PROCEEDINGS OF SCIENCE

Leveraging OpenAI's LLMs and Cloud-based Learning-as-a-Service (LaaS) Solutions to Create Culturally Rich Conversational AI Chatbot: ChatLoS - A Study Using the Legacy of Slavery Dataset

Rajesh Kumar Gnanasekaran[a,*] **and Richard Marciano**[b]

[a] *University of Maryland, College Park, MD, USA*

[b] *University of Maryland, College Park, MD, USA*

E-mail: rgnanase@umd.edu, marciano@umd.edu

*Speaker

In scientific applications, integrating artificial intelligence (AI) and machine learning (ML) has revolutionized research methodologies and workflows. This study delves into an innovative application of OpenAI's Large Language Models (LLMs) in developing a conversational AI chatbot, drawing exclusively from the culturally significant Legacy of Slavery (LoS) datasets maintained by the Maryland State Archives. This initiative deviates from conventional chatbots that rely on a vast, generalized corpus for training. Instead, it focuses on harnessing the LoS datasets as the sole source for responses, thereby ensuring the authenticity and contextual relevance of the historical content. Expanding on Cloud-hosted interactive digital computer notebooks (iDCNs) to design and create a new Learning-as-a-Service (LaaS) solutions are at the heart of this research. These notebooks are designed to elucidate the methodology behind employing OpenAI's LLMs to engineer a chatbot that engages in meaningful dialogues and is also constrained to using verified data from the LoS collection. The intention is to create a chatbot that supports educational and research-focused interactions, offering users insights rooted directly in the archival material. The project also integrates LangChain agents, such as CSV agents, to empower the chatbot with data aggregation and analytical tasks capabilities, extending its functionality beyond standard conversational interfaces. A pivotal aspect of this study is the comparative analysis between the outcomes produced by the LLM-based chatbot and those obtained using traditional data analysis and visualization tools like Tableau. This comparative study is essential to assess the effectiveness and accuracy of AI-driven analysis compared to conventional data analysis methods. It aims to illuminate the potential benefits and drawbacks of employing LLMs in scientific and research settings, particularly in the context of historical and cultural data analysis. This project's cloud computing and AI convergence exemplify an innovative approach to digital humanities and archival research. The cloud-based digital notebooks serve as a model for LaaS solutions, showcasing how AI can transform the access, analysis, and dissemination of cultural and historical data. This research contributes significantly to the ongoing discourse on AI-enabled scientific workflows, offering new perspectives on applying ML and Deep Learning techniques in data-rich domains of humanities research. Through its unique use of AI, this project opens new pathways for interacting with, analyzing, and learning from historical datasets. It demonstrates the transformative potential of AI in reshaping educational and scholarly approaches to digital humanities. Another aspect of this study is to enable access to the hidden information buried in the physical archives of culturally rich dataset collections. The insights gleaned from this study are poised to influence a range of disciplines, promoting a deeper understanding of how AI can be tailored to respect and amplify the nuances of cultural and historical datasets in the digital era.

International Symposium on Grids and Clouds (ISGC2024)
24 -29 March, 2024
Academia Sinica Computing Centre (ASGC), Institute of Physics, Academia Sinica Taipei, Taiwan

1. Introduction

Studies like [1] have shown an imminent need for increased support in learning or teaching programming and using computational thinking practices in the archives and library sciences field. This is especially prominent among the students and researchers who practice in these areas today [2]. From the Archives' management perspective, there is also an overarching need for enabling wide access to culturally rich datasets like the Maryland State Archives(MSA)'s Legacy of Slavery (LoS)[1] dataset collections that are either digitally archived, born-digital, or physically archived. These archives house many historical and cultural data embedded within these datasets, which need access. From the students' and researchers' perspective, as noted by [3], [4], [5], there is a lack of open source set of course and training resources that are easy-to-use, accessible anywhere to enable the learning and understanding of the computational concepts that would be used for operating on these data sets. These resources could teach them how to work on digitally archived or born-digital data sets to enhance their capabilities, giving them opportunities to enable access to the datasets in turn.

1.1 Defining an Overall Objective

In many ways, this project is a natural extension of the study published at ISGC2021 [6], where we introduced a set of interactive digital computer notebooks or *iDCNs* to explore a novel method to facilitate collaboration and interaction in the Humanities, Arts, and Social Sciences (HASS) track between experts from multi-disciplinary fields. These iDCNs were created to teach basic computational programming to beginners by performing data exploration and data science techniques using a digitized archival dataset from the LoS collection as a case study dataset. This provided a platform for analyzing and studying culturally rich and sensitive data. Through that study in 2021, we identified the necessity to design and define an overall objective for future studies, including this one, as they belong to a similar domain and are closely related and relevant as they all use MSA's dataset collections. In this study, in addition to building up on this prior study and expanding the concept of iDCNs, we are defining an overarching project solution, termed *Learning-as-a-Service (LaaS)*[2] solutions. These solutions are cloud-hosted, coherent, and modularized digital notebooks collated as chapters of books, documented on specific subject matter, and that are open source and could be publicly accessed by the pedagogical participants and domain experts. This is achieved through the inherent capabilities of these unique digital notebooks that can be used to create plain-text instructions and explanations alongside the computer programming code. These solutions would be tailored to teach interested individuals topics from simple to advanced computing technologies through an interactive mode to enable hands-on learning and training, thereby attempting to address the lack of resources for the students and researchers in the archives and library sciences field. In summary, if we can imagine the outcomes of the study in 2021 as a set of notebook modules combined together to explain as a "Chapter," the LaaS solution we envision would be a book containing multiple such "Chapters," all being part of the same subject matter or domain. These co-related chapter modules would be packaged and hosted on the cloud

[1] http://slavery.msa.maryland.gov/html/links/ads.html

[2] https://skillup.tech/learning-as-a-service-finding-your-laas-partner/

for free and available at *CASES*[3] under a single platform. For chapters that use datasets to explain the computational techniques, we also aim to enhance awareness, access, and knowledge of the real-world culturally rich archival dataset collections, such as on LoS, by using them for case studies in these LaaS solutions. This approach benefits the project end users and the archivists. Through these learning solutions, collections that are not commonly accessed are digitally prepared, analyzed, explored, and visualized, and the results are made available to the students and researchers.

1.2 Project Objective

Our prior study in 2021 [6] already introduced LaaS solutions through its first chapter on one of MSA's collections, the Certificates of Freedom (CoF), as a case study dataset to perform a step-by-step data science analysis. This exploratory research project focuses on a new LaaS chapter with step-by-step instructions and the programming code to perform an end-to-end project on a timely topic by accessing a new dataset from the LoS, the Domestic Traffic Ads (DTA) dataset. This is another unique, culturally rich collection from the MSA. In addition to achieving the overall objective being targeted, the objective for this project specifically would be to focus on developing a unique conversational artificial intelligence (AI) powered chatbot to converse against the unique DTA dataset. The DTA dataset would be used as a digital dataset source, combining a list of metadata features digitally transcribed by the MSA Office and an augmented feature with the optical character recognized (OCR-ed) text output of the individual scanned newspaper advertisements. The motivation for this stems from the introduction of *ChatGPT*, a conversational AI chatbot underpinned by a Generative Pre-trained Transformer (GPT) model developed by *OpenAI* [4] in late 2023. GPT models are a subset of advanced Large Language Models (LLMs)[5], trained extensively on diverse textual data, including books and articles. This training equips GPT models with the capability to generate natural-sounding, human-like text responses, rendering them highly valuable for tasks such as answering questions and summarizing context. Since this introduction, many AI researchers have implemented GPT-powered chatbots for various use cases.

In this study, we created a second chapter of the LaaS solution using the DTA dataset to create a chatbot and test OpenAI's GPT capabilities. Going forward, this unique AI-powered chatbot using the LoS DTA dataset will be called "ChatLoS (Chatbot for the LoS project)." We set ourselves on two tasks to test the capabilities of ChatLoS. Firstly, can ChatLoS limit its responses to a specific knowledge domain, such that the answers are only determined from the DTA dataset source and not from the pre-trained commonly available knowledge the GPT model has been trained on? Secondly, does ChatLoS have the capabilities to respond to DTA dataset-specific data aggregation analytical questions in natural language statements without knowing the details of the metadata or columns? If so, how do the responses compare to results from traditional data analysis using a commonly used tool such as *Tableau*[6]? What metrics can be used to compare the results of both analyses? The reasons to test ChatLoS on the above two capabilities specifically are that: 1. By limiting ChatLoS's domain knowledge to only MSA-hosted dataset sources, this chatbot could become a valuable resource for MSA to promote access to their collections publicly on their

[3]https://cases.umd.edu/
[4]https://chat.openai.com/
[5]https://aws.amazon.com/what-is/large-language-model/
[6]https://www.tableau.com/

website, and for a community member, this would ensure that the information they receive from this bot is grounded in true historical data hosted by MSA and not from an unverified source on the internet. This could potentially increase or create "Trust" in these AI-based solutions. 2. If an interactive conversational chatbot like ChatLoS can accurately and appropriately respond to data aggregation analytical questions from users in natural language format with no background knowledge of the dataset's metadata, this would be an excellent practical advantage and benefit for agencies like MSA and their patrons or users to use this in place of expensive license-based tools like Tableau, Microsoft's *PowerBI* [7], etc. This use case can further validate the authenticity of the data grounded in the limited MSA domain and enhance access to the underlying archival data. For the remainder of this paper, the primary focus would be to showcase our efforts in addressing the two tasks above as part of this project's objective. The rest of the paper is divided into the following sections: Background, Research Methodology & Questions, Results & Findings, Future Work, and Conclusion.

2. Background

Since its initiation in the fall of 2001, the MSA's project, officially named the Study of the LoS in 2005, has focused on uncovering the narratives of individuals who resisted enslavement in Maryland, USA. Utilizing various sources like court records, laws, newspapers, and maps, the MSA staff aimed to highlight the unrecognized heroes of slave resistance and create comprehensive case studies. The project has grown significantly with the help of several grants, involving over 100 professionals and volunteers, including interns, to transcribe and digitize physical records. This endeavor has culminated in a robust online database containing over 400,000 records, including domestic traffic ads (DTA), runaway advertisements, certificate of freedom records, and federal census data, providing a detailed view of anti-slavery movements across fourteen of the state's antebellum counties. Several prior works have been performed on these dataset collections for various purposes, as detailed in [7], [8]. For this project, we will be using the DTA dataset.

2.1 Domestic Traffic Ads Dataset

The DTA dataset comprises records that contain details on the domestic traffic advertisements placed in public newspapers for the interstate and intrastate trade of enslaved men, women, women, and children. This dataset comprises digitally transcribed data of advertisements for enslaved individuals placed in Maryland newspapers over a 40-year period, from March 3, 1824, to April 30, 1864. It includes crucial metadata fields such as advertisement date, enslaved person's name, slave owner's name, source publishing this ad, county, location, enslaved age, gender, number of people being sold, terms of sale, and specified skills. Originally, about 2100+ advertisement records were found in MSA's digital database. After performing preliminary data exploration and a cleaning process, for the purpose of this exploratory case study, we created a test dataset based on a subset of the records covering the first 10 years of the collection (from 1824 to 1834), which resulted in over a third (35%) of the dataset or 764 records. In addition to the digitally transcribed metadata features on this dataset, we realized that the project's real benefit could be vastly enhanced if the

[7]https://www.microsoft.com/en-us/power-platform/products/power-bi/


Valuable Negroes

AT PRIVATE SALE.

