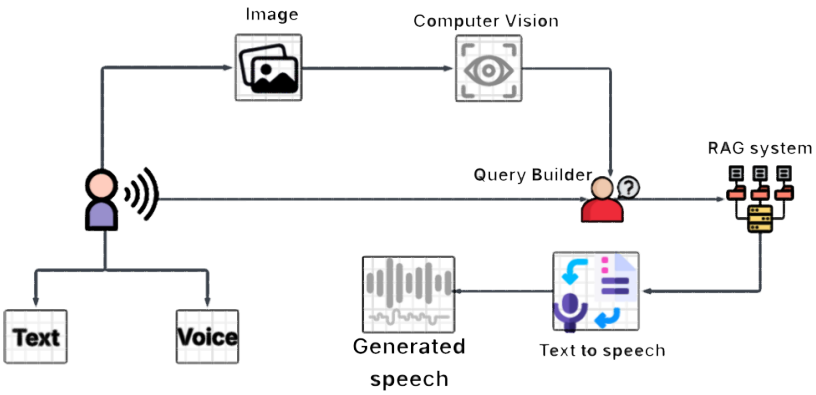


Figure 4. The `MuseePal` speech-to-speech pipeline as shown in Abdelfattah & Atef (2025). Voice or image input is processed via computer vision and RAG, with responses returned through text-to-speech

## 4.4 Further Details & App Integration

During the early development stages, the application was entirely built within a single monolithic Python script. That prototype, while functional, lacked the scalability and modularity

required for checking reliability and for integration with a mobile frontend. The backend was refactored into a lightweight, API-based architecture using Flask. This design decouples the core logic from the user interface, enabling seamless communication. The backend now exposes distinct RESTful API endpoints for handling key functionalities, such as audio transcription, RAG-based querying, and audio generation. This modular design ensures that the system is scalable, easy to test, and deployable across various environments, from local machines to cloud infrastructure. The mobile frontend was developed using React Native with Expo, selected for its rapid prototyping capabilities and cross-platform support (cf. Abdelfattah & Atef, 2025).

## 4.5 Evaluation & Results

MuseePal was evaluated through a combination of quantitative performance metrics and qualitative user experience testing across its image detection, voice processing, and conversational components. It demonstrated strong performance, as shown in Tables 1 and 2. While these results show excellent recognition capability, the near-perfect precision also indicates possible overfitting on the limited dataset, highlighting the need for further validation with larger and more diverse artifact collections. The RAG component significantly improved contextual accuracy compared to LLM-only responses, while the end-to-end voice pipeline enhanced accessibility for users preferring auditory interaction. Overall, evaluation confirmed the feasibility of `MuseePal` as a proof-of-concept system, while also identifying the importance of dataset expansion and user studies for broader deployment.

Table 1. Overall Model Performance Metrics as shown in Abdelfattah & Atef (2025)

| *Metric:* | *Accuracy* | *Macro Precision* | *Macro Recall* | *Macro F1 Score* | *mAP@0.5* | *mAP@0.5:0.95* |
|---|---|---|---|---|---|---|
| Value: | 98.36\% | 1.00 | 0.975 | 0.987 | 0.995 | ~0.84 |

Table 2. Per-Class Evaluation Metrics (Actual Results) as shown in Abdelfattah & Atef (2025)

| *Class* | *TP* | *FP* | *FN* | *Precision* | *Recall* | *F1 Score* |
|---|---|---|---|---|---|---|
| Seated Statue of King Amenhotep III | 9 | 0 | 0 | 1.00 | 1.00 | 1.00 |
| Standing Statue of King Thutmose III | 15 | 0 | 0 | 1.00 | 1.00 | 1.00 |
| Statue Head of King Akhenaten | 12 | 0 | 0 | 1.00 | 1.00 | 1.00 |
| Statue of King Sety II Holding Standards | 17 | 0 | 0 | 1.00 | 1.00 | 1.00 |
| Statue of Ramesses II | 7 | 1 | 0 | 0.875 | 1.00 | 0.933 |
| Background | 0 | 0 | 1 | 0.00 | 0.00 | 0.00 |

# 5. FUTURE WORK

While the current system relies on set theoretic constraints, future iterations will require a probabilistic approach to quantify uncertainty in Cultural AI. We propose the following extensions for future research.

## 5.1 Theoretical Expansion: Probabilistic Hallucination Modeling

We intend to move from binary verification to a Bayesian Confidence Model. Let $H$ be the event of a hallucination. We can model the probability of a truthful response $T$ given the context $C$

and query $Q$ as a Bayesian update: $P(T|Q, C)=\frac{P(Q|T,C)P(T|C)}{P(Q|C)}$, where $P(T|C)$ is the prior probability that the text is true given the context (consistency check). The term $P(Q|T,C)$ represents the likelihood that the query would generate this response if it were true. Future work will involve developing a "Verification Agent" that estimates these probabilities numerically. If the posterior probability $P(T \mid Q, C)$ falls below a safety threshold $\lambda$, the system would be programmed to withhold the answer rather than speculate.

