Generative AI Model for Archaeology-based Learning

FYP24087

Submitted by: Yian Fan

UID: 3035952886

Abstract

This report explores the development of a Generative AI (Gen AI) system tailored for archaeological research, with a focus on the limitations of existing large language models (LLMs) in accurately synthesizing and generating domain-specific knowledge. By leveraging Retrieval-Augmented Generation (RAG), the system enhances the AI's ability to process, interpret, and synthesize archaeological texts. Through a combination of advanced text chunking strategies, metadata extraction, and the development of a specialized dataset, this project aims to create a tool that supports archaeologists in analyzing and interpreting complex archaeological data with greater accuracy and efficiency. The system's performance is evaluated using a custom benchmark set of 125 questions, yielding an F1-score of 93.44%, demonstrating its effectiveness. The report outlines the methodology, challenges faced, and future improvements to enhance the system's applicability in real-world archaeological research.

Acknowledgements

I want to express my gratitude to Professor Peter J. Cobb for providing me with the chance to work on this project. His assistance during the process has been crucial in determining the course of my investigation. I am appreciative of the opportunity to work on such a fascinating and difficult project under his direction.

I want to sincerely thank Dr. Schnieders Dirk, whose knowledge and careful guidance have been invaluable. I was able to overcome many of the theoretical and technical obstacles I faced thanks to his perceptive suggestion. I am grateful to have had the chance to learn from him.

Table of Contents

List of Figures

List of Tables

List of Abbreviations or Symbols

Abbreviation	***Full Term***
RAG	Retrieval-Augmented Generation
Gen AI	Generative Artificial Intelligence
LLMs	Large Language Models
GANs	Generative Adversarial Networks
NMT	Neural Machine Translation
UI	User Interface
UAVs	Unmanned Aerial Vehicles
AI	Artificial Intelligence
OCR	Optical Character Recognition

1. Introduction

1.1. Background

Archaeology plays a crucial role in understanding human history and cultural development by providing tangible evidence of past societies' lifeways and beliefs [1]. The field of archaeology has traditionally relied on the analysis and interpretation of artifacts and historical records to reconstruct ancient civilizations, with the physical alteration and excavation of archaeological research often described as "destructive" [2]. However, in recent years, the field has witnessed significant shifts towards digitalization and non-invasive methodologies—such as digital photogrammetry, laser scanning, unmanned aerial vehicles (UAVs), and geophysical surveys—that enable detailed site documentation without destructive excavation. In addition, there are some noteworthy applications such as utilizing Generative Adversarial Networks (GANs) to restore ancient Roman coins [3], or developing Neural Machine Translation (NMT) models to aid in the translation of Akkadian texts from cuneiform script and transliteration into English [4]. These technologies not only improve efficiency and accuracy but also foster new paradigms in archaeological inquiry, facilitating large-scale landscape analysis and real-time data processing [1].

Artificial Intelligence (AI) has made significant advancements in natural language processing and data analysis, yet its application in archaeology currently remains limited [5], highlighting a significant gap in research opportunities that could further enrich archaeological research and methodologies. The integration of various technologies into archaeological research represents a paradigm shift, with our project aiming to develop and leverage Generative Artificial Intelligence (Gen AI) to enhance archaeological research by processing and generating accurate and meaningful information, addressing a gap in AI utilization in the field of archaeology.

1.2. Problem Statement

"AI programs will need to have a better grasp of current archaeological knowledge and theory before they can synthesize or build new ideas" [6]. Currently, the primary challenge identified is the inadequacy of existing Large Language Models (LLMs) in synthesizing and generating accurate archaeological knowledge. AI models, such as

GPT-4o, are trained on broad datasets but lack specialization in archaeological texts and research papers, leading to challenges in accurately retrieving and interpreting domain specific information, historical data, and recent discoveries. Many models provide inconsistent or contextually inaccurate information, particularly when faced with complex inquiries that require nuanced understanding of historical contexts, material culture, or archaeological methods. For instance, when tasked with interpreting ambiguous archaeological data, LLMs may generate overgeneralized conclusions or incorrect attributions, such as misidentifying artifacts, drawing parallels between unrelated cultures, or offering outdated interpretations that do not align with current archaeological consensus.

Another major limitation is the context window constraint in AI models. GPT 4o, for example, has a 128K token limit [7], which means it can typically process fewer than ten full text research papers at a time when accounting for system usage and user interaction. This restriction makes it difficult to analyze large datasets or conduct comprehensive cross referencing of multiple sources. Without additional retrieval mechanisms, AI models struggle to maintain continuity and coherence when handling extensive academic research, limiting their usefulness for archaeologists in developing broader analyses using multiple sources of information.

These issues limit the models' utility for researchers and practitioners, as they often provide misleading or overly simplistic answers to inquiries ranging from artifact classification to cultural reconstruction and historical interpretation.

1.3. Motivation

Our group is motivated by a deep interest in the intersection of technology and archaeology, particularly the potential of Gen AI in archaeological research.

Existing LLMs often struggle to accurately reproduce correct archaeological knowledge, leading to vague or sometimes erroneous information. This limitation highlights a significant gap in the application of AI within the field, as while AI has been successfully utilized for tasks such as translation and artifact reconstruction [4], the development of specialized Gen AI tools tailored to archaeology remains underexplored. As current AI models still face limitations in accurately synthesizing

domain specific archaeological knowledge, this project aims to address such gaps through tailored development [6].