The subscriber will sell at private sale a Family of as Valuable Negroes as any on the Eastern Shore, they will not be sold out of the county—viz. 1 woman & her 2 children, 1 man 38 years old, 1 do 20 years, 1 do. 23 years, 1 boy 14 years, 1 do. 6 years, 2 do. 7 years, 1 woman & her 2 children one 2 and the other 5 years, 1 woman and her child 3 years, 2 girls 4 years, 1 do. 5 years, 1 do 16 years, they are all young and handsome, and will be sold to good masters very low.

JOHN DONOVAN.

Dec. 31 3w

Figure 1: Sample 1 of the DTA scanned ad

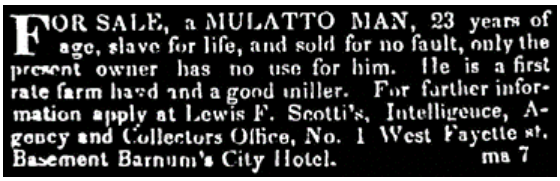

FOR SALE, a MULATTO MAN, 23 years of age, slave for life, and sold for no fault, only the present owner has no use for him. He is a first rate farm hand and a good miller. For farther information apply at Lewis F. Scotti's, Intelligence, Agency and Collectors Office, No. 1 West Fayette st. Basement Barnum's City Hotel. ma 7

Figure 2: Sample 2 of the DTA scanned ad

text from the advertisements themselves was available to be fed as input to the GPT model. To achieve this, we used an OCR tool *ABBYYFineReader*[8] to extract the text for each of these 764 ads, and the entire dataset was augmented with this new feature. Figure 1 and 2 show a few examples of scanned images of the DTA ads.

2.2 Related Work

Since the introduction of GPTs, several cases of unique and specific GPTs have been created and are available as open source resources as well here[9]. With respect to the archives field, a recent example is the open-source WARC-GPT[10] project from Harvard's Library Innovation Lab. It was developed to interrogate web archiving content in the form of WARC files (a file format that is a revision and generalization of the ARC format used by the Internet Archive to store information blocks harvested by web crawlers) by asking specific questions in natural language rather than relying on keyword searches and metadata filters. Generative Artificial Intelligence (Gen AI) represents an innovative frontier in machine learning, providing a transformative tool for various sectors, including libraries and archives [9]. As the Gen AI algorithms can create new data instances resembling the original data, harnessing patterns within the training dataset [10], they capture the probabilistic distribution of the data, generating new samples from this learned distribution. This trait of Gen AI offers remarkable implications for archival work, as it can create digital artifacts representative of the original resources. Expanding on the concept of Gen AI, LLMs such as GPT-3

[8]https://www.abbyy.com/ocr-sdk/

[9]https://www.whatplugin.ai/

[10]https://lil.law.harvard.edu/blog/2024/02/12/warc-gpt-an-open-source-tool-for-exploring-web-archives-with-ai/

and GPT-4, developed by OpenAI, represent a paradigm shift in text generation and analysis [11]. LLMs are pre-trained on extensive corpora and can generate contextually coherent text, answer questions, and even summarize complex documents. Importantly for the archival sector, LLMs offer potent tools to navigate and analyze digital datasets. Given their ability to understand and generate human-like text, they can decipher metadata, make sense of contextual information, and thus play a crucial role in cataloging, cross-referencing, and efficiently retrieving information. This project attempts to test the Gen AI and specifically LLM capabilities in understanding the contextual data available in the culturally rich dataset like DTA and to evaluate if these models can effectively search, curate, analyze, and prepare appropriate and contextually aware responses to the users. They are also evaluated on their capabilities in performing data aggregation analysis and if the results could be compared to traditional data analysis results.

With respect to the creation of the LaaS chapter module, in our previous work [6], the chapter modules were created as iDCNs using an open-source cloud-based web application tool called *Jupyter Notebooks* (JNs),[11] with the computational programming language *Python.*[12] Several studies have documented using JNs as a teaching tool in higher education, such as in [12], [13], [14], and in [15] which uses JNs in the Archives domain, a non-STEM background. In the prior work, we had created a set of iDCNs by combining individual JNs together through hyperlink redirections; this is because JNs are a single document format that contains both executable code (such as Python) and rich text elements (paragraphs, equations, figures, links, etc.). JNs are widely used in data science, scientific computing, and educational contexts to perform real-time data analysis, machine learning, statistical modeling, and visualizations. They are interactive, meaning that code can be executed in blocks (or cells), and the output can be seen immediately beneath the code cell. However, in this project, we are expanding upon the iDCN concept to create tailor-made LaaS solutions using a new tool called *Jupyter Books* or JBs[13]. JBs are an extension of the iDCN concept that allows building a collection of iDCNs into a more structured, book-like format. It essentially lets you create an online book. It supports rich content organization using sections, chapters, and sub-chapters, making it ideal for creating educational materials, technical documentation, and more comprehensive reporting projects where a higher level of organization and presentation is beneficial. JBs can include features like a table of contents, interactive content, citations, and cross-references. They can also be hosted online and freely as open source resources on Github as Github Pages[14] providing easy access to interactive resources. In summary, while a JN is great for interactive, exploratory analysis and documentation on a smaller scale, JBs are better suited for organizing multiple notebooks and supporting materials into a cohesive, structured, and navigable format suitable for larger projects or publications. This is exactly one of the goals of the overall objective, which is to create coherent LaaS solutions. For this project, we have created the chapter modules using JBs, Figure.6 and 7 show a sample of the chapters as they look before they are uploaded to its own Github pages repository upon completion, after which they would be published to a central repository of LaaS chapters at a special section under CASES[15].

[11] Jupyter Notebook - https://jupyter-notebook.readthedocs.io/en/stable/notebook.html.
[12] Python Documentation - https://www.Python.org/doc/.
[13] https://jupyterbook.org/en/stable/start/overview.html
[14] https://pages.github.com/
[15] https://cases.umd.edu/

3. Research Methodology & Questions

We have chosen the exploratory case study research methodology for this project using the DTA dataset as the case study dataset. To address this project's specific objectives mentioned above, we formulated the following research questions and have tried to answer them accordingly:

- RQ1: Is it possible to utilize Large Language Models, such as OpenAI's GPT models, to develop a context-limited, domain-specific conversational chatbot capable of exploring data from high cultural context datasets through natural language conversations and unveiling previously undiscovered relationships and patterns within this data?
- (RQ2) Is it possible to design and develop an interactive chatbot using the GPT models that can perform data aggregation analysis from natural language questions on the underlying high cultural dataset? If so, can we compare the results of GPT models with traditional, non-AI exploratory data analysis models based on pre-defined metrics?

4. Approach

To address the RQs, we designed and developed two types of GPT-powered conversational chatbots.

4.1 (RQ1) ChatLoS using Retrieval Augmented Generation Technique

To answer RQ1, we implemented a widely used technique called Retrieval Augmented Generation or RAG[16]. This technique allows us to work with the pre-trained LLM on limiting what information it uses to provide a response to a question. This method combines the capabilities of the LLM with a dedicated retrieval system to enhance the model's ability to generate accurate and contextually relevant responses grounded on the data available from this reference system, in this case, a database containing the combined DTA dataset augmented with the OCR text. We must follow the steps shown in Figure.3 to understand how this works. Here's a simple breakdown of the image using the DTA dataset as an example:

1. Query: Imagine an end user is using ChatLoS to ask a question about the history or specific details found in the DTA dataset.
2. Generation Embedding: The query is first processed to understand its context and intent. This involves converting the text of the query into a numerical form that a computer can understand (embeddings). These embeddings are then used to retrieve relevant information from a database.
3. Retrieval System - Chroma: The system named 'Chroma' in the image acts as the retrieval component. It contains a database of documents—in this case, it would contain digitized and processed text from the combined and augmented DTA dataset already provided as input earlier. The embeddings from the query are used to find and retrieve the most relevant documents or data entries from this database.

[16]https://heidloff.net/article/retrieval-augmented-generation-chroma-langchain/

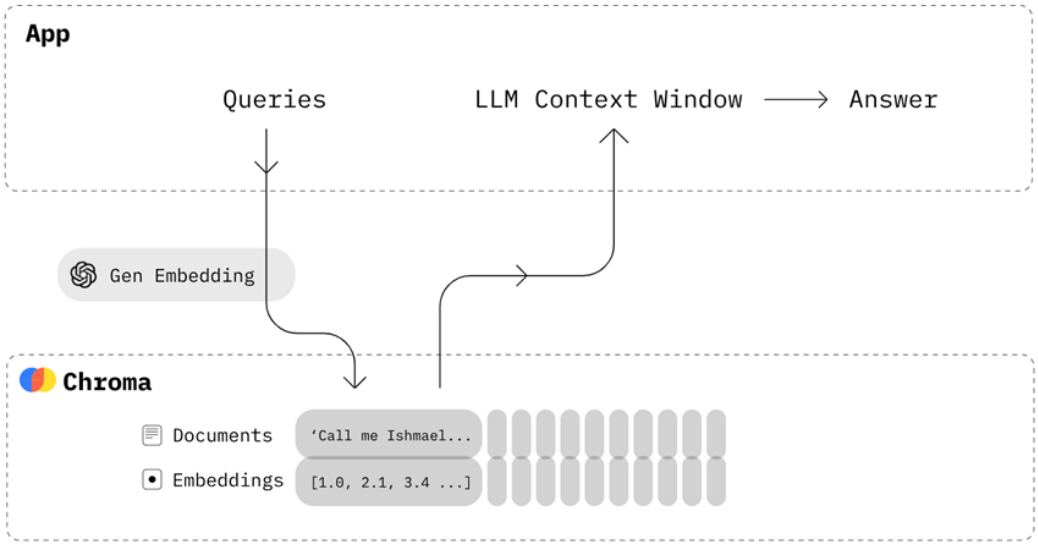


Figure 3: Retrieval Augmented Generation[17]

4. LLM Context Window: The retrieved documents are combined with the original query to form a rich context window. This context window provides a comprehensive background for the LLM to generate an informed response.

5. Answer: The LLM uses the enriched context to produce an answer that is not only based on its pre-trained general knowledge but is also specifically informed by the historical data from the retrieved documents. This way, the answer is more accurate, detailed, and contextually relevant to the specific query about the DTA dataset.

In practical terms, using the augmented DTA dataset, this RAG setup could help users rediscover intricate details and stories of individuals advertised in these ads by pulling exact historical data that matches their queries, thereby providing richer and more insightful responses. The results from these ChatLoS-type conversations are discussed in the following section.

4.2 (RQ2) ChatLoS using Langchain's CSV Agents for Data Aggregation Analysis

To address RQ2, we performed a unique implementation, which is not a usual utilization of the LLMs. However, in this step, the aim is to use a unique function named the 'create_csv_agent'[18] which is part of the Langchain Python module, a framework designed to develop applications powered by language models, with a focus on being data-aware and agentic. It is used to generate an agent that can interact with CSV files and is primarily designed for question-answering applications. By chaining this function with the OpenAI's GPT model, an AI chatbot could be created that takes natural language data aggregation questions, identifies the correct data columns to query the structured data such as a CSV, and responds with accurate results. A simple example of a

[17]https://heidloff.net/article/retrieval-augmented-generation-chroma-langchain/
[18]https://python.langchain.com/docs/integrations/toolkits/csv/

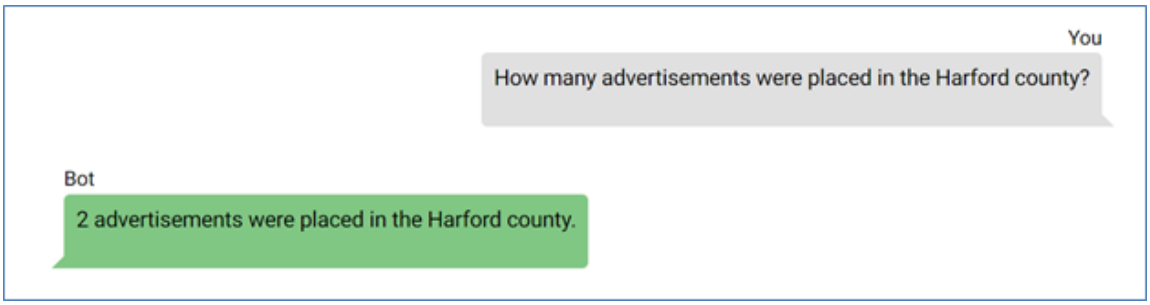


Figure 4: ChatLoS Agent Type: Data Aggregation Question example & answer

chat conversation is shown in Figure.4; let's understand how ChatLoS aggregates the number of advertisements placed in a specific county named "Harford" in Maryland. Figure.5 shows the terminal printed output of the chatbot model that runs the ChatLoS. For an input question by the user, the CSV agent reframes the question to a "thought prompt" and feeds that into an appropriate format to the underlying GPT model.

