FYP24087 Interim Report

Jiho Lam 3035824273

Generative AI for Researching Archaeology

Yian Fan 3035952886

Project Objective

Under the supervision of Professor Peter Cobb, this project seeks to integrate Artificial Intelligence with the field of Archaeology to develop a Generative AI that addresses the limitations of available AI models while exploring the current potential of generative AI's abilities for archaeological research. By enhancing AI’s ability to process and interpret archaeological literature, this project seeks to improve research efficiency and support researchers in their work.

Thus, 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 various 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.

Project Background

Artificial Intelligence has made significant advancements in natural language processing and data analysis, yet its application in archaeology currently remains limited. Current 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. Additionally, access to specialized archaeological research papers is often restricted, with many key texts behind paywalls or difficult to source. This creates a barrier for researchers seeking to leverage AI for efficient literature review and data synthesis.

Another major limitation is the context window constraint in AI models. GPT-4o, for example, has a 128K token limit, 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. As current AI models still face limitations in accurately synthesizing domain-specific archaeological knowledge, this project aims to address such gaps through tailored development [1].

This project seeks to address these challenges by developing a Generative AI system specifically tailored for archaeological research. By integrating Retrieval-Augmented Generation (RAG) and fine-tuning AI models with domain-specific literature, the system will enhance AI's ability to process, interpret, and synthesize archaeological texts, enabling for 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.

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. Despite some successful AI applications in areas like artifact restoration and text translation [2][3], specialized Gen AI for archaeology has been underexplored, creating a gap for further innovation in archaeological methodologies. 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, developing AI applications, which will not only enhance our technical expertise but also position us at the forefront of innovative research methodologies.

Project Methodology

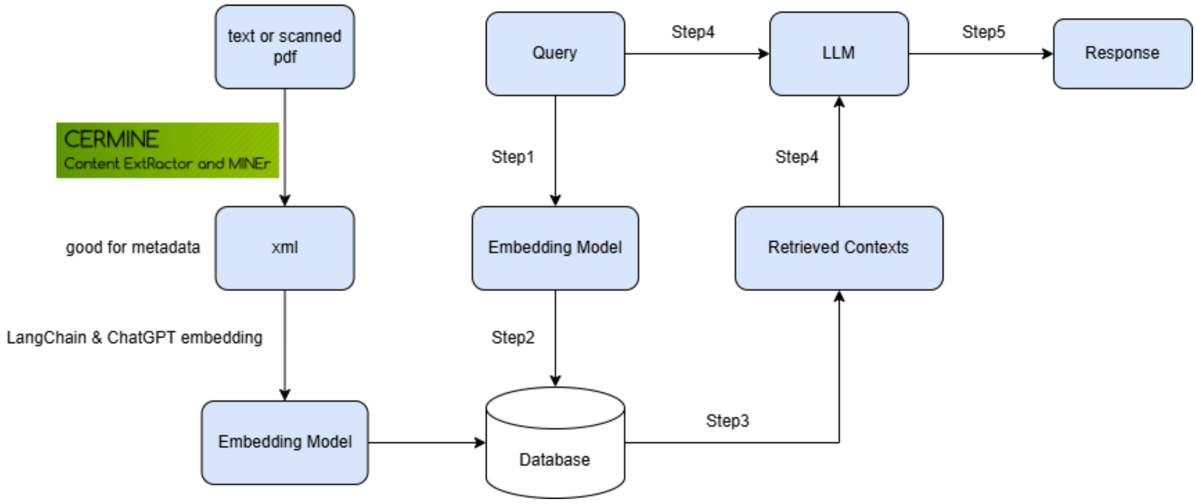


Figure 1. Overview of the RAG process used in training an archaeology-focused LLM

This project follows a structured approach to developing a Generative AI system optimized for texts and paper digestions, or in this case, archaeological research. The methodology involves multiple stages, including document conversion, text processing, embedding storage, and retrieval-augmented response generation. By leveraging Retrieval-Augmented Generation (RAG), which offers significant improvements in grounding model responses by retrieving relevant information, and integrating fine-tuning techniques to align the model with specialized domain knowledge, this system enhances AI's ability to efficiently process and synthesize archaeological literature [4][5]. The overall process has been visualized in *figure 1*.

The first step in the pipeline is document conversion and structuring, where research paper PDFs are converted into XML format using CERMINE, a content extraction API. XML is chosen over plain text because it retains extensive metadata and allows for the categorization of different types of text through structured labels. This structured format facilitates better text processing and improves the AI system's ability to interpret and utilize the content more effectively.

Following this, the text undergoes splitting and embedding using LangChain. Currently, the text is divided into chunks of 1,000 characters, with an overlap of 200 characters to maintain context. Once split, each text chunk is embedded using ChatGPT's embedding model, and both the embedding vectors and corresponding text chunks are stored in ChromaDB, a vector database. In this process, RAG enhances the quality of the generated responses by retrieving relevant data chunks, thus reducing hallucinations and improving the accuracy and transparency of the output [6].