By undertaking this project, we aim to fill this gap by developing a model that generates accurate, contextually relevant, and actionable insights to support archaeologists in their work. These insights could help clarify complex archaeological questions, such as interpreting the use of artifacts or understanding ancient social structures, by offering detailed, evidence-based explanations. Ultimately, we seek to create a tool that not only aids in knowledge discovery but also contributes directly to the advancement of archaeological theory and practice.

In addition to advancing AI's role in archaeology, this project presents a valuable learning opportunity in AI development, retrieval based processing, and interdisciplinary research applications. With the rapid rise of generative AI technologies, our team saw this as a chance to gain hands on experience in AI development. Through this project, our aim is to acquire practical skills in working with large language models and developing AI applications, which will not only enhance our technical expertise but also position us at the forefront of innovative research methodologies.

1.4. Objectives

The objective of our project is to develop a Gen AI model specifically designed to address the challenges faced in LLM use for archaeological research, particularly the limitations of existing LLMs in accurately synthesizing and generating relevant archaeological knowledge. By developing the model, enhancing the quality and reliability of information available for the researchers, enabling a research environment that would empower researchers to explore and analyze complex questions with higher efficiency.

This project seeks to address these challenges by developing a Generative AI system specifically tailored for archaeological research. By combining retrieval-augmented generation (RAG) with domain-specific literature, the system will enhance AI's ability to process, interpret, and synthesize archaeological texts, enabling more reliable and precise outputs. The goal is to create a tool that streamlines the research

process, enabling archaeologists to efficiently gather, analyze, and extract insights from vast amounts of literature.

This project aims to provide a tool that enhances the research process for archaeologists by streamlining literature and data gathering, ultimately improving information digestion. By leveraging AI-driven techniques, the system will enable archaeologists to efficiently access, summarize, and interpret vast amounts of research material. Through automated document processing, contextual search capabilities, and knowledge extraction, the tool will facilitate a more efficient workflow, reducing the time spent on manual literature reviews while ensuring comprehensive analysis. By addressing these challenges, the project seeks to bridge the gap between cutting-edge AI technologies and archaeological scholarship, fostering a more dynamic and data-driven approach to research, utilizing the various latest available advancements in the AI-sphere.

To achieve these objectives, the project will involve several key steps. First, a set of archaeological literature will be collected and cleaned to create a specialized text dataset. This dataset will form the foundation for building a RAG model, which will combine information retrieval with generation capabilities to improve the accuracy and relevance of responses. Additionally, we will develop a problem set for testing the model. Finally, a demo UI will be created to facilitate easy interaction with the model, allowing researchers to test it easily.

1.5. Outlines

This report provides an overview of a Generative AI system developed for archaeological research, focusing on the limitations of current large language models (LLMs) and the application of Retrieval-Augmented Generation (RAG). It begins with an introduction to the challenges in archaeology and the potential of AI. The methodology section outlines the development process, including text chunking, metadata extraction, and dataset creation. The results are presented with an evaluation of the system's performance based on a custom benchmark set, showing strong effectiveness with an F1-score of 93.44%. The discussion highlights the challenges faced, such as hallucinations and dataset limitations, and compares the RAG-enhanced model to the baseline. Finally, the report concludes with future plans,

including expert collaboration and improvements to dataset quality and platform scalability.

2. Methodology

This project follows a structured approach to developing a Generative AI system optimized for texts and paper digestions, or in this case, archaeological research. By leveraging Retrieval-Augmented Generation (RAG), which offers significant improvements in grounding model responses by retrieving relevant information to align the model with specialized domain knowledge, this system enhances AI's ability to efficiently process and synthesize archaeological literature [8][9].

For this project, we focus on optimizing the RAG process itself rather than fine-tuning the model. We hypothesize that even a less powerful base model, when paired with an effective RAG system that extracts the most relevant and useful information, can still generate satisfactory results. Fine-tuning requires annotated corpora, extra training cost, and risks overfitting to narrow domain texts—whereas here we focus our resources on maximizing the RAG pipeline's ability to fetch the most relevant evidence [10].

2.1. Developmental Overview of Setup

Python 3.11 will serve as the primary programming language, leveraging its extensive libraries for natural language processing and machine learning to support model training and experimentation.

LangChain will streamline interactions with LLMs, allowing us to efficiently manage input and output pipelines. The vector database ChromaDB will store embeddings and metadata generated from archaeological research papers, providing fast and relevant context retrieval during inference. This RAG process will be critical to improving the accuracy and relevance of model outputs by combining retrieved data with generative responses.

For metadata extraction, we have selected the KeyBERT library and spaCy's en_core_web_sm model. KeyBERT is a simple and effective method for extracting keywords and concepts from text, while spaCy's pre-trained model provides useful named entity recognition (NER), part-of-speech tagging, and dependency parsing. These tools will help in identifying and extracting important concepts and entities from the collected archaeological literature to further enhance the RAG system's ability to understand and generate contextually accurate responses.