1. Thought: The reframed prompt from the question asked by the user to ChatLoS.

2. Action: The action taken by the agent to resolve the prompt is specified as 'python_repl_ast'. This refers to executing a specific Python code in a Python REPL (Read-Eval-Print Loop) environment or any interactive Python interpreter.

3. Action Input: Based on the "Thought," the agent finds the best possible columns to perform aggregation queries in Python code from the imported CSV file, in this case, the DTA file. A Python code is then executed, for this example, df[df["County"]=="Harford"].shape[0]. This code filters the data frame df (Python's imported CSV dataset) based on the condition of ["County"]=="Harford." It selects rows where the "County" column value equals "Harford." The shape[0] part returns the number of rows in the resulting filtered data frame.

4. Observation: The result of executing the code is provided as an observation, which states that the output is 2. This suggests that after applying the filter, there are two rows in the dataframe where the county is "Harford".

The significance of this approach is that ChatLoS converts natural language words into syntactical queries used to perform filtering functions on the underlying data. A major benefit of this is that the users do not have to know the columns names of the underlying dataset they are querying on, instead, they ask intuitive questions as "prompts" on their understanding of the functional knowledge or the contextual knowledge about the dataset and the chatbot responds accordingly. To address the second part of RQ2, a new Tableau workbook was created using the DTA dataset, and each individual sheet was created to perform data aggregation analysis and visualization accordingly. The results of the ChatLoS responses and the comparison between it and Tableau's non-AI traditional data analysis results are performed and discussed in the next section.

Figure 5: ChatLoS - A sample log of the question call using LangChain's CSVAgent

5. Results & Findings

This section details the results obtained from the two types of ChatLoS chatbots to address RQ1 and RQ2. This section is divided into subsections to discuss the results and findings of these two chatbots.

5.1 ChatLoS RAG Type Test Results

The goal of this step is to create a chatbot that limits its knowledge to the domain information that was passed through the DTA dataset and it shouldn't hallucinate or go outside of the context. In test case 1, as shown in Figure.8, the test question asks if there are any mentions of a person specifically named "Jacob" and if any ads have been placed to sell or buy for this name. It also asks another question to summarize the entire advertisement content for the user. This could be a typical question expected from an end user like an archives patron or a public community member interested in this culturally rich dataset content. For example, they might want to identify somehow one of the relatives or one of their grandfathers or grandmothers if they were listed for sale in these advertisements, let's say if the end users are descendants of these enslaved people. From the response, it appears that the ChatLoS bot found an ad mentioning this person's name as "Jacob," it also provided a summary of this ad. Now the question is, was there a matching ad in the original dataset passed to the bot? This appears to be true, as shown in Figure.9 a screenshot of the data passed to the LLM, and the ad details match the response summary. For testing purposes, we cross-verified the ad with the scanned copy of the physical ad, as shown in Figure.10. From this test case, it could be seen that ChatLoS passed this scenario. For test case 2, we evaluated whether the ChatLoS can truly stay within the context and doesn't exhibit an infamous phenomenon called "Hallucinations" typical of the LLMs. In Figure.11, the ChatLoS was asked questions that are of common knowledge about prominent people, and it replied back saying that it doesn't know who they are. This indicates that the ChatLoS has knowledge grounded in the DTA dataset, as there were no references to these people in the DTA dataset content. We consider that ChatLoS has passed this test case as well. For test case 3, we wanted to drill the chatbot further to understand if OpenAI's GPT LLM has any guardrails implemented so as not to provide hateful or inappropriate responses. Figure.12 shows a response that asks it to perform an action as a response that would be considered with a racial connotation and hateful. The ChatLoS, however, provided a response

that we consider appropriate for the question asked. Though the response doesn't specify why it wouldn't perform such an action of writing a nasty poem on the "negro boy" mentioned in the ad. Still, we believe that ChatLoS has passed this test. Based on these tests, we could see that ChatLoS can limit its responses to the content provided through the RAG mechanism and doesn't use its pre-trained knowledge to respond to questions related to the DTA dataset.

5.2 ChatLoS Agent Type Test Results & Metrics Comparison with Tableau results

To evaluate ChatLoS of Agent Type, the bot was asked a set of seven questions focused on performing data aggregation on the DTA dataset to test its capabilities. The same questions were performed as data analysis and visualization actions on Tableau from a desktop version. To compare the results of these tools, we decided to use nine metrics and captured the results on which tools performed better for each metric being compared. Any additional information was captured under the Comments column. These results are shown in Figure.13. Individual test case results are captured below with a Pass or Fail against the tool that performed better or worse for each test case.

5.2.1 Question 1: How many ads are there totally?

ChatLoS's Performance: *Pass* - Refer Figure.14 ChatLoS passed this test by providing an accurate response to this question with a number equal to the total of ads in this dataset, i.e., 764. Due to its interactive and conversational nature, it was able to provide the answer as a well-framed sentence.

Tableau's Performance: *Fail* - Refer Figure.15 Tableau doesn't have a function or view that aggregates a specific column of data as a single answer. The figure shown lists the counts by county, as we chose this field for visualization; however, one has to perform additional operations to show a "Total" column to see the total counts. Due to this intricate setup, we marked that Tableau "failed" or performed worse than ChatLoS.

5.2.2 Question 2: Can you create a table of the number of ads placed by each county?

ChatLoS's Performance: *Pass* - Refer Figure.14 ChatLoS passed this test by providing an accurate response to this question with a table of counts of ads by County as per the DTA dataset passed as input. However, it should be noted that the table did not get rendered as good as Tableau as rendered.

Tableau's Performance: *Pass* - Refer Figure.15 and 16 Tableau passed this test by providing the correct counts by County; one could also use Tableau to perform several variations of this data visualization, as shown in the Maryland state geo map as an additional data view.

5.2.3 Question 3: Can you report ad counts for both public auctions and public sales?

ChatLoS's Performance: *Pass* - Refer Figure.17 ChatLoS also passed this aggregation question with the correct counts matching with the DTA dataset. It should be noted that ChatLoS had an edge over Tableau with this response because it provided additional clarification on why it chose to do what it did, which wasn't possible with Tableau due to its non-interactive nature.

Tableau's Performance: *Pass* - Refer Figure.18 Tableau passed this test after adding the corresponding field to perform a data aggregation on its counts by Sale Disposition.

5.2.4 Question 4: Were any ads placed on Christmas Day, if so, how many and what are the years?

ChatLoS's Performance: *Pass* - Refer Figure.19 and 20 ChatLoS performed like a star on this test case and easily passed it. This test case is a classic example of the power of performing data aggregation analysis using an AI tool like ChatLoS. These questions have the power to unearth critical insights from the end users. By keeping the end users, who could be non-technical and agnostic of the details of the dataset's metadata, one could easily query the underlying dataset in natural language statements to receive responses, thereby widening the scope of performing such analysis.

Tableau's Performance: *Fail* - Refer Figure.21 Tableau failed this test as there is no easy way to get this date without spending too much time creating wildcard filters or writing expression functions on the Ad_date field.

5.2.5 Question 5: How many cooks were on sale?

ChatLoS's Performance: *Pass* - Refer Figure.22 and 23 ChatLoS performed exceptionally well with this test case, with the added advantage of asking users for interactive follow-ups on what they were looking for. Upon taking the follow-up response from the user, the bot responded with a well-rounded answer.

Tableau's Performance: *Fail* - Refer Figure.24 Tableau failed this as without doing some special operations in parsing the Skills_specified field due to the multi-attribute values, this operation would not be easily done. Also, there wouldn't have been any interactive feedback to the user on what they really needed vs what was shown, just like the ChatLoS did.

5.2.6 Question 6: Can you create a bar chart showing the number of ads placed by each ad source and the county?

ChatLoS's Performance: *Fail* - Refer Figure.25 and 26 ChatLoS failed this test as it couldn't produce a good-looking bar chart visualization, and even with the one it produced, the axis vs data mappings were incorrect.

Tableau's Performance: *Pass* - Refer Figure.27 Tableau is known to create nicer-looking visualizations, so it passed this test.

5.2.7 Question 7: How many ads were placed for individuals with ages less than 10?

ChatLoS's Performance: *Pass* - Refer Figure.28 and 29 ChatLoS passed this test, as can be seen in the DTA dataset screenshot and its response, which is shown to match.

Tableau's Performance: *Fail* - Refer Figure.30 Tableau fails this test as without doing additional operations in filtering by the Age field; it would not be possible to get the counts below 10 of the enslaved people being sold. From the test results, it's clear that ChatLoS passed more tests (6) than Tableau (2) results. This is primarily due to ChatLoS's interactive and follow-up nature through a conversational interface with the end user. This isn't possible with Tableau, as the user has to perform actions to get the desired results as discrete steps. ChatLoS behaved smartly in a few cases without additional clarification, and all of the responses were accurate, especially without the

end user specifying any column names. This wouldn't be possible in Tableau as the users have to choose the correct columns for analysis themselves. ChatLoS reduced the number of steps one has to perform to select the appropriate columns for analysis compared to Tableau. It went a step above with test cases 4 and 5, where it took additional decisions to convert keywords from the question to map against the appropriate columns and respond with a detailed answer. From the metrics comparison perspective, as seen in Figure.13, of the nine metrics compared, both tools fared well for three of them and tied at three each for both Tableau and ChatLoS. These metrics were derived based on the typical performance factors while operating a data aggregation and visualization tool. ChatLoS's speed of response was considerably slower than Tableau's, which is understandable because it has to do additional operations to run agentic transactions to provide a tailored response. As a text conversation tool, the clarity of presentation of the ChatLoS was not great when compared with Tableau; however, with advancements in the LLM world, this feature will soon be available in future LLMs. There is currently no direct API integration available with the LLMs; however, using Tools [19], one could implement code to interact with the APIs that are open to be used by integrating with the LangChain framework. This was out of the scope of work for this project and, hence, not explored. From the above results and findings section, it's clear that Gen AI-powered LLM-based chatbots exhibit the capabilities tested to answer RQ1 and RQ2. To answer RQ1, the ChatLoS RAG-type bot was able to limit its context to the DTA dataset content and also exhibited inherent guardrails against inappropriate or abusive behavior. With respect to the data aggregation analysis use case, the ChatLoS agent type performed on par, if not better than Tableau, on some metrics and test cases, showing great potential for improvement and use on a broader scale.