A vector database is a specialized type of database designed to store and manage vectors, numeric representations of data, such as the embeddings generated 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.

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. 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. To further improve retrieval capabilities, we plan to integrate advanced retrieval techniques, such as joint-optimization and query rewriting, which will help refine the RAG framework [7][8].

These retrieved text chunks are then integrated into a structured prompt using LangChain's templating tools, which is subsequently sent to ChatGPT-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, though more permanent UI interfaces are planned.

What has been accomplished

The first semester of this project primarily focused on the exploration and testing of various AI models, investigating and implementing retrieval-based strategies, and developing initial infrastructure to support the development of Generative AI for archaeological research. Key milestones include evaluating different AI models, integrating Retrieval-Augmented Generation (RAG), addressing OCR challenges, and developing a temporary web interface for testing.

A primary objective was to identify which AI models were best suited for our project. We tested Google Gemini, SciBERT, GPT-4, among others, comparing their ease of use, accuracy on a small sample testing set, and integration with other tools and technologies. OpenAI's GPT-4 was eventually selected for our core model due to its robust technical capabilities, ease of use, and compatibility with advanced AI frameworks such as LangChain.

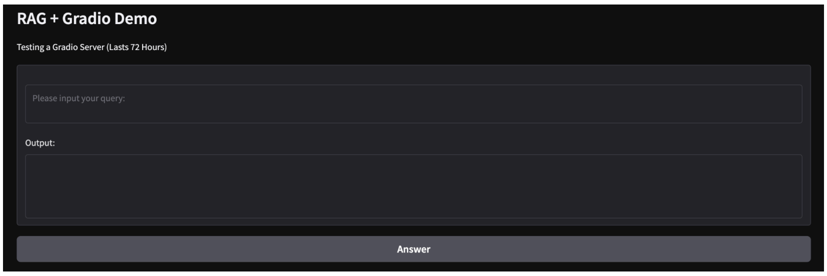
To enhance the model's ability to process archaeological texts, we explored different retrieval and training strategies, including Retrieval-Augmented Generation (RAG) and fine-tuning. The RAG framework was successfully integrated, allowing the system to retrieve relevant text chunks from a vector database to be used to generate responses. This improved the AI's ability to generate contextually correct and accurate answers rather than relying solely on pre-trained knowledge.

A vector database (ChromaDB) was implemented to store and efficiently retrieve embedding vectors of research texts. This enables similarity-based searches, ensuring that responses are based on semantically relevant text segments rather than simple keyword matching. The integration of vector search significantly improves the accuracy and relevance of retrieved information.

Processing scanned archaeology papers required overcoming OCR (Optical Character Recognition) challenges. Initial attempts at utilizing OCR for PDF-to-text conversion were inefficient and often inaccurate due to the unique layouts and templates of various research texts. Furthermore, text within diagrams and images caused discrepancies, leading to loss of critical information. The manual process of converting each PDF was also highly time-consuming. Additionally, there were no cost-efficient OCR solutions that provided both high accuracy and reliable formatting preservation, making it necessary to explore alternative automated methods for document processing.

Therefore, we opted to utilize CERMINE, a publicly available API for PDF-to-XML conversion, specifically designed for research papers and academic texts. To efficiently process large volumes of PDFs, we developed a script that automates the conversion process, enabling efficient batch processing.

Finally, to facilitate early-stage testing, a temporary web interface was developed using Gradio. This allows users to interact with the system, input queries, and evaluate the AI's responses in real-time. A screencap of the web interface is shown below:



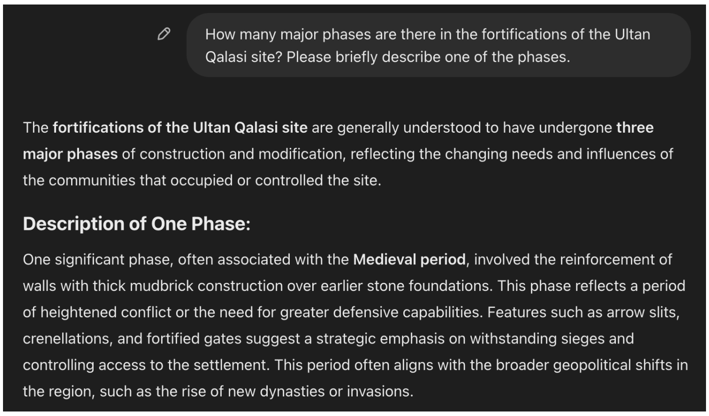
RAG + Gradio Demo

Testing a Gradio Server (Lasts 72 Hours)

Preliminary Results and Observations

To evaluate the system's performance, we conducted tests using 15 research papers, with each paper assessed using five questions we created manually by parsing through the papers ourselves. The primary focus was on straightforward data extraction tasks, assessing the accuracy of retrieval and response generation. Out of 75 total responses, 68 were correct, yielding an approximate accuracy of 91%. The remaining responses contained either incomplete or inaccurate information, highlighting areas for further refinement.