2.1.1. Model Selection

In the early phase, we tried offline BERT-based models (e.g., “bert-base-uncased”) for generating embeddings, hoping to avoid API dependencies. However, retrieval benchmarks on a small validation set revealed relatively poor semantic alignment, making BERT unsuitable for our needs.

Recognizing that retrieval quality drives overall RAG performance more than the underlying generative model, we shifted to using OpenAI’s embedding model (“text-embedding-3-small”), which delivered significantly better semantic similarity scores. LangChain’s robust support for OpenAI—complete with built-in client wrappers and streamlined prompt handling—further simplified development.

We also experimented briefly with Google Gemini via its beta API. While Gemini showed promise, the lack of mature Python libraries and community examples made integration more cumbersome. Given limited development time, we opted to standardize on OpenAI's models (“gpt-4o” and “text-embedding-3-small”) for both embedding and generation, balancing performance, reliability, and developer convenience.

2.1.2. Abandoned Trial

We also experimented with a keyword-based retrieval strategy using BM25 to rank relevant documents before trying the metadata. Initially, we used BM25 independently to identify top-ranked passages based on term frequency and inverse document frequency. While this method worked reasonably for surface-level keyword matches, it consistently failed to capture semantically relevant content where the terminology used in the query did not exactly match the document phrasing [11].

To improve performance, we then tested a hybrid retrieval approach: BM25 and dense embeddings were used together, with top-ranked chunks from both sources merged or re-ranked based on combined scores. However, even with this hybrid setup, retrieval benchmarks revealed that keyword-based components often introduced noise or prioritized overly generic content. In particular, we observed that keyword-based candidates tended to dilute the semantic precision offered by the embedding model. Ultimately, this combined method showed lower top-1 and top-3 retrieval accuracy compared to pure embedding-based retrieval, especially in domain-specific queries

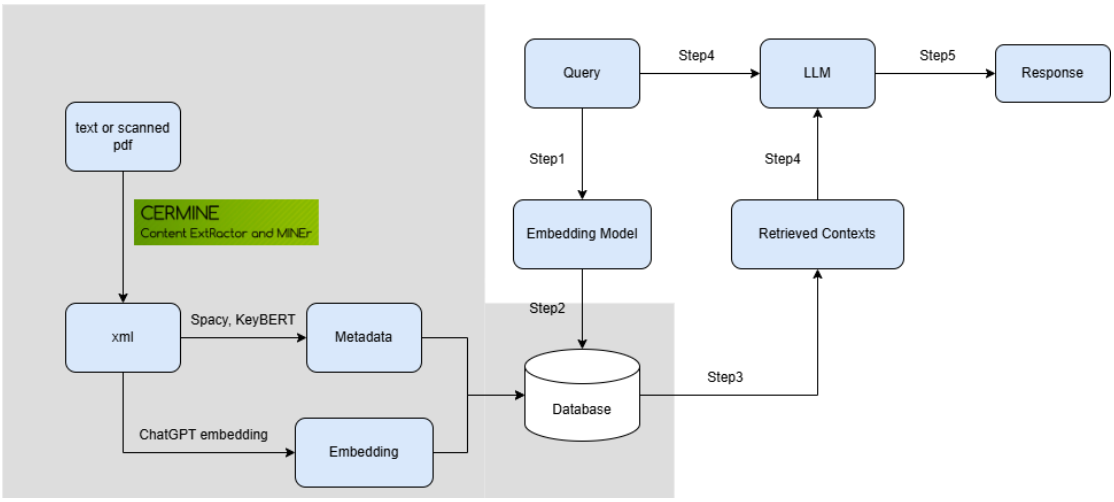

Figure 1. Overview of the RAG process (gray background: database preparation)

involving less frequent terms. Based on these findings, we abandoned the BM25-enhanced retrieval route in favor of fully dense retrieval with optional metadata weighting.

2.2. Data Collecting, Extraction and Preparation

For the data collection phase, we collaborated closely with archaeologists to identify and select relevant textual data that can support the development of our AI model. This will include a wide range of archaeological research materials such as academic papers, books, reports, and other pertinent resources. The selected data will be converted into XML format to facilitate efficient processing and integration with the model.

As shown in figure 1's gray part. The first step in the pipeline is document conversion and structuring, where research paper PDFs are converted into XML format using CERMINE [12], a content extraction API. Before split the xml into chunks, the spaCy's "en_core_web_sm" and KeyBERT model will be used to analysis the file to extract the metadata. Following this, the text undergoes splitting and embedding using LangChain. Currently, the text is divided into chunks 3800 characters, with an overlap of 500 characters to maintain context. Once split, each text chunk is embedded using ChatGPT's embedding model "text-embedding-3-small", and both the embedding vectors, metadata and corresponding text chunks are stored in ChromaDB, a vector database. A vector database is a specialized type of database designed to store and manage vectors, numeric representations of data, such as the embeddings generated