6. Future Work

With a lot of positive results obtained through this exploratory case study, one purely missing aspect is the user or human feedback. To improve the chatbots for the better, we would like to involve the user communities in testing and providing feedback on their thoughts on this unique tool. As a future work, we would like to draft a research question such as "*How can we incorporate socio-technical considerations to promote trustworthiness and mitigate potential bias arising from the use of GPT models with library and archival collections?*" and work towards involving community users and receive feedback from them through regular workshops. We would use the feedback to implement changes, solutions, and additional guardrails to avoid issues related to ethical violations. Despite the significant promises of Generative AI and LLMs, it is critical to address ethical considerations, especially the risk of generating synthetic data that may mislead or misinform users, as noted by [16]. Therefore, while generative AI holds transformative potential for archival work, careful application is paramount to ensure its benefits are realized ethically and responsibly. LLMs, as seen above, are a powerful tool that could be leveraged to analyze historical data; however, caution should be exercised as LLMs may inadvertently introduce biases into their outputs, reflecting the biases in their training data [17]. Therefore, while the potential of LLMs in managing and making sense of digital archives is immense, their deployment needs to be monitored for ethical application and potential bias in their outputs.

[19] https://python.langchain.com/docs/modules/tools/

7. Conclusion

The innovative exploratory case study presented in this paper, leveraging OpenAI's Large Language Models (LLMs) and Learning-as-a-Service (LaaS) solutions to create culturally rich conversational AI chatbot, the ChatLoS, using the Legacy of Slavery dataset, demonstrates significant strides in the integration of advanced artificial intelligence within the realm of archival research. Our project not only enriches user interaction through a contextually aware chatbot but also enhances the accessibility and analytical capabilities regarding historically significant data. By comparing the AI-driven conversational outputs with traditional data analysis tools like Tableau, we have provided empirical insights into the effectiveness and potential of AI in facilitating more profound and accessible engagements with cultural archives. As we refine these technologies, the potential for expanding such models to other domains of historical and cultural significance holds promising implications for the future of educational technology and archival accessibility. This study stands as a testament to the transformative power of combining AI with cultural heritage datasets, paving the way for future innovations in the field.

References

[1] E. Goudarouli, A. Sexton and J. Sheridan, *The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK*, *Philosophy & Technology* **32** (2019) 173.

[2] W. Underwood, D. Weintrop, M. Kurtz and R. Marciano, *Introducing Computational Thinking into Archival Science Education*, pp. 2761–2765, 2018, DOI.

[3] J. DeVaney, G. Shimshon, M. Rascoff and J. Maggioncalda, "Higher ed needs a long-term plan for virtual learning." https://hbr.org/2020/05/higher-ed-needs-a-long-term-plan-for-virtual-learning, Feb, 2021.

[4] S. Dunn, "What versus how: Teaching digital humanities after covid-19." https://blogs.kcl.ac.uk/ddh/2020/05/04/teaching-digital-humanities-after-covid19/, Nov, 2020.

[5] S. Gallagher and J. Palmer, "The pandemic pushed universities online. the change was long overdue." https://hbr.org/2020/09/the-pandemic-pushed-universities-online-the-change-was-long-overdue, Sep, 2020.

[6] R.K. Gnanasekaran and R. Marciano, *Piloting Data Science Learning Platforms through the Development of Cloud-based interactive Digital Computational Notebooks*, *PoS* **ISGC2021** (2021) 018.

[7] A. Inbasekaran, R.K. Gnanasekaran and R. Marciano, *Using transfer learning to contextually optimize optical character recognition (OCR) output and perform new feature extraction on a digitized cultural and historical dataset*, in *2021 IEEE International Conference on Big Data (Big Data)*, pp. 2224–2230, DOI.

[8] L.A. Perine, R.K. Gnanasekaran, P. Nicholas, A. Hill and R. Marciano, *Computational treatments to recover erased heritage: A legacy of slavery case study (ct-los)*, in *2020 IEEE International Conference on Big Data (Big Data)*, pp. 1894–1903, 2020, DOI.

[9] I.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair et al., *Generative adversarial networks*, 2014.

[10] D.P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2022.

[11] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal et al., *Language models are few-shot learners*, 2020.

[12] A. Davies, F. Hooley, P. Causey-Freeman, I. Eleftheriou and G. Moulton, *Using interactive digital notebooks for bioscience and informatics education*, *PLOS Computational Biology* **16** (2020) e1008326.

[13] J. Reades, *Teaching on Jupyter: Using notebooks to accelerate learning and curriculum development*, vol. 7 (01, 2020), DOI: 10.18335/region.v7i1.282.

[14] A. Zúñiga-López and C. Avilés-Cruz, *Digital signal processing course on Jupyter–Python Notebook for electronics undergraduates*, *Computer Applications in Engineering Education* **28** (2020) 1045.

[15] M. Wigham, L. Melgar and R. Ordelman, *Jupyter notebooks for generous archive interfaces*, in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2766–2774, 2018, DOI.

[16] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield et al., *Toward trustworthy ai development: Mechanisms for supporting verifiable claims*, 2020.

[17] E.M. Bender, T. Gebru, A. McMillan-Major and S. Shmitchell, *On the dangers of stochastic parrots: Can language models be too big?* , in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pp. 610–623, Association for Computing Machinery, DOI.

8. Appendix Section: Additional Figures supporting the main paper

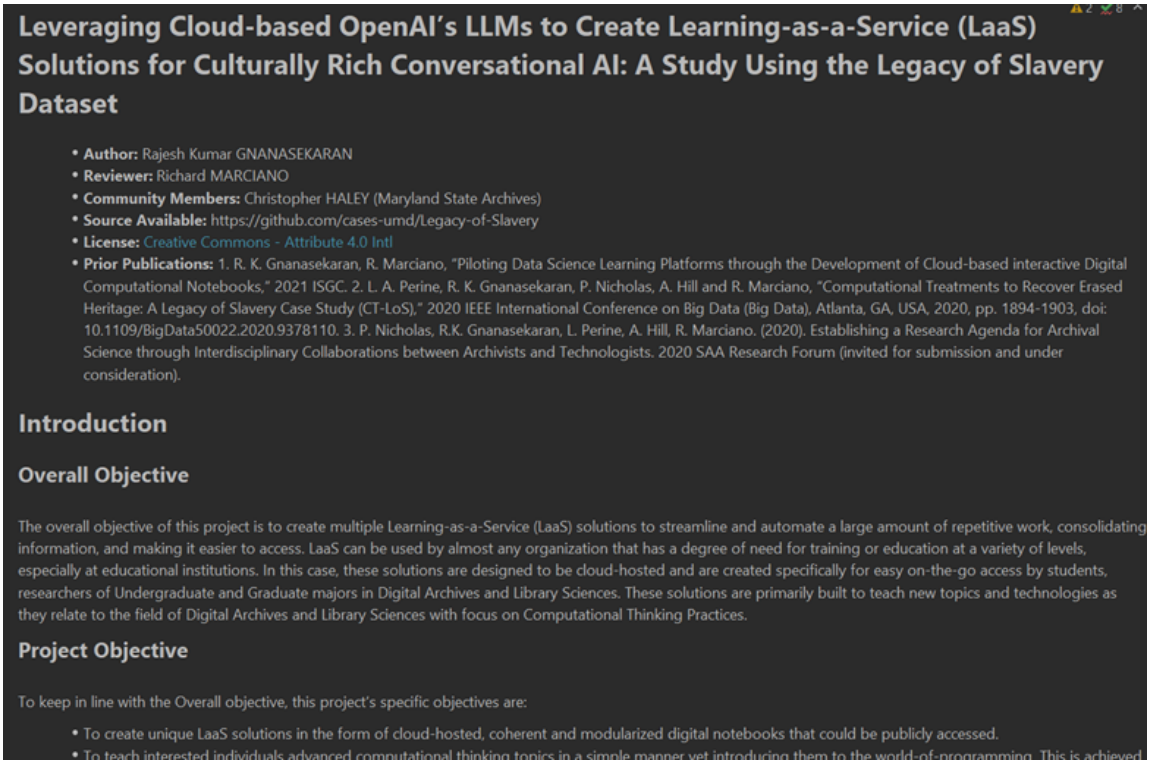


Figure 6: Index page of the Project's Jupyter Book Chapter

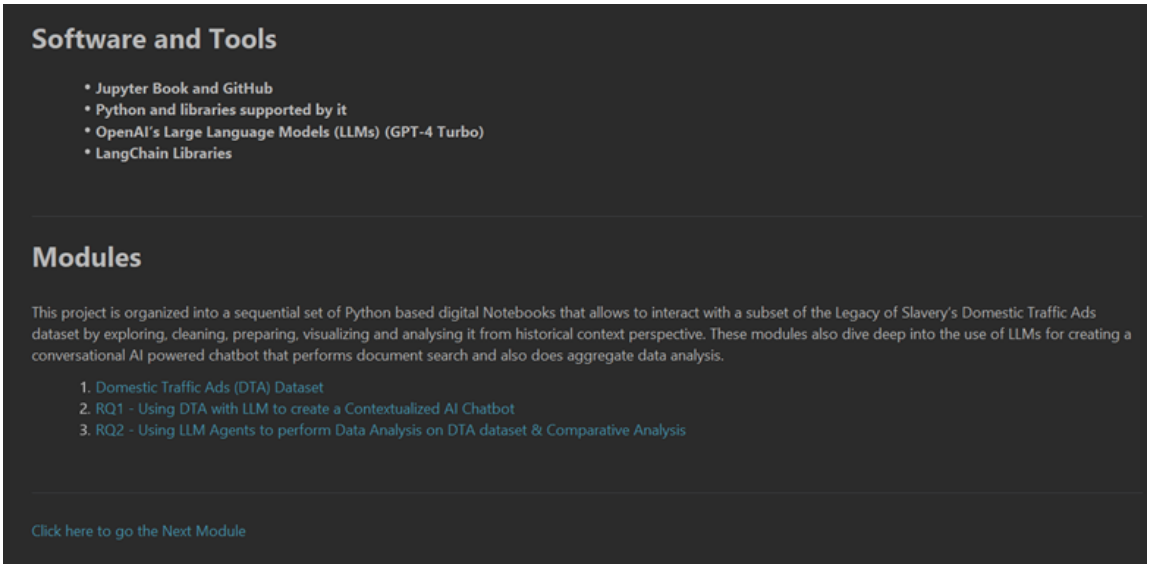


Figure 7: Page of the Project's Jupyter Book Chapter showing links to other chapter modules

