ABSTRACT

Title of Dissertation: ENHANCING ACCESS TO CULTURAL ARCHIVES THROUGH DATA SCIENCE, GENERATIVE AI, AND KNOWLEDGE GRAPHS

Rajesh Kumar Gnanasekaran
Doctor of Philosophy, 2025

Dissertation Directed by: Professor Richard Marciano
College of Information

Cultural archives hold invaluable historical records, yet outdated cataloging methods and access barriers limit their usability. My dissertation addresses these challenges by integrating data science, generative AI, and knowledge graphs to enhance engagement with Maryland's Legacy of Slavery *(LoS)* project collections. My two-pronged strategy focuses on (1) developing computational tools to empower researchers and students to access and perform detailed data analysis on archival datasets independently and (2) leveraging generative AI and knowledge graphs to improve accessibility and contextual analysis. This dissertation synthesizes five interconnected studies conducted between 2021 and the present, structured under three research objectives. Research Objective I supports the first prong, while Objectives II and III collectively support the second.

Research Objective I – Empowering Archival Practitioners through Data Science and Computational Thinking: In Study 1 (published, 2021), I used a *mixed-method exploratory case*

study approach to develop *interactive Digital Computational Notebooks (iDCNs)* as educational tools for archival studies through a step-by-step data science based analysis on one of the LoS datasets. Designed based on a well-established computational thinking (CT) framework, iDCNs integrate Python scripts, narrative explanations, and visualizations to guide students and archivists in cleaning, analyzing, and interpreting LoS datasets. An IRB-approved user survey among students and educators confirmed that iDCNs enhance technical proficiency and critical thinking, making archival data more accessible for independent data analysis.

Research Objective II – Designing and Evaluating Generative AI Solutions for Enhanced Access: Addressing the second prong of my strategy, in Study 2 (peer-reviewed, 2023; to be published, 2025), I employed *design science and an exploratory case study* approach to develop *ChatLoS*, a chatbot powered by Retrieval-Augmented Generation (RAG) and OpenAI's GPT models, allowing users to query one of the LoS datasets in natural language. By chunking text for retrieval, ChatLoS preserved contextual relevance and eliminated the need for users to understand data schemas. While it significantly improved accessibility, its reliance on RAG introduced limitations, including ethical concerns, bias, data privacy risks, and constraints that restricted the chatbot's ability to handle complex, multi-step analytical tasks beyond targeted searches. To address these, in Study 3 (published, 2024), I explored *a comparative empirical design* in leveraging a generative AI Agent that enables dynamic complex data analysis beyond targeted semantic search retrieval. Findings revealed that while RAG-based retrieval is optimal for targeted semantic search with explainability, an AI agentic approach enhances exploratory analysis, trend identification, and multi-step reasoning also with explainability. However, both studies highlighted concerns about limited scalability across more extensive archival collections, inaccuracies, inconsistencies, usage of proprietary GPT models, including ethical risks, and data

control.

Research Objective III – Evaluating and Optimizing Generative AI for Ethical and Scalable Archival Access: To address issues identified in Objective II, the first study in this objective focused on completing the development of a *Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG)* enhanced ChatLoS. This system integrates multiple LoS datasets—Certificates of Freedom, Domestic Traffic Advertisements, and Manumissions—into a unified knowledge graph that supports explainable, multi-hop reasoning grounded in archival provenance and explainability. This ChatLoS version was further enhanced to promote transparency by providing source links to scanned archival documents and surfacing internal query logs in a human-readable format. This design enables users to understand how answers are generated and to verify the archival trail. Following this, two systematic user evaluations were conducted using an evaluation rubric grounded in the *Activity Theory* framework. These evaluations compared the current LoS access systems to the three evolving versions of ChatLoS, revealing how each Gen-AI enhanced iteration progressively mediated user interaction and resolved long-standing usability and interpretability contradictions. This study demonstrated how a KG-RAG enhanced generative AI system can resolve key contradictions in existing access methods by aligning tool behavior with archival principles such as context, trust, and traceability. This also identified opportunities for future work by introducing new contradictions. The second part of this objective evaluated four leading LLMs—GPT-4o, Claude, Llama, and Gemini—across criteria such as security, accuracy, guardrail customization, and multi-user deployment capabilities. Findings identified enterprise-grade models like Azure OpenAI GPT-4o as more appropriate for sensitive archival applications.

The dissertation concludes with a synthesis of findings that integrates the results of the three research objectives. It also highlights promising directions for future work, including participatory

community studies and interface enhancements to ensure equitable, transparent, and culturally sensitive access to digital archives.

ENHANCING ACCESS TO CULTURAL ARCHIVES THROUGH DATA SCIENCE, GENERATIVE AI AND KNOWLEDGE GRAPHS

by

Rajesh Kumar Gnanasekaran

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2025

Advisory Committee:
Professor Richard Marciano, Chair/Advisor
Professor William Regli, Dean's Representative
Professor Vanessa Frias-Martinez
Dr. William Underwood
Professor Jane Greenberg

Dedicated to the loving memory of

Late Mr. Gnanasekaran Mannu

A devoted father, mentor, and friend

whose wisdom, kindness, and unwavering support

continue to guide me every day.