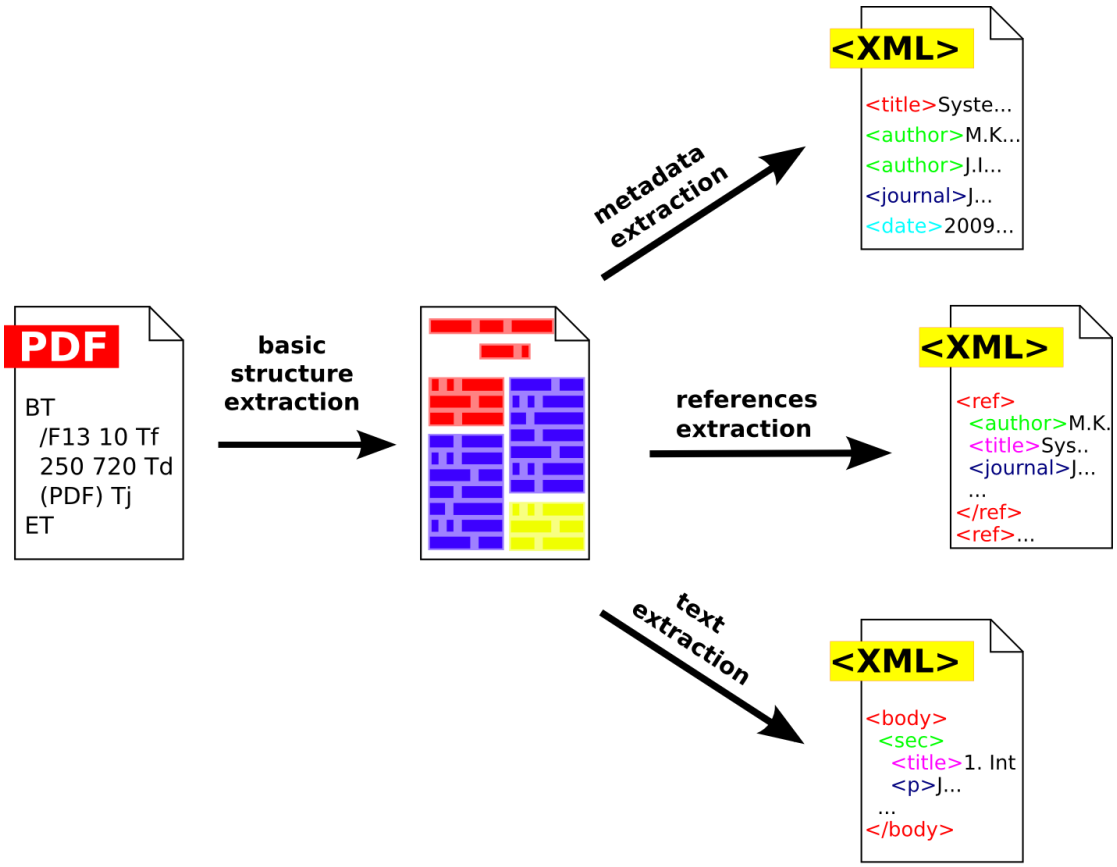

Figure 2. CERMINE Workflow

from text. This is useful as it captures the semantic meaning of text, transforming it into a high-dimensional numerical format that allows for efficient similarity searches. Thus, this implementation ensures efficient and accurate retrieval of relevant information when responding to queries.

Initially, we tried using the Umi-OCR [13] for optical character recognition (OCR), but we faced several issues. The OCR system frequently stalled during processing, and it often misidentified characters that didn't exist, which led to significant errors and hindered progress. We switched to CERMINE [12], which offers relatively better stability and, also, outputs data in an XML structure. The XML format's structured approach has the advantage of preserving metadata and text categorization, which enhances both the extraction and RAG process [14]. However, we did not fully utilize the advantages of XML tags. Since the XML structure tree generated from the PDF could not ensure proper content allocation, it often resulted in highly inconsistent chunk sizes. I believe this would impact the performance of the embeddings, so I decided to temporarily abandon the use of these tags. As a result, we had to resort to

Figure 3. Converted XML file with Metadata

chunking data based on the size of the chunks (defined by chunk size) rather than utilizing the XML structure tree to guide this process. This limitation will be addressed in our future plans, where we aim to refine the XML processing to ensure that we can take full advantage of the semantic hierarchy within the XML tags for more accurate data chunking and retrieval.

2.2.1. OCR Testing and Selection

Initially, we experimented with several Optical Character Recognition (OCR) tools, including Adobe Acrobat's built-in OCR engine [15]. However, this tool did not perform well due to frequent errors and inconsistencies in character recognition, especially when handling more complex documents. We then switched to Umi-OCR [13] for OCR processing, but this also came with several issues. The OCR system frequently stalled during processing, and it often misidentified characters that didn’t exist, which led to significant errors and hindered progress.

In light of these challenges, we transitioned to using CERMINE, which offered better stability and more reliable output. CERMINE outputs data in an XML structure, which provided a more structured approach to data extraction and ensured that metadata such as section headings, bibliographic references, and author information was preserved. This structure greatly enhanced the extraction and RAG (Retrieval-Augmented Generation) process, as the preserved metadata provided additional semantic context, improving the accuracy and relevance of the responses generated by the model.

2.2.2. XML and Metadata

The XML format offered a significant advantage by retaining key metadata elements, such as section headers and bibliographic references, which can provide additional semantic context, improving the accuracy and relevance of the responses generated by the model [14]. However, while CERMINE provided robust XML output, we found that the XML structure tree generated from the PDFs did not always ensure proper content allocation. This led to inconsistent chunk sizes during the document processing phase, which affected the performance of the embedding models. As a result, we decided to temporarily abandon the use of XML tags for guiding chunking. Instead, we used a fixed chunk size, defined by the number of characters, to split the documents.