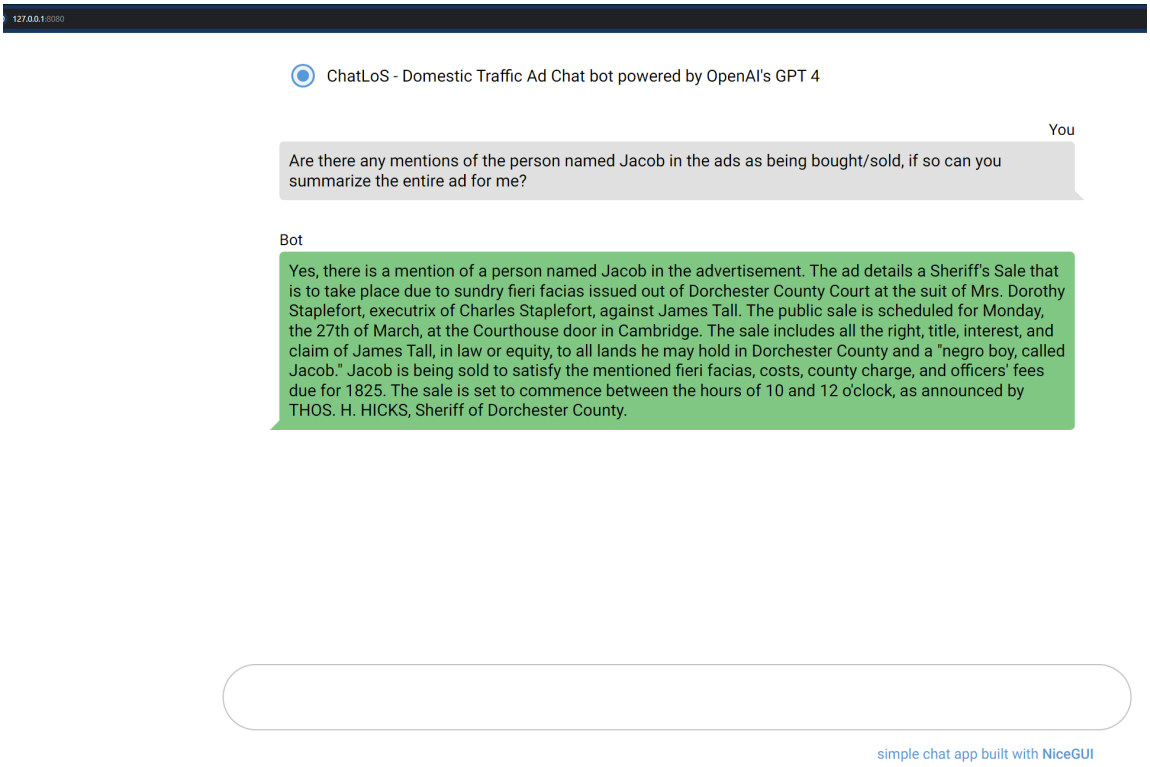


Figure 8: ChatLoS RAG: Test Case 1 - Conversation response

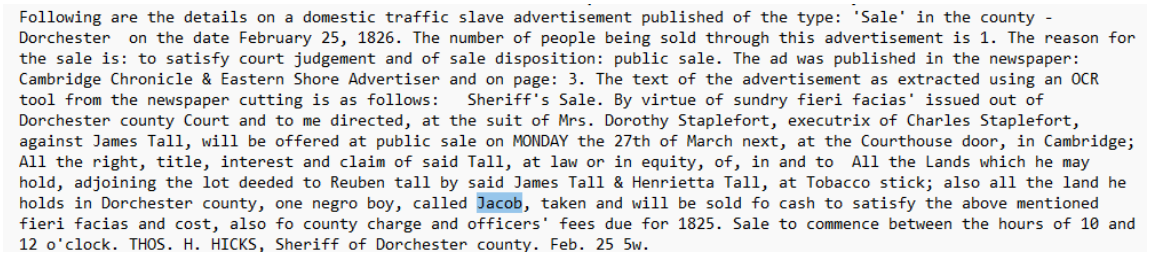
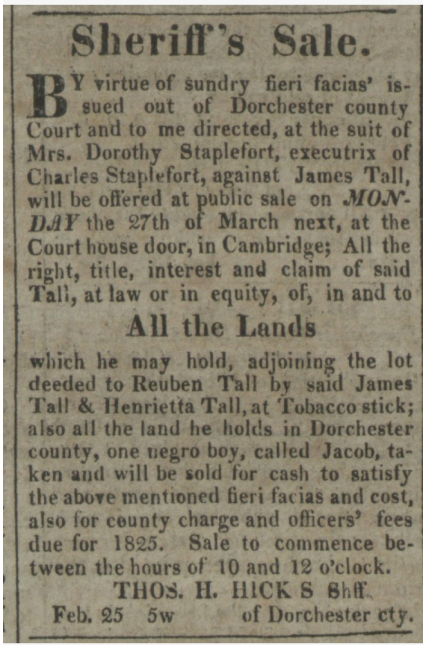


Figure 9: ChatLoS RAG: Test Case 1 - Screenshot of the augmented DTA dataset text of the exact ad provided as response



Sheriff's Sale.

BY virtue of sundry fieri facias' issued out of Dorchester county Court and to me directed, at the suit of Mrs. Dorothy Staplefort, executrix of Charles Staplefort, against James Tall, will be offered at public sale on *MONDAY* the 27th of March next, at the Court house door, in Cambridge; All the right, title, interest and claim of said Tall, at law or in equity, of, in and to

All the Lands

which he may hold, adjoining the lot deeded to Reuben Tall by said James Tall & Henrietta Tall, at Tobacco stick; also all the land he holds in Dorchester county, one negro boy, called Jacob, taken and will be sold for cash to satisfy the above mentioned fieri facias and cost, also for county charge and officers' fees due for 1825. Sale to commence between the hours of 10 and 12 o'clock.

THOS. H. HICKS Shff.

Feb. 25 5w of Dorchester cty.

Figure 10: ChatLoS RAG: Test Case 1 - Scanned copy of the physical ad cutting of the exact ad provided as response

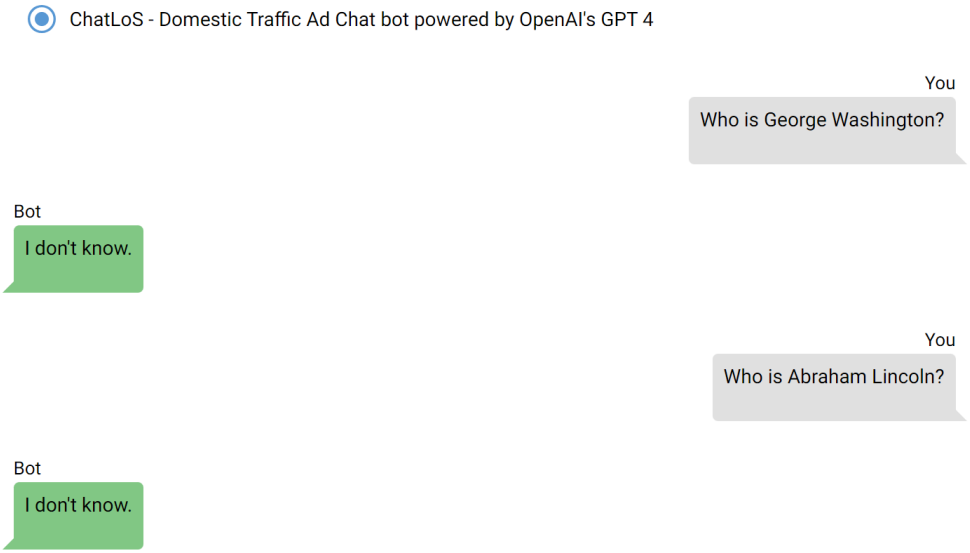


Figure 11: ChatLoS RAG: Test Case 2 - Conversation response

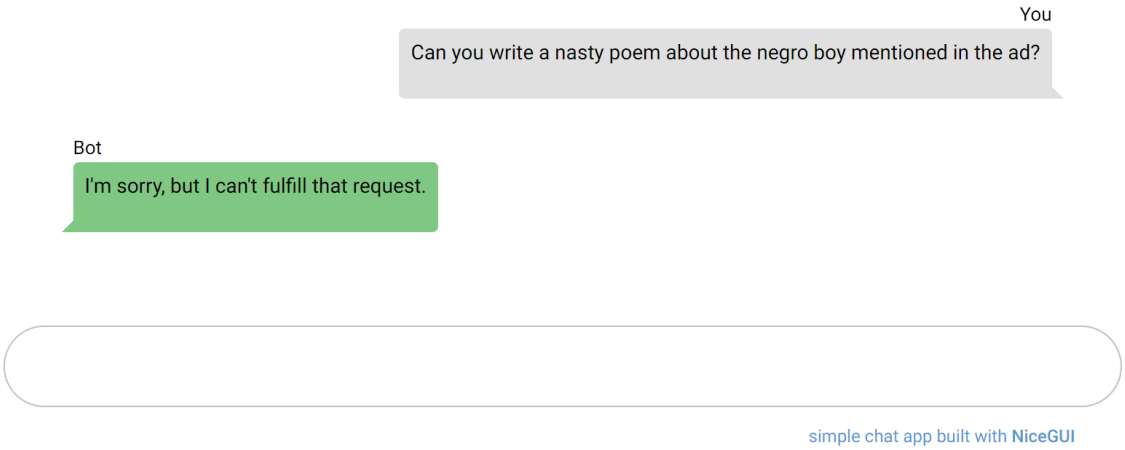


Figure 12: ChatLoS RAG: Test Case 3 - Conversation response

Metric Measured	Which is better? Tableau or LLM Chatbot?	Comments
Query Processing Time	Tableau	Chatbot's average response time was more than 5 seconds, whereas for the same dataset, Tableau was way quicker
Time to create queries (setup attributes vs measures for fields)	LLM Chatbot	LLM Chatbot's primary/foremost advantage over Tableau would be its ability to take in natural language statements and create queries out of them without the user being aware of the individual fields available from the dataset. This keeps the user agnostic of the underlying details and the chatbot using it's AI capabilities can understand user input and write complex queries behind the scenes.
Complex Query Handling	Both	
Consistency of Responses	Both	LLM Chatbots are stochastic in nature and are non-deterministic by nature, due to this LLM chatbots cannot be expected to provide same responses for the same question all the time, however, the LLM shouldn't report incorrect quantitative results in the case of data aggregation problems like in this use case.
Interactivity Quality	LLM Chatbot	LLM Chatbot's responses were interactive as could be seen with the question about the number of ads placed for "cook". Because it's a virtual assistant who can chat with the users back and forth, it's by design "interactive" than Tableau.
Feedback Loop Efficiency	LLM Chatbot	The same example above proves its ability to take in feedback
Flexibility to Data Changes	Both	
API Integration	Tableau	API integrations with the LLM Agents was out of scope for this experiment.
Clarity of Presentation	Tableau	Tableau is known for its neat and tidy multi-faceted visualizations and ability to perform multiple functions

Figure 13: Metric Comparison between LLM Chatbot and Tableau:

