FYP24087 Project Plan

Jiho Lam 3035824273

Yian Fan 3035952886

Project Background

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" (Calugay, 2015).

However, the integration of various technologies into archaeological research represents a paradigm shift, with our project aiming to develop and leverage generative AI to enhance archaeological research under the supervision of Professor Peter Cobb by processing and generating accurate and meaningful information, addressing a gap in AI utilization in the field of archaeology.

Problem Identification

"AI programs will need to have a better grasp of current archaeological knowledge and theory before they can synthesize or build new ideas." (Cobb, 2023)

Currently, the primary challenge we have identified and are aiming to address is the inadequacy of existing LLMs in effectively synthesizing and generating accurate archaeological knowledge, with many providing vague or misleading information when confronted with archaeological inquiries of various degrees of complexity, limiting their utility for researchers and practitioners in the field.

Motivation

Our group is motivated by a deep interest in the intersection of technology and archaeology, particularly the potential of generative AI in archaeological research. As (Cobb, 2023) suggests, 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 (But, 2024), the development of specialized generative AI tools tailored to archaeology remains underexplored. By undertaking this project, we aim to address this gap, creating a model that will be able to generate meaningful insights that can provide relevant and useful information for archaeologists in the field.

Additionally, with the recent extreme growth of interest surrounding generative AI, our group sought a project that would allow for a dynamic learning experience where we would be enabled to gain hands-on experience in AI development, acquiring practical skills that will not only enhance our technical expertise, but also position ourselves at the forefront of innovative research methodologies.

Review

Currently, in the field of archaeology, numerous applications of AI are being used. One notable area is artifact reconstruction, such as utilizing generative adversarial networks (GANs) to restore ancient Roman coins (Altaweel et al., 2024), or developing neural machine translation (NMT) models to aid in the translation of Akkadian texts from cuneiform script and transliteration into English (But, 2024). However, despite these advancements, the use of generative AI specifically remains underexplored, highlighting a significant gap in research opportunities that could further enrich archaeological research and methodologies.

Project Objective

The objective of our project is to develop a generative 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 our model, we aim to enhance the quality and reliability of information available to researchers, enabling a research environment that would empower researchers to explore and analyze complex questions more efficiently. Additionally, we will create a user-friendly graphical

user interface (GUI) to ensure usability and accessibility for researchers, allowing them to interact seamlessly with the AI model and derive insights from the generated data.

Project Methodology

As the implementation of our project will focus on leveraging generative NLP AI to enhance archaeological research, we will begin by testing various open-source LLMs to determine which best suits our objectives. This will involve evaluating models based on their ability to generate coherent, contextually relevant responses to archaeological inquiries. We will utilize Python as our primary programming language, taking advantage of its extensive libraries for natural language processing and machine learning.

To facilitate interactive experimentation and visualization, we will employ Google Colab, an accessible platform that allows for the integration of executable code and rich text in a single document. Utilizing tools such as LangChain, we can streamline the process of working with LLMs. Additionally, we plan to incorporate a vector database to efficiently manage and retrieve embeddings generated from our dataset of archaeological research papers. This combination of technologies will enable us to train and develop a robust AI model which will be able to generate accurate knowledge and information.

Evaluation

The evaluation of our generative NLP AI model will be comprehensive, focusing on several key performance metrics to ensure a robust and well-rounded testing scheme.

Beginning with accuracy and precision, by comparing the model's responses to a predefined set of questions and their corresponding expected answers, we can calculate the percentage of correct responses generated, as well as the ratio of relevant responses to a total number of responses generated through qualitative assessments and automated scoring systems.

To further evaluate the model's ability to retrieve relevant information, we will measure recall and compute the F1-score, which balances precision and recall, providing a holistic view of the model's performance. Additionally, we will test the model's generalization by applying it to a

validation dataset that was not included in the training phase, determining its ability to perform on unseen data.

Robustness will be evaluated by introducing slight variations in input data to observe how performance is affected, while interpretability will be assessed through user evaluation to determine how easily the model's outputs are understood.

The conciseness of responses will be measured by analyzing response length relative to information density, ensuring that answers both informative and succinct.

Finally, we will gauge user satisfaction through feedback sessions, including our supervisor, archaeologist Professor Cobb, with discussions after their interactions with the model to gather qualitative insights into their experiences.

Finally, we will conduct a thorough analysis for bias detection to identify any skewness in responses related to specific archaeological perspectives or datasets.

Project Schedule and Milestones

Below is our current project schedule and milestones, with approximate completion dates:

Stage	Milestone	Approximate Completion Date
Project Planning	Project scope and objectives finalized	October 1st, 2024
Data Collection	Initial data sources identified and collected	Early October, 2024
Model Set Up	Initial model selection and set up	Mid October, 2024
Model Development	First round of model testing and evaluation using collected data	End of October, 2024
Model Refinement	Adjustments and improvements based on initial testing, regarding data used, etc. Alongside this, a rudimentary GUI for user testing.	Early November, 2024

Second Evaluation	Comprehensive evaluation of model performance	Mid November, 2024
Model Refinement	Another stage of model training and development, as well as GUI improvements.	End of November, 2024
Third Evaluation	Another comprehensive evaluation of model performance	December, 2024
Final refinements and development	The last stage of model training and development.	January, 2025
Final Evaluation	Final evaluation for finetuning, last adjustments to be made here.	February, 2025
Documentation	Final project report and documentation completed	May ~ April 2025.

We hope to perform the bulk of our model training before early 2025 to account for potential delays which will inevitably occur and ensure ample time for thorough testing and refinement. Additionally, we cannot provide specific values regarding precision, F1-scores, or other performance metrics at this stage due to current uncertainties before testing and development. However, once we begin training the model, we will implement these values into our milestones to establish a clear development plan and track progress effectively.

Reference

Altaweel, M., Khelifi, A., & Zafar, M. H. (2024). Using generative AI for reconstructing cultural artifacts: Examples using Roman coins. *Journal of Computer Applications in Archaeology*, *7*(1), 301–315. https://doi.org/10.5334/jcaa.146

But, T. (2024, September 2). *Ai revolutionizes archaeology: Discoveries & challenges*. AI Revolutionizes Archaeology: Discoveries & Challenges. https://www.historica.org/blog/the-latest-ai-innovations-in-archaeology

Calugay, S. (2015, October). The destructive nature of archaeology. https://www.theposthole.org/read/article/343

Cobb, P. J. (2023, September 22). *Large language models and generative AI, oh my!: Advances in archaeological practice*. Cambridge Core. https://www.cambridge.org/core/journals/advances-in-archaeological-practice/article/large-language-models-and-generative-ai-oh-my/314BA1339E6908606B90202C0DEF266E