Despite these limitations, the XML structure remained beneficial, and we plan to address these challenges in the future. Specifically, we aim to refine the XML processing to ensure that the semantic hierarchy within XML tags can be fully utilized, allowing for more accurate data chunking and improving the retrieval process.

2.2.3. Data Quality and Cleanup

After selecting CERMINE for document conversion, we encountered several data quality issues related to the XML output generated by the tool. While CERMINE proved to be more stable and efficient compared to earlier OCR tools, it still presented challenges in handling complex documents. In particular, about 15.5% of the processed documents required cleanup due to issues such as garbled text, missing sections, and non-standard formatting.

To mitigate these issues, we developed an automated cleanup process integrated into the document conversion pipeline. This process was specifically designed to address the flaws introduced during the XML conversion phase. The cleanup algorithm detected and removed corrupted content, handled placeholder characters, and normalized the formatting of different text structures. If the conversion resulted in text that could not be preserved or confidently recovered, the cleanup process would either discard that portion of the document or remove the entire document from the dataset.

While this step improved the reliability of our data, it also meant that several documents lost large portions of content, rendering them partially or entirely unusable

for downstream AI tasks. However, this tradeoff was essential to ensure the quality and academic rigor of the content used by the system, as it prioritized working with verified, high-quality source materials. By filtering out compromised documents, we maintained the integrity and accuracy of the responses generated by the AI, which is particularly crucial for research domains like archaeology [16].

2.3. Query Processing with RAG Workflow

As illustrated in Figure 1, when a user inputs a query, the system first embeds the query using ChatGPT's embedding model. The resulting embedding is then used by the local RAG (Retrieval-Augmented Generation) model to retrieve the most relevant text chunks from the database based on vector similarity and metadata.

RAG is an AI technique that combines retrieval-based search with generative AI, allowing the model to pull in relevant external information rather than relying solely on its pre-trained knowledge. This enhances the accuracy and relevance of responses, especially in specialized domains like archaeology, where up-to-date and domain-specific information is crucial [17].

These retrieved text chunks are then integrated into a structured prompt using LangChain's templating tools, which is subsequently sent to GPT-4o. The model then generates a response based on the RAG-enhanced prompt, ensuring that the output is contextually accurate, grounded in the retrieved research data, and more informative than what the model could generate from its base knowledge alone. Finally, to allow for convenient testing of the model, a temporary UI, currently a web interface, was developed through Gradio, a Python package for swift web application building, thus allowing users to conduct context-free tests either locally or remotely to verify its functionality.

This RAG-enhanced workflow ensures that the generated answers are grounded in specific archaeological data, thereby minimizing the risks of hallucination and improving accuracy. As highlighted in [18], RAG reduces hallucinations by grounding responses in retrieved evidence, offering both accuracy and transparency.

2.4. Prompt Engineering and User Interface

To allow for convenient testing of the model, a temporary UI was initially developed using Gradio, a Python package for swift web application building. This Gradio-based interface enables users to conduct context-free tests either locally or remotely to verify the functionality of the system, providing a flexible and rapid means of testing during development.

However, for more permanent deployment, the system will be transitioned to a webpage as the final version. The webpage-based interface will allow for more robust and scalable user interaction, making it easier for users, such as archaeologists, researchers, and students, to submit queries and interact with the AI system in a more user-friendly environment. The webpage will integrate the core functionalities of the RAG pipeline, allowing users to retrieve detailed, domain-specific information from the research database with ease, and will be designed to support long-term usage and system scalability.

Prompt Engineering plays a critical role in guiding the system's response accuracy and relevance [19]. In our interface, we implemented a dynamic prompt system that allows users to choose between simple or detailed response modes using a simple button interface. These two modes correspond to different prompt templates that adjust the level of detail, tone, and complexity of the response. For simpler queries, the system generates brief and concise answers using a basic prompt template, while more complex queries trigger detailed, multi-layered prompts that encourage in-depth responses, tailored to the needs of researchers or archaeologists seeking comprehensive insights.

This dynamic approach to prompt engineering is essential for improving both the quality and usability of the system. It ensures that the generated responses align with the expectations of the user—whether they require quick, factual information or need a more nuanced, scholarly response grounded in archaeological research. By giving users control over the response format, we can better cater to the varying demands of different user groups, ensuring the system meets their specific needs [19].