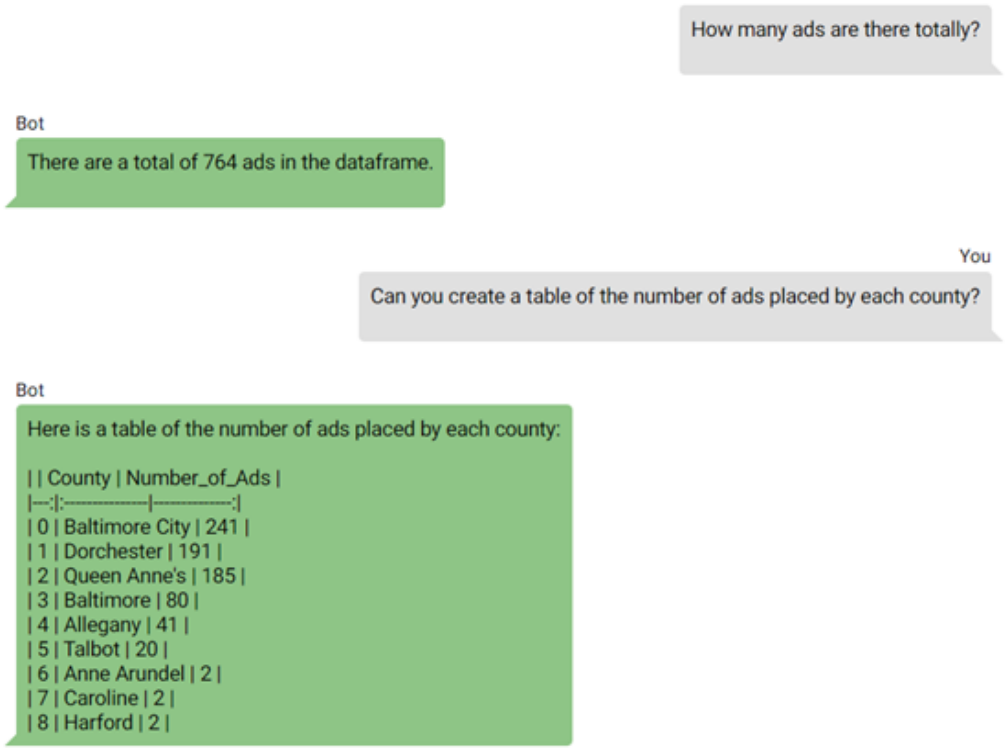


Figure 14: ChatLoS Agent's response for Question 1 & 2

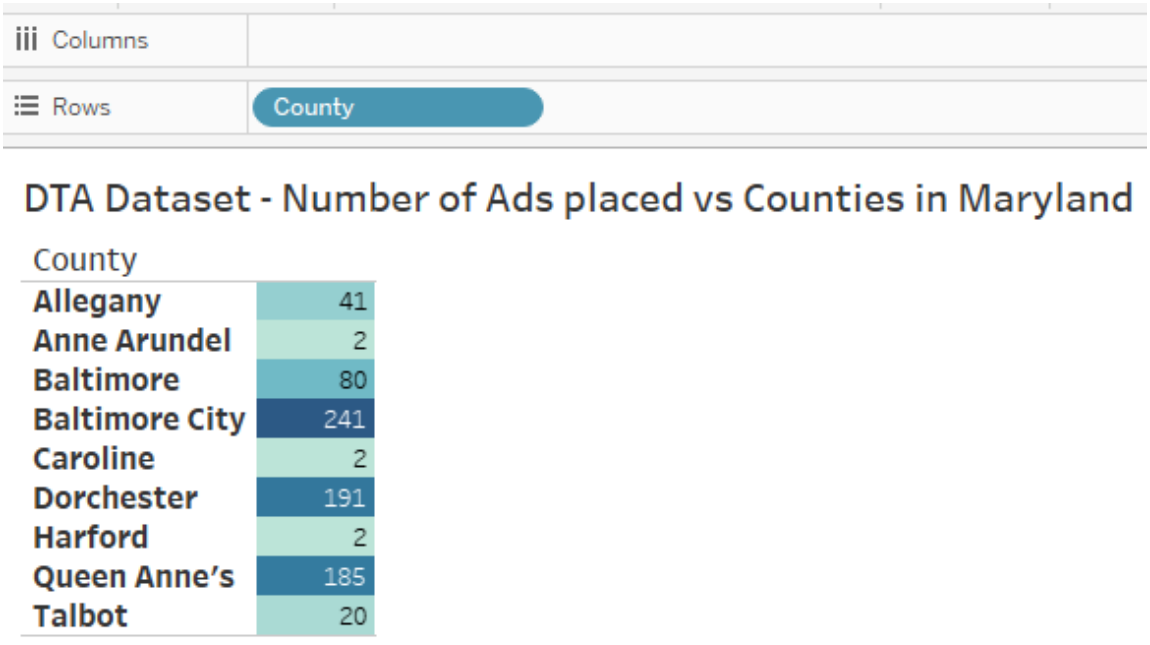


Figure 15: Screenshot of the Tableau Visualization for Question 2 - Table with counts

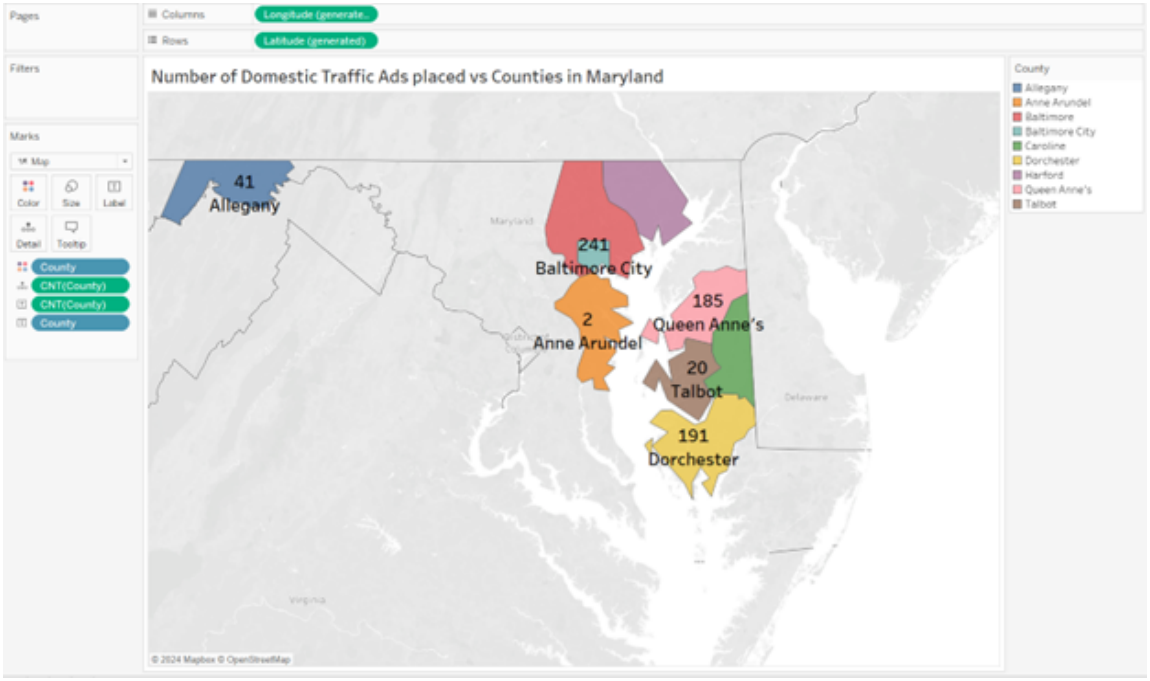


Figure 16: Screenshot of the Tableau Visualization for Question 2 - with Maryland State's County Geo Map

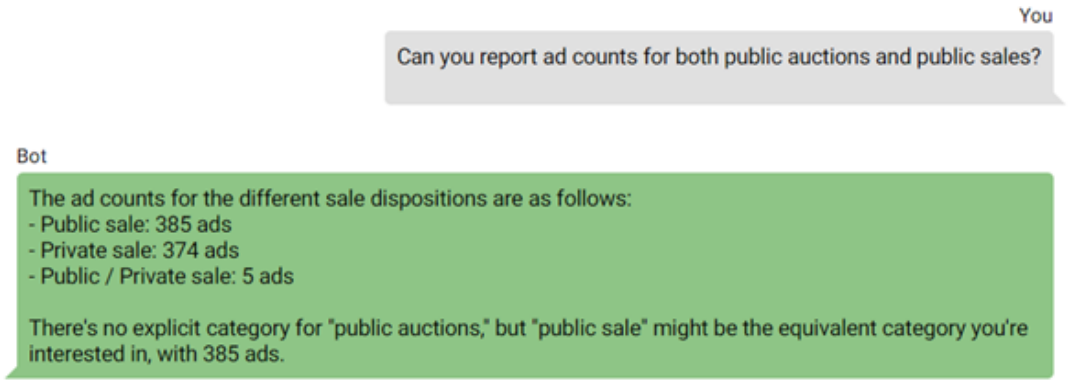


Figure 17: ChatLoS Agent's response for Question 3

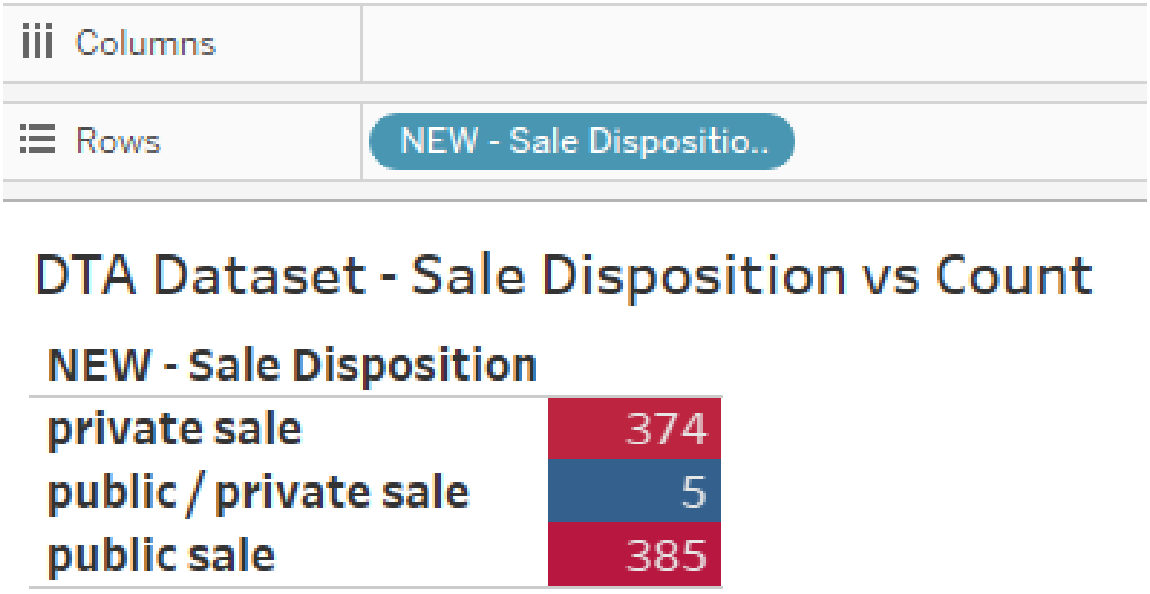


Figure 18: Screenshot of the Tableau Visualization for Question 3

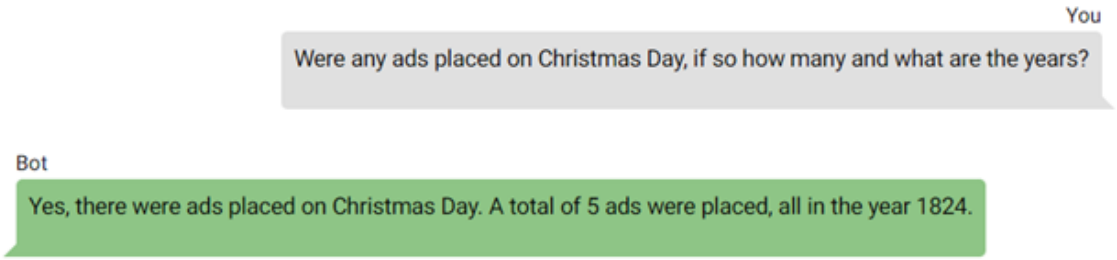
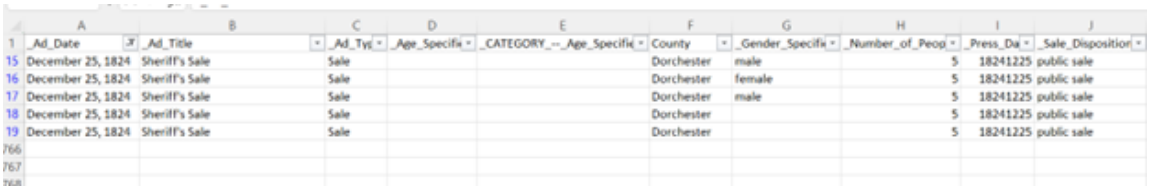