This dissertation is for you, Dad.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AI	Artification Intelligence
AT	Activity Theory
CAS	Computational Archival Sciences
CoF	Certificates of Freedom
CT	Computational Thinking
DTA	Domestic Traffic Adverstisements
Gen-AI	Generative AI
GPT	Generative Pre-Trained Transformer
KG	Knowledge Graph
LoS	Legacy of Slavery
LLM	Large Language Models
MSA	Maryland State Archives
RAG	Retrieval Augmented Generation

Chapter 1: Introduction

1.1 Background and Significance

Cultural archives serve as custodians of collective memory, preserving narratives that shape historical understanding and contemporary identity [2]. Yet these archives are often underutilized by the public and researchers due to barriers in access and interpretation [3]. Traditional archival practices rooted in manual cataloging and reading struggle to keep pace with the scale and complexity of digital-era collections. Many archives have become essentially large troves of data preserved but not readily accessible to users [3]. For example, digitized collections like the Maryland's Legacy of Slavery (LoS) project contain extensive data on enslaved individuals and emancipated communities. However, unlocking insights from such data requires computational analysis beyond conventional methods. This gap between archival wealth and accessibility has been widely recognized as a critical challenge [3] [4]. Significant portions of cultural heritage remain unexplored without new approaches, and essential stories within these records stay untold. At the same time, the rise of data science and artificial intelligence (AI) offers promising avenues to bridge this gap. In particular, computational methods can dramatically enhance how we process, analyze, and engage with archival materials [4]. AI has proven "*essential for cleaning, exploring, and visualizing archival and special collections*" [4]. These technologies can facilitate what was once impossible: rapidly extracting patterns from millions of records or conversing with an

archive in natural language. However, realizing this potential in archives comes with its challenges. On one hand, cultural heritage institutions face a skills gap – many archivists, librarians, and humanities scholars have not been trained in computational thinking or data science [3] [5]. On the other hand, AI systems (especially generative AI (Gen-AI) [6] powered solutions like large language models (LLMs) [7]) carry risks of ethical pitfalls – for example, they might introduce inaccuracies or biases ("hallucinations" [8]) when applied to sensitive historical data [9]. The convergence of these trends has created an urgent need for research that empowers users with computational skills and develops AI-driven tools ethically and responsibly to enhance archival access by overcoming these challenges.

This dissertation responds to that need by proposing a comprehensive approach to enhance access to cultural archives through data science and Gen-AI, a two-pronged approach explained in detail below. The significance of this work lies in bridging the gap between traditional archival practices and cutting-edge computational techniques [10]. In doing so, it contributes to the emerging field of computational archival science (CAS), which formally blends archival science and computational methods to tackle modern archival challenges [10]. Ultimately, the goal is to transform how cultural archives are accessed and interpreted (e.g., those of enslaved individuals in Maryland's history from the LoS collections). This transformation is not just technical – it has profound implications for education, research, and community engagement around archives. Researchers and students can ask new questions about historical data; librarians and archivists will gain new skills and tools to serve the public; and interested patrons and communities could find new pathways to connect with their heritage.

1.2 Problem Statement and Research Objectives

The core problem addressed in this research is the limited accessibility and interpretable engagement with large-scale, culturally sensitive archival datasets under current practices. Key contributing issues include: (a) insufficient computational skill sets among many archival sciences scholars and information professionals, leading to underuse of data-driven methods in archives [5]; (b) a lack of tailored tools that allow non-technical users to explore archival data in meaningful ways, beyond simple catalog searches; and (c) the nascent state of applying Gen-AI to archives, which has shown promise but also revealed challenges in accuracy, context-awareness, and ethical sensitivity [4] [9]. These gaps mean that vast archives (for instance, records of slavery, historical newspaper content) are not yielding their complete potential insights to historians, educators, or the communities they represent. To address these challenges, this dissertation pursues a two-pronged strategy:

- First prong: Develop hands-on learning tools to equip students, archivists, and community members with computational thinking (CT) and data science techniques for working with archival collections independently.

- Second prong: Leverage, evaluate and optimize Gen-AI (via agents and knowledge graphs) to create innovative applications (such as interactive conversational platforms aka chatbots [6]) for engaging with digital archives. This strategy fosters interactive, context-rich engagement with cultural archives, including enhancements and optimizations to the user experience and the ethical safeguards surrounding LLMs.