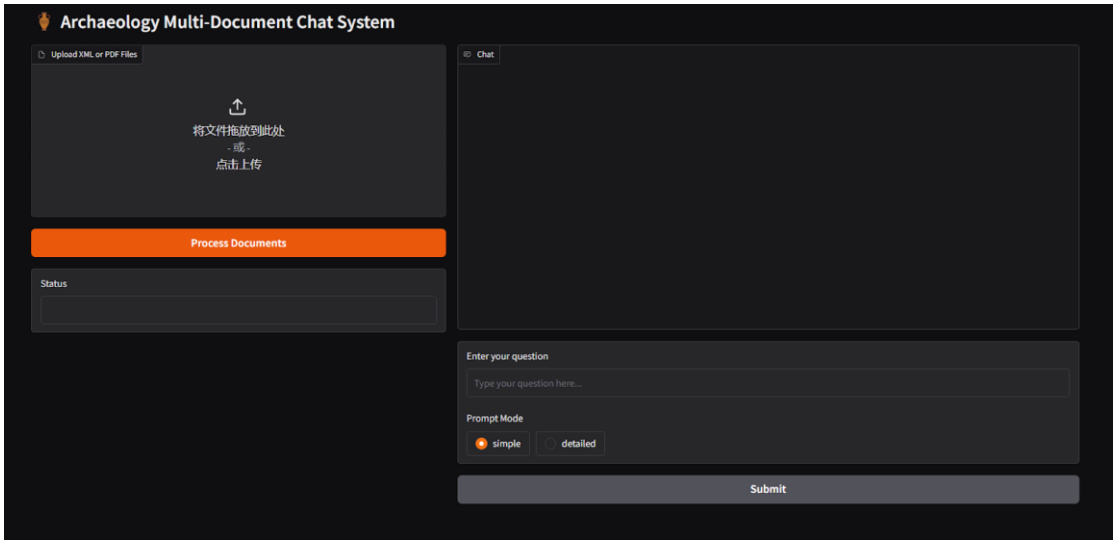

Figure 4. Webpage-based User Interface

3. Tests and Results

This section outlines the test case design, progress and results of the project, focusing on the advancements made and the challenges encountered. Although the results are satisfactory, they cannot reflect the actual level well due to the lack of an expert-proposed problem set for performance testing.

3.1. Test Design

The system will be evaluated using a custom benchmark set of 125 questions, which includes 100 questions based on text content and 25 from papers that was not converted properly to test hallucinations. The results will be classified into correct and incorrect, and for fake questions, either avoided fake questions or answered fake questions.

The report will use F1 scores as the evaluation metrics. These metrics will help assess the model's performance on the test set. As the output is text-based and the dataset is relatively small, automated comparison can be complex and error-prone. Therefore, each response will be manually checked against the original sources.

Initially, we intended to work with archaeologists to design a more comprehensive test set that would better capture the nuances of archaeological research. However, due to difficulty in recruiting specialists, we created a simplified set ourselves.

Without the participation of experts, the test set may not fully reflect the complexity of the field.

Type	Question	Expect Answer	Source
True	Which century is the earliest mention of the Ghilghilchay defensive long wall dated to?	7th century AD	AlievEtal2006-GhilghilchayDefensiveLongWall.pdf
True	Which river is mentioned as potentially causing the flooding that led to the abandonment of Sasanian settlements and irrigation systems in the Mughan Steppe?	The Aras River	AlizadehEtal2021-SasanianCollapseMughan.pdf
Fake	What are the four main factors that explain the Elymaeans' rise as a major power in Khuzestan?	NA	Alizadeh1985-ElymaeanOccupationofLowerKhuzestan.pdf

Table 1. Example Questions (Type indicates whether it can be answered)

3.2. Results

The system's performance on the benchmark set of 125 questions is summarized below.

	Correct/Avoid	Incorrect/Answered
True Question	92 (TP)	8 (FN)
Fake Question	6 (FP)	19 (TN)

Table 2. Classification Breakdown of Testing Results

Based on these data, we can calculate the following metric:

Metric	Formula	Value
Accuracy	$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{111}{125} = 88.80\%$
Precision	$\frac{TP}{TP + FP}$	$\frac{92}{98} = 93.88\%$
Recall	$\frac{TP}{TP + FN}$	$\frac{92}{100} = 92\%$
F1-Score	$\frac{2 * Precison * Recall}{Precision + Recall}$	92.93%

Table 3. Performance Metrics

3.3. Discussion

The following section presents an analysis of the performance metrics derived from the test cases outlined in the previous section. By examining the precision, recall, and F1-score, we can assess the strengths and weaknesses of the system and highlight areas for further improvement.

In our test set, the model achieved an F1-score of 93.44%, which is a strong indication of overall performance, particularly in applications like archaeology, where both accuracy and caution are critical.

The F1-score is important because it gives equal weight to precision and recall, ensuring that a model is not just accurate but also able to retrieve as much relevant information as possible without generating false responses. The relatively high F1-score suggests that this model can managing this balance well.

While, comparing the current results with those from the mid-term report, we observed a drop in accuracy from 91% to 88.8%. This decline can primarily be attributed to two factors:

Increased Complexity of the Test Set: The current test set is more complex than the previous one, as it includes questions of greater difficulty and nuance. The shift from a simpler set of 75 questions to the more challenging 125-question benchmark likely contributed to the drop in performance.

Performance on Fake Questions: The model's performance on fake questions (those designed to test hallucinations) is particularly concerning. Out of the 25 unanswerable questions, the model made 6 errors, resulting in a 24% error rate. This indicates that while the model performs well on factual questions, it struggles with distinguishing between answerable and non-answerable questions. This highlights the need for further optimization, particularly in handling hallucinations and ensuring that the model refuses to generate speculative or false answers.

The significant error rate on the fake questions highlights a flaw with the current system. While the model correctly retrieves chunks based on vector similarity to the prompt, the issue arises when the retrieved chunks are semantically related to the prompt, but not directly relevant to the actual question.