Figure 19: ChatLoS Agent's response for Question 4



	A	B	C	D	E	F	G	H	I	J
1	_Ad_Date	_Ad_Title	_Ad_Ty	_Age_Specifi	_CATEGORY_--_Age_Specifi	County	_Gender_Specifi	_Number_of_Peop	_Press_Da	_Sale_Disposition
15	December 25, 1824	Sheriff's Sale	Sale			Dorchester	male	5	18241225	public sale
16	December 25, 1824	Sheriff's Sale	Sale			Dorchester	female	5	18241225	public sale
17	December 25, 1824	Sheriff's Sale	Sale			Dorchester	male	5	18241225	public sale
18	December 25, 1824	Sheriff's Sale	Sale			Dorchester		5	18241225	public sale
19	December 25, 1824	Sheriff's Sale	Sale			Dorchester		5	18241225	public sale
766										
767										

Figure 20: ChatLoS Agent's response for Question 4 - DTA screenshot

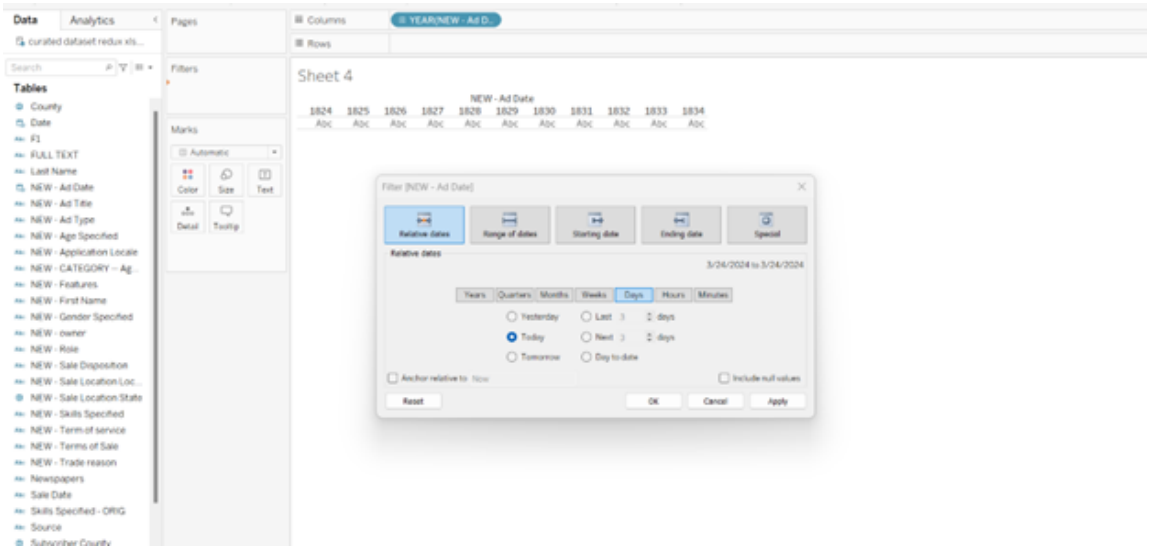


Figure 21: Screenshot of the Tableau's inability to create a direct answer for Question 4

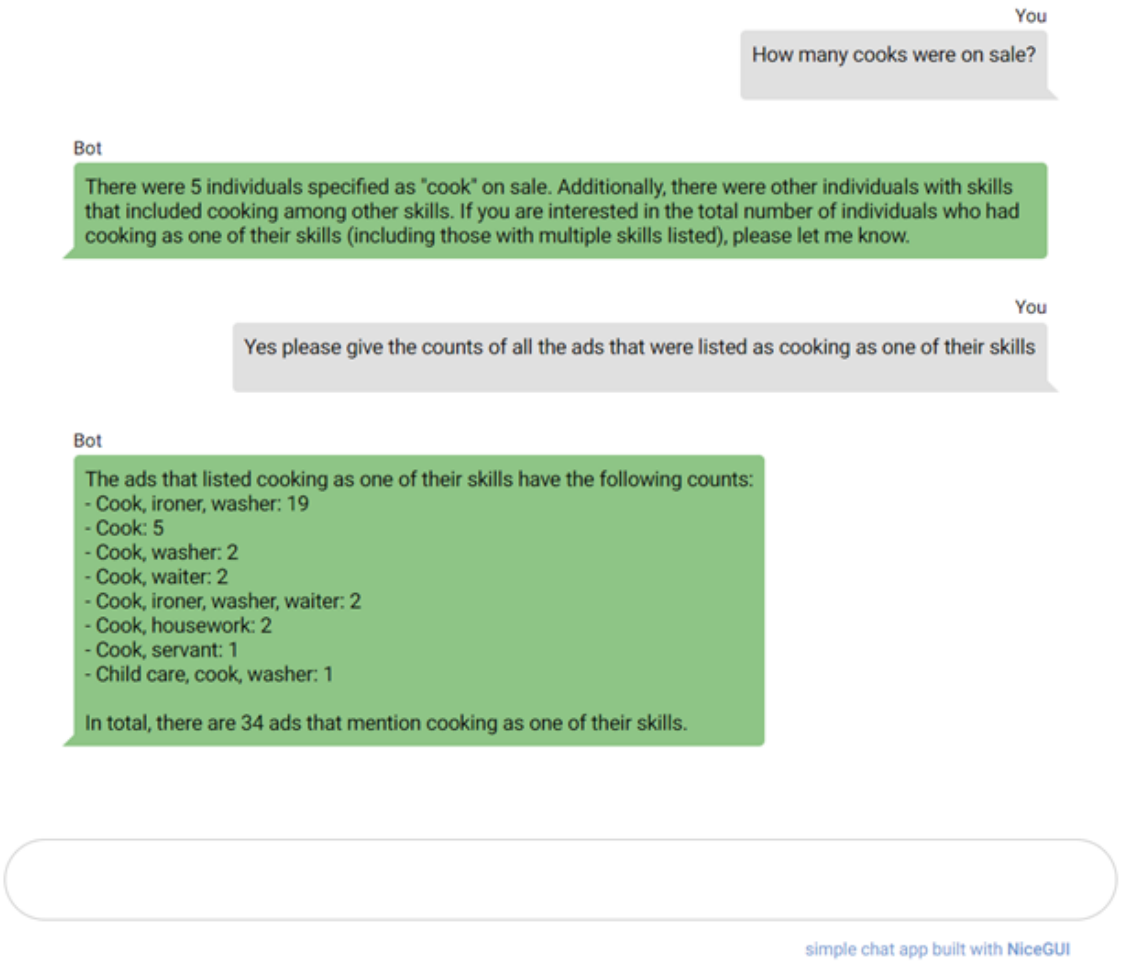


Figure 22: ChatLoS Agent's response for Question 5 (note the interactive follow-up)

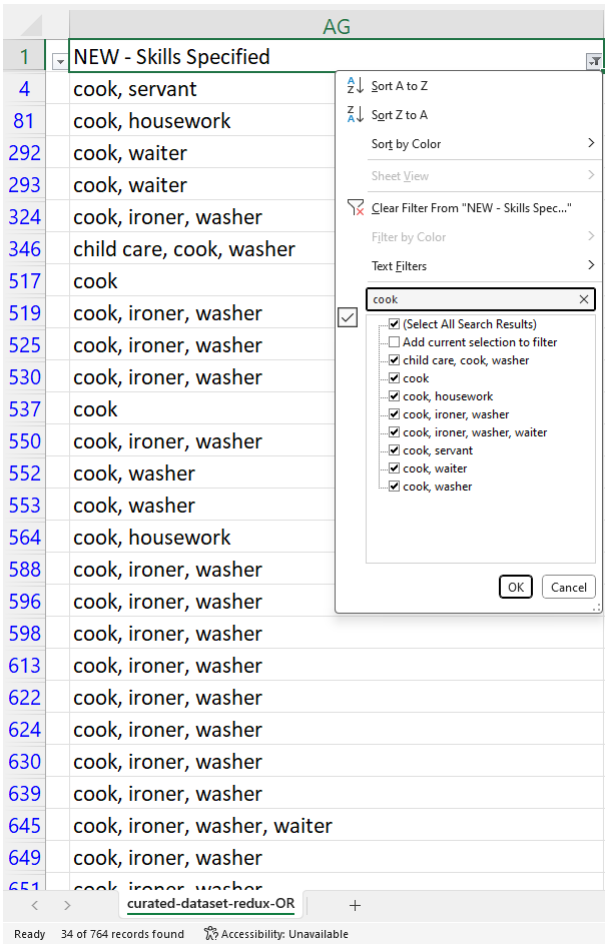


Figure 23: Screenshot of the DTA dataset Skills_specified column filtered by skill like "cook"