For comparison, we also tested the same GPT-4 without our RAG-based enhancements. The untrained model often provided incorrect or incomplete information, as it lacked direct access to the source texts, demonstrating the improvements provided by our RAG approach and implementation. Below is an example, where when asked the same question, our model provided the correct answer using our provided source texts.

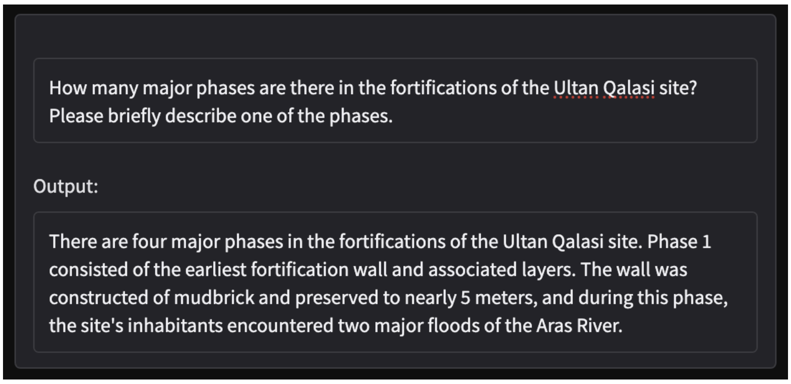


How many major phases are there in the fortifications of the Ultan Qalasi site? Please briefly describe one of the phases.

The **fortifications of the Ultan Qalasi site** are generally understood to have undergone **three major phases** of construction and modification, reflecting the changing needs and influences of the communities that occupied or controlled the site.

Description of One Phase:

One significant phase, often associated with the **Medieval period**, involved the reinforcement of walls with thick mudbrick construction over earlier stone foundations. This phase reflects a period of heightened conflict or the need for greater defensive capabilities. Features such as arrow slits, crenellations, and fortified gates suggest a strategic emphasis on withstanding sieges and controlling access to the settlement. This period often aligns with the broader geopolitical shifts in the region, such as the rise of new dynasties or invasions.



How many major phases are there in the fortifications of the Ultan Qalasi site? Please briefly describe one of the phases.

Output:

There are four major phases in the fortifications of the Ultan Qalasi site. Phase 1 consisted of the earliest fortification wall and associated layers. The wall was constructed of mudbrick and preserved to nearly 5 meters, and during this phase, the site's inhabitants encountered two major floods of the Aras River.

While the system demonstrated relatively high accuracy in extracting specific information, certain challenges were identified, particularly in handling large-scale overview analyses. One key limitation might be the context window size of the model, which restricts the number of tokens that can be processed in a single query. This constraint makes it difficult to analyze extensive documents or synthesize broad summaries effectively. Additionally, the number of tokens used per query may not always be sufficient to capture complex relationships between different sections of a paper, leading to incomplete or less comprehensive responses.

Furthermore, while the system achieved 91% accuracy, the remaining 9% of responses were either incomplete or incorrect, indicating room for improvement. Enhancing text chunking strategies, refining retrieval mechanisms, and optimizing prompt engineering could further improve accuracy and ensure more contextually aware responses. Future iterations will focus on expanding token utilization, refining retrieval strategies, and exploring additional methods to enhance the system's ability to perform large-scale document analysis.

What will be done

During the second semester, our primary focus would be on improving our ability to evaluate system performance, enhance retrieval capabilities, improving the user interface, and refining answer generation to ensure the AI system becomes a more effective tool that aligns closer to our project goal.

A key priority will be implementing a categorized question set to systematically assess the AI's accuracy and reasoning abilities. With the help of trained archaeologists, developing an extensive graded question bank would assist in testing and measuring the model's ability to extract direct information, as well as multi-step reasoning tasks, which evaluate its capacity to synthesize data from multiple sources. With this, we can ensure that not all feedback is subjective analysis by our supervisor, but real tangible values that we can work towards, ensuring progress is truly being made.

Additionally, we will introduce multi-turn interactions, allowing the system to retain contextual information across consecutive queries, improving response coherence in ongoing discussions.

To further enhance retrieval accuracy, we will refine our RAG implementation by leveraging metadata utilization, which helps categorize and rank retrieved documents more effectively. Additionally, distributed filtering and ranking mechanisms will be implemented to prioritize the most relevant text chunks. Improvements in dynamic query expansion will also ensure better retrieval of contextually relevant information, leading to more precise and well-structured responses.

Beyond technical enhancements, we aim to develop a more user-friendly interface by transitioning from the current Gradio demo to another more permanent interface. This will provide better accessibility, improved usability, and a more intuitive interaction experience for researchers.

Lastly, we will refine answer generation through prompt optimization and potential fine-tuning, ensuring that responses are not only factually accurate but also structured, contextually aware, and suitable for archaeological research and discussion.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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

[5] 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

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

[7] 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.

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