In the current setup, when the model generates an answer, it extracts the most relevant chunks from the database based on the prompt's content. However, the retrieved chunks, although closely related to the terms in the prompt, might contain information that is contextually unrelated to the specific question being asked. For unanswerable questions, this creates a risk where the model might misinterpret these related but irrelevant chunks as valid sources of information, leading to hallucinations.

The real challenge here is not just the extraction of these chunks, but the model's ability to distinguish between relevant information and irrelevant, yet semantically similar, content. The model needs to recognize when there is no appropriate data available to answer the question and refuse to generate an answer instead of relying on the retrieved chunks, which may be contextually close but not actually answering the query.

This issue could be addressed by refining the prompt design and introducing a mechanism that makes the model aware that some queries do not have corresponding answers in the database [20][21]. This awareness could help the model recognize when to avoid answering entirely rather than generating a response based on irrelevant or tangentially related content. Additionally, introducing specific markers or conditional logic in the prompt might guide the model to more accurately handle unanswerable questions, reducing the number of hallucinated responses.

3.3.1. Without RAG

When comparing the RAG-enhanced model to the baseline GPT-4o model (without RAG), significant improvements were observed. The baseline model, which relied solely on its pre-trained knowledge, often struggled with domain-specific queries in archaeology. Without access to external sources or reference material, the model produced incomplete or factually incorrect answers. These shortcomings were particularly evident in the handling of questions that required precise and contextually grounded information from the documents.

In contrast, the RAG-enabled model performed much better by incorporating document grounding. By retrieving relevant text chunks from the indexed database, the system was able to provide contextually accurate and source-backed responses. This dramatically reduced the occurrence of hallucinations, showcasing the tangible benefits of RAG in reducing errors and improving response reliability.

In conclusion, the RAG enhancement helped improve the model's ability to retrieve verifiable, relevant information, making it significantly more reliable than the baseline GPT-4o model. This underscores the importance of integrating external knowledge sources in specialized domains like archaeology, where factual accuracy and reliability are paramount.

4. Future Plan

While the current implementation of the RAG system has shown promising results, there are several areas that can be improved to further enhance its performance and real-world applicability. Below are the key directions for future development:

1. Collaboration with Archaeological Experts: A major limitation in the current phase of the project is the lack of input from archaeological experts. Due to this, the test set was simplified and may not reflect the true complexity of archaeological research. More insightful and analytical questions designed by experts are needed to better evaluate the model's performance on complex, research-driven inquiries.

2. Improving Dataset Quality: The current dataset used for training and evaluation is not of optimal quality, which could impact the system’s overall performance. To address this, we plan to work on both new data acquisition and refining our

existing data-cleaning methods. By improving the dataset, we aim to reduce errors, enhance retrieval quality, and ultimately boost the precision and recall of the system.

3. Transition to an Online Platform: Although we have developed a functional web-based user interface (UI), the current infrastructure lacks the necessary server support for wider deployment. As a result, the system is limited to local testing and demonstration. Moving forward, we plan to transform the UI into an online platform, which will enable remote access for a broader audience. This transition will allow for more extensive testing, user interaction, and potential scaling of the system for practical use in archaeological research.

4. Enhancing Text Chunking Strategies with XML Tags: Another important aspect of the system that can be improved is text chunking. By leveraging XML tags from the structured documents, we can enhance the chunking strategy by considering semantic metadata such as headings, references, and citations. This will allow the system to process the text more effectively, preserving the context and relationships between sections. We plan to refine chunking strategies using these XML tags to improve text segmentation, ensuring that relevant information is better preserved and retrieved during the process.

5. Adopting Advanced Retrieval Techniques: To further optimize RAG's performance, we aim to implement more advanced retrieval techniques such as joint optimization and query rewriting [22][23]. Joint optimization will help improve the alignment between the retrieval and generation processes, ensuring more accurate responses. Query rewriting can enhance the system's ability to better understand and rephrase user queries, thereby improving the retrieval of relevant information. These techniques are crucial for refining the system's ability to handle complex queries and improve the quality of the generated responses.

By addressing these aspects, we are confident that the RAG system can be significantly enhanced, making it a more powerful tool for archaeological research and other specialized fields. The ultimate goal is to create a robust, accurate, and scalable AI system capable of providing real-time, domain-specific insights, grounded in verified academic content.

5. Conclusion

This project explored the potential of Gen AI in addressing challenges faced by archaeological research, particularly in synthesizing and generating accurate, domain-specific knowledge. By developing a RAG framework, the goal was to improve AI models' ability to provide meaningful and relevant insights for archaeological inquiries. While the results showed some progress, they also highlighted many limitations that need to be addressed for the model to meet expectations, especially the hallucinations.

Despite the challenges faced, this project lays the groundwork for a specialized Gen AI tool for archaeology. By continuing to refine the retrieval logic, enhancing the dataset, and incorporating expert feedback, we are optimistic that the model can evolve into a valuable resource for researchers. This project contributes to the growing intersection of AI and archaeology, with the potential to enhance the quality and depth of archaeological knowledge synthesis, offering new insights into ancient civilizations.