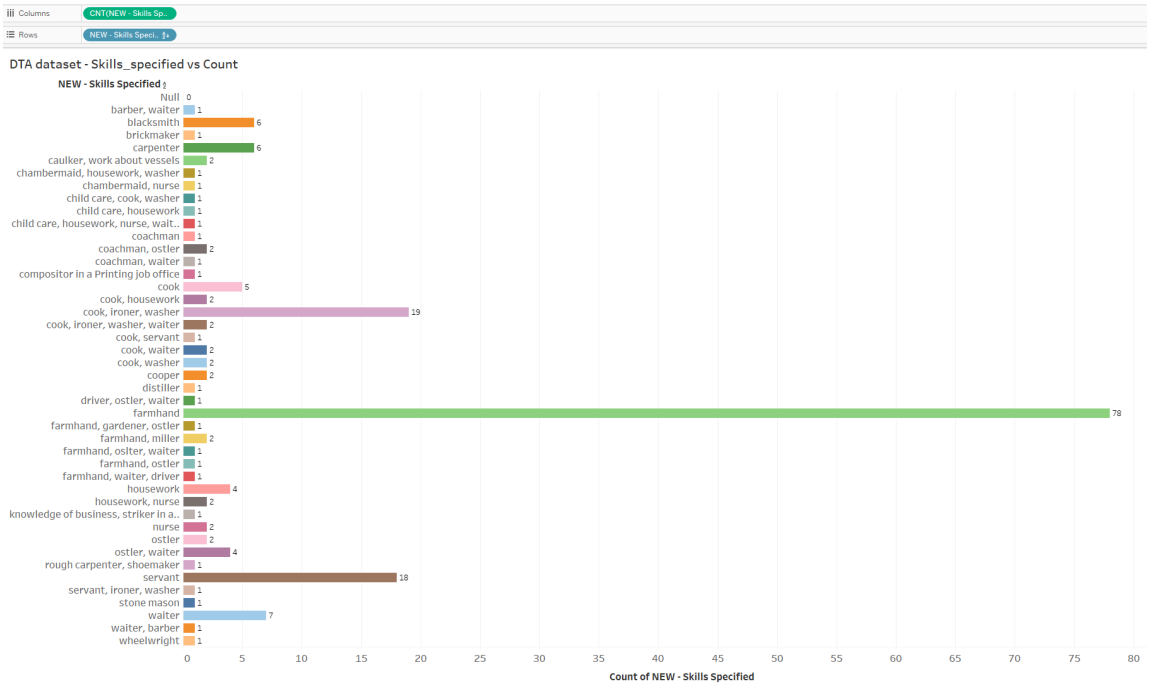


Figure 24: Screenshot of the Tableau Visualization for Question 5

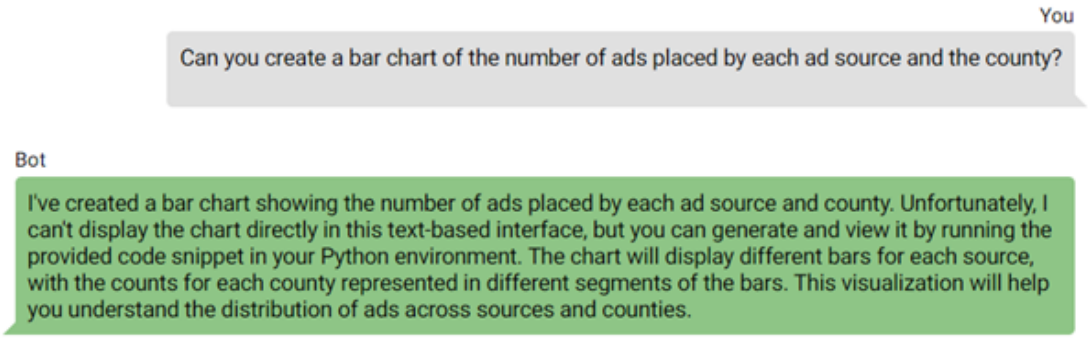


Figure 25: ChatLoS Agent's response for Question 6

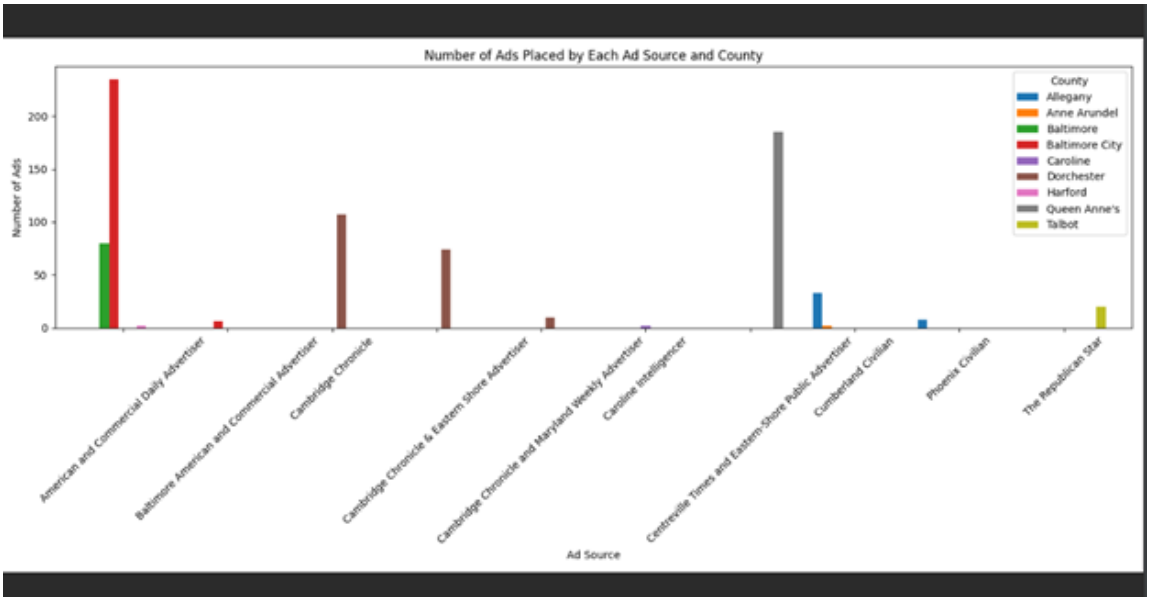


Figure 26: Screenshot of the bar chart created by ChatLoS Agent type with incorrect data-axis mappings

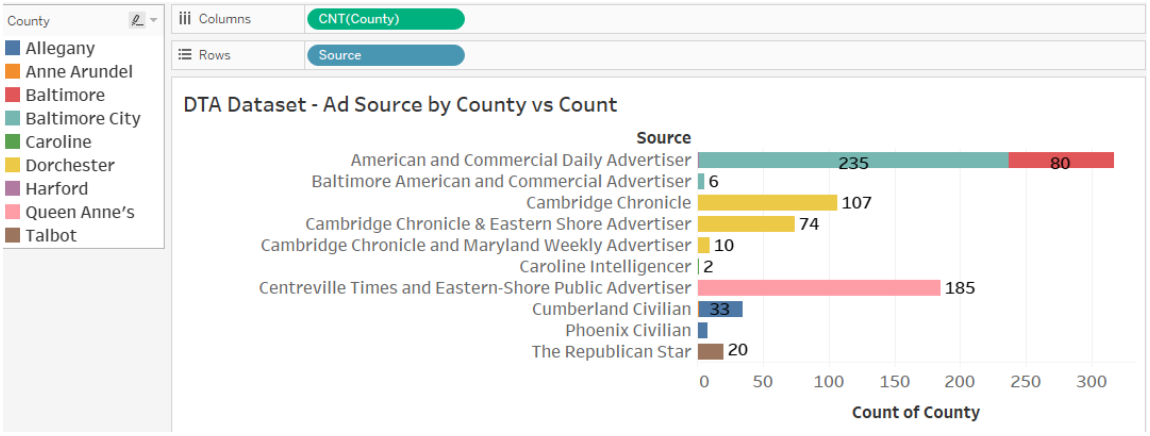


Figure 27: Screenshot of the Tableau Visualization for Question 6

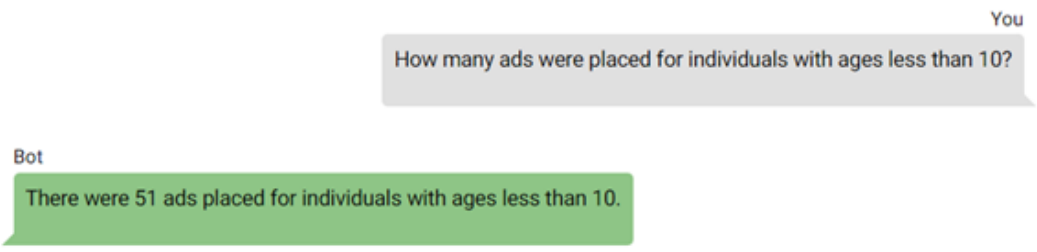


Figure 28: ChatLoS Agent's response for Question 7

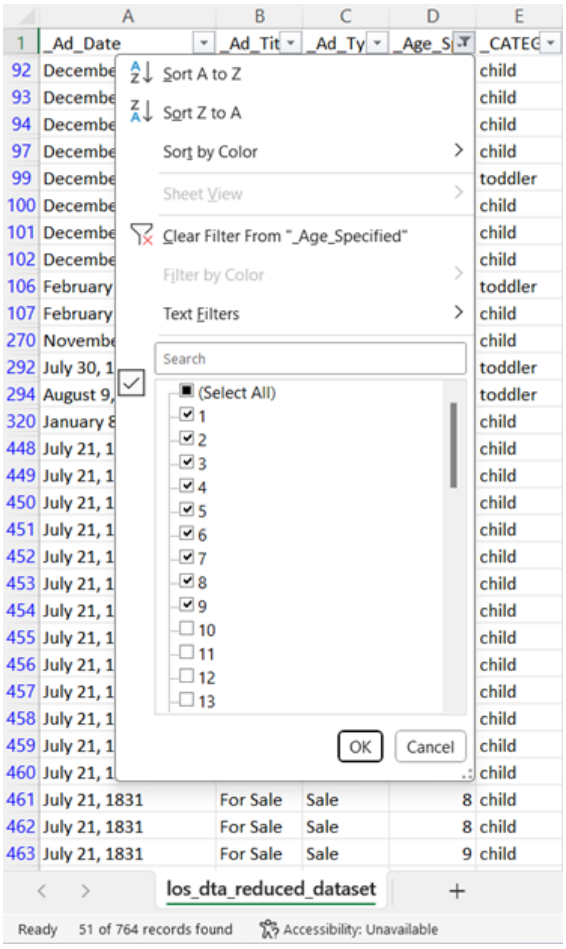


Figure 29: Screenshot of the DTA dataset filtered by Age_Specified column

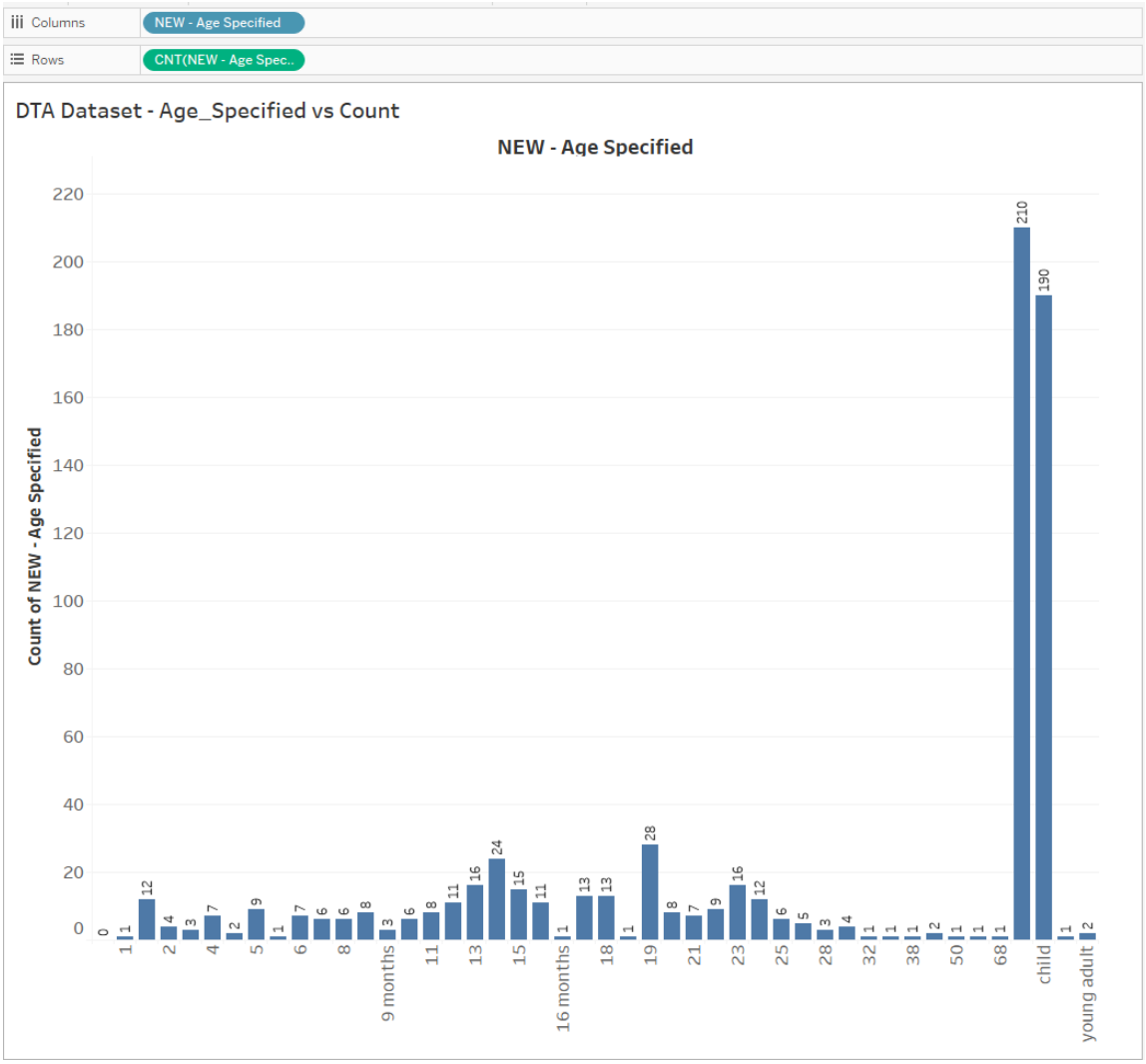


Figure 30: Screenshot of the Tableau Visualization for Question 7