## 5.2 Theoretical Expansion: Graph-Based Retrieval (GraphRAG)

The current vector-based retrieval assumes independence between data chunks. However, history is relational (e.g., *Father-Son* relationships between *Pharaohs*). We propose transitioning the Knowledge Space $\mathcal{K}$ from a set of discrete vectors to a Knowledge Graph $K_G= (V, E)$, where the nodes ($V$) represent artifacts, historical figures, and locations; and the edges ($E$) represent semantic relationships (e.g., "*Built By*", "*Succeeded By*", "*Located In*"). Retrieval would then be defined as a "subgraph extraction" problem. Instead of finding similar points in space, the system would traverse edges to answer complex relational questions like "*How did this king's father influence his art style?*", which vector similarity alone cannot resolve.

## 5.3 Practical Expansion: Cross-Museum Generalization

We plan to test the transferability of the mapping function $\Phi$ and retrieval function $\mathcal{R}$ to different cultural contexts (e.g., Greco-Roman museums). This will validate the universality of the proposed Schema-Guided Grounding Model beyond the specific instance of Ancient Egyptian heritage.

# 6. CONCLUSION

The synthesis of the `MuseePal` architecture with the presented trustworthy principles reveals an effective, generalizable paradigm for TAI in Cultural Heritage. By integrating the "Cultural Schema" ontology (Zakzouk & Abdelfattah, 2025), `MuseePal` avoids the common pitfall of generic AI responses (e.g., "*This is a nice statue*"). The schema forces the retrieval system to seek specific dimensions—such as the *political implications* of a statue's stance or the *religious significance* of its inscriptions. This ensures that the system functions as a domain expert rather than a generalist chatbot. In addition, the adoption of the "Source Attribution" as a trust mechanism (Gebrail & Abdelfattah, 2025) transforms the user experience. When the system states a historical date or interprets a hieroglyph, it implicitly links that claim to a specific document in the vector store. This creates a "Chain of Trust": User →Agent →Document →Museum Curator. This chain is essential for educational deployment, where misinformation can have pedagogical consequences.

# REFERENCES


Abdelfattah, A. & Atef, D. G., 2025, MuseePal: An AI-Powered Cultural Guide For Museums (The Case of The Grand Egyptian Muesuem). *Proceedings of the 24th International Conference on WWW/Internet (ICWI 2025)*. Porto, Portugal, pp. 197-204.

American Alliance of Museums, 2022. *Museum facts & data*. [online]. Available at: https://www.aam-us.org/programs/about-museums/museum-facts-data/

Andelkovic, Z. et al., 2022**.** Museum tour guide performance: A visitor perspective. *Sustainability,* Vol. 14, No. 16.

Becattini, F., Bongini, P., Bulla, L., Bimbo, A. D., Marinucci, L., Mongiovì, M. & Presutti, V., 2023. VISCOUNTH: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, *19*(6), pp. 1-20.

Béchard, P. & Ayala, O. M., 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. *arXiv preprint arXiv:2404.08189*

Bongini, P., Becattini, F., Bagdanov, A. D. & Del Bimbo, A., 2020, November. Visual question answering for cultural heritage. *IOP Conference Series: Materials Science and Engineering*, *949*(1), 012074.

Bown, C., 2020. How much information is too much on a guided tour? *Thinking Museum Blog*. [online] Available at: https://thinkingmuseum.com/how-much-information-is-too-much-on-a-guided-tour

Davey, G., 2005. What is museum fatigue. *Visitor Studies Today*, *8*(3), 17-21.

Falk, J. H. & Dierking, L. D., 2018. *Learning from museums: Visitor experiences and the making of meaning*. Rowman & Littlefield, Lanham, MD.

Gebrail, M. & Abdelfattah, A., 2025. AI-Powered Pedagogy: Generating Reliable and Context-Aware Educational Content from Instructor-Curated Materials. *Proceedings of the 22nd International Conference on Applied Computing (AC 2025)*. Porto, Portugal, pp. 91-98.

Ji, Z. et al., 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38.

Karpukhin, V. et al., 2020. Dense passage retrieval for open-domain question answering*. Proceedings of EMNLP* (1). [Online], pp. 6769-6781.

Lewis, P. et al., 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Wang, B. et al., 2023. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713*

Wang, Z. et al., 2024, May. Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1-20.

Xie, S., Liu, Y., Lee, J. S. & Dong, H., 2025. MetaDecorator: Generating Immersive Virtual Tours through Multimodality. *arXiv preprint arXiv:2501.16164*

Zakzouk, A. & Abdelfattah, A., 2025, Smart Tourism Meets AI: A Multi-modal Assistant for Exploratory Cultural Engagement. *Proceedings of the 22nd International Conference on Applied Computing (AC 2025)*. Porto, Portugal, 260-264.

Zhou, X. & Zhang, M., 2025. Fusion of Virtual Human Guides and Digital Twins in Building a Tourism Metaverse: A Research on Cultural Tourism Product Development Based on Multimodal Interaction and Intelligent Dissemination Paths. *Applied and Computational Engineering*, Vol. 163 (Proceedings of the 3rd International Conference on Software Engineering and Machine Learning), pp. 42-48. doi: 10.54254/2755-2721/2025.24193