6. Reference

[1] C. Papadopoulos, C. Ioannidis, and K. Maria, “Digital Tools for Data Acquisition and Heritage Management in Archaeology and Their Impact on Archaeological Practices,” *Heritage*, vol. 7, no. 1, pp. 107–121, 2024. doi: 10.3390/heritage7010107

[2] S. Calugay, "The destructive nature of archaeology," *The Post Hole*, Oct. 2015. [online]. Available: https://www.theposthole.org/read/article/343.

[3] M. Altaweel, A. Khelifi, and M. H. Zafar, "Using generative AI for reconstructing cultural artifacts: Examples using Roman coins," *Journal of Computer Applications in Archaeology*, vol. 7, no. 1, pp. 301–315, 2024. doi: 10.5334/jcaa.146.

[4] T. But, "The Latest AI Innovations in Archaeology," *Historica*, Aug. 15, 2024. [online]. Available: https://www.historica.org/blog/the-latest-ai-innovations-in-archaeology.

[5] A. Al-Sabaawi, L. Bai, and X. Zhang, “Managing Artificial Intelligence in Archaeology: An overview,” *J. Archaeol. Sci.*, vol. 160, p. 105432, 2024.

[6] P. J. Cobb, "Large language models and generative AI, oh my!: Advances in archaeological practice," *Advances in Archaeological Practice*, vol. 11, no. 3, pp. 314–330, 2023. [online]. Available: https://doi.org/10.1017/aap.2023.20.

[7] OpenAI, “GPT-4o,” OpenAI Platform, 2025. [online]. Available: https://platform.openai.com/docs/models/gpt-4o. [Accessed: Apr. 21, 2025].

[8] Y. Hoshi, D. Miyashita, Y. Ng, K. Tatsuno, Y. Morioka, O. Torii, and J. Deguchi, “Retrieval-Augmented Generation for Large Language Models: A Survey,” *arXiv preprint*, 2023. [online]. Available: https://arxiv.org/abs/2312.10997

[9] H. Soudani, E. Kanoulas, and F. Hasibi, “RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture,” *arXiv preprint*, 2024. [online]. Available: https://arxiv.org/abs/2401.08406

[10]Soudani, H., Kanoulas, E., & Hasibi, F. (2024). *Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge*. arXiv. https://arxiv.org/abs/2403.01432

[11]M. Mehta, "Understanding Okapi BM25: A Guide to Modern Information Retrieval," *AdaSci*, 2021. [online]. Available: https://adasci.org/understanding-okapi-bm25-a-guide-to-modern-information-retrieval. [Accessed: 20-Apr-2025].

[12] CeON, "CERMINE: Content ExtRactor and MINEr," 2025. [online]. Available: http://cermine.ceon.pl/index.html. [Accessed: 20-Apr-2025]

[13] hiroi-sora, "Umi-OCR," GitHub, 2025. [online]. Available: https://github.com/hiroi-sora/Umi-OCR. [Accessed: 20-Apr-2025]

[14] P. Badakhchani, "XML: Documents with semantics," *Hashnode*, May 7, 2024. [online]. Available: https://pedbad.hashnode.dev/xml-documents-with-semantics. [Accessed: Apr. 20, 2025].

[15] Adobe. "Adobe Acrobat." *Adobe*, [online]. Available: https://www.adobe.com/sg/acrobat.html. [Accessed: 20-Apr-2025]

[16] L. Casini, N. Marchetti, A. Montanucci, V. Orrù, and M. Roccetti, "A human–AI collaboration workflow for archaeological sites detection," Scientific Reports, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-36015-5.

[17] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020. [online]. Available: https://arxiv.org/abs/2005.11401. [Accessed: 22-Apr-2025].

[18] "RAG vs. Fine-Tuning vs. Both: A Guide for Optimizing LLM Performance," Galileo, 2023. [online]. Available: https://www.rungalileo.io

[19] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," *arXiv*, 2022. [online]. Available: https://arxiv.org/abs/2210.03629. [Accessed: 22-Apr-2025].

[20] L. Cao, "Learn to Refuse: Making Large Language Models More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism," *arXiv preprint arXiv:2311.01041*, 2023. Available: https://arxiv.org/abs/2311.01041

[21] S. Diao et al., "R-Tuning: Instructing Large Language Models to Say 'I Don't Know'," *arXiv preprint arXiv:2311.09677*, 2023. Available: https://arxiv.org/abs/2311.09677

[22] W. Zheng and L. Yin, "Characterization inference based on joint-optimization of multi-layer semantics and deep fusion matching network," *PeerJ Computer Science*, vol. 8, p. e908, Apr. 2022, doi: 10.7717/peerj-cs.908.

[23] J. Liu, and B. Mozafari, "Query Rewriting via Large Language Models," *arXiv preprint*, 2024. [online]. Available: https://doi.org/10.48550/arXiv.2403.09060

7. APPENDIX

LIST OF LIBRARIES AND VERSIONS

Libraries	Version
ChromaDB	0.5.15
Gradio	5.9.1
spaCy (en_core_web_sm)	3.8.0
KeyBERT	0.9.0
LangChain	0.3.18
LangChain-OpenAI	0.3.4
LangChain-Community	0.3.17
OpenAI	1.61.1