The research addresses both sides of the problem by integrating these prongs – building human

capacity to work with data and building advanced technical systems to augment what humans can do. I adopt a mixed-methods research design to achieve these objectives. The guiding philosophy is that each study addresses a facet of the overall problem, using methods suited to that facet, while informing the subsequent studies in an iterative, reflective process. The two-pronged approach (education and Gen-AI tools development) is operationalized through five interconnected studies conducted from 2021 to 2025, corresponding to each of this dissertation's three core research objectives. Here I summarize each of these objectives and their associated studies with the approach taken in each study and how they are interrelated:

1.2.1 Objective I: *Empowering Archival Practitioners through Data Science and Computational Thinking:*

Through Study 1, *a mixed-method exploratory case study* (peer-reviewed and published in October 2021 [11]), I design and develop interactive, cloud-based learning platforms (specifically, interactive Digital Computational Notebooks, or iDCNs) aimed to teach students and professionals computational skills and data analysis techniques in the context of real archival datasets. These tools were developed by following an established CT framework and lower the barrier to applying data science methods on archival records, by providing guided, narrative-rich environments where users can practice data cleaning, analysis, and visualization on cultural heritage data independently [11]. The effectiveness of these tools in improving users' computational skills and confidence in working with archives has been assessed through an IRB-approved case study survey, and optimistic user feedback was received. Study 1 establishes the feasibility of using computational notebooks as a learning tool to bridge archives and data science. It addresses Prong

1 of my approach by focusing on human learning and empowerment. For this study, the research questions identified and answered are:

- RQ1A: Can a novel set of cloud-based iDCNs be developed with content used as a learning tool to teach computing for higher education students from non computational backgrounds?
- RQ1B: Can the Maryland State Archives (MSA)'s digitized Certificates of Freedom (CoF) dataset from the LoS collection be used to create the computational content as a case study through a step-by-step contextual data science driven analytical treatment?
- RQ1C: Can the CT practices be applied to non-STEM backgrounds in creating the computational content?
- RQ1D: Can the resulting novel learning environment be useful for Library and Archival Science educators and students? How would they react to such changes to their pedagogical learning environment?

1.2.2 Objective II: *Designing and Evaluating Generative AI Solutions for Enhanced Access:*

To achieve this objective, I conduct two studies, Study 2 and 3, focusing on leveraging and evaluating Gen-AI to iteratively design and develop interactive solutions to enhance user engagement with cultural archives. Study 2 (peer-reviewed in 2023 and to-be published in June 2025), *an exploratory case study and design science* approach guided the development of an AI-powered chatbot ("*ChatLoS*"). It was built using a unique Retrieval-Augmented Generation (RAG) mechanism and OpenAI's Generative Pre-Trained Transformer (GPT) models to establish a

context-aware chatbot. In ChatLoS v1, I implemented a RAG-based retrieval mechanism to enable natural language querying of one of the LoS datasets. This chatbot eliminated the need for users to understand the underlying data schema, making archival data more accessible to researchers and the public [12] [13]. By chunking large text into manageable units, ChatLoS preserved contextual relevance while optimizing computational efficiency. Results showed ChatLoS significantly improved accessibility for non-technical users while maintaining contextual relevance. Despite its success in enhancing access, RAG-based retrieval introduced limitations, including ethical sensitivity concerns, bias, data privacy risks, and context window constraints that restricted the chatbot's ability to handle complex, multi-step analytical tasks beyond targeted searches. The outcome of this study is a functional prototype and initial evidence of its capabilities. This study answered the research question:

- RQ2: What are the potential benefits and challenges associated with utilizing LLMs, such as OpenAI's GPT, to develop a tailored chatbot capable of exploring data from high cultural context datasets and unveiling previously undiscovered relationships and patterns within this data?

Drawing from the initial positive results of Study 2 and to address its limitations, I conducted Study 3. As *a comparative empirical design* study, Study 3 (peer-reviewed and published in March 2024 [14]) took a step further to assess the strengths and limitations of the Gen-AI tool in performing advanced data analysis and visualizations as explored in Study 1. To do this, Study 3 employed Gen-AI agents [15] in ChatLoS as v2 and empirically compared its advanced data analysis capabilities to the traditional data-visualization and exploratory tool, Tableau. By measuring accuracy, efficiency, and ethical sensitivity, the study demonstrated that an Gen-AI

agentic approach overcomes some limitations of RAG by allowing users to perform complex analyses, trend identification, and multi-step reasoning on archival datasets. By systematically comparing ChatLoS v2 with a traditional data visualization tool on the same archival dataset, the research identified scenarios where Gen-AI provides added value and where it falters [4] [14]. ChatLoS v1 and v2 with RAG and Gen-AI agents serve distinct yet complementary purposes: RAG-based retrieval is optimal for targeted semantic search, whereas Agentic Gen-AI excels in exploratory and advanced data analysis. The objective was to leverage and evaluate whether a Gen-AI chatbot can improve accessibility and interactivity in archival research, compared to traditional search or static data visualization interfaces, which was achieved through these two studies. The research question answered by this study is:

- RQ3: 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?

The findings from Study 2 and Study 3 confirmed that Gen-AI significantly enhances archival access and usability, with context-aware search and advanced data analysis. However, both studies exclusively used OpenAI's GPT models, raising concerns about data privacy, proprietary control, and the potential use of user interactions in model retraining. These issues inspired me to evaluate multiple LLMs to determine the most ethically sound, technically robust, and scalable Gen-AI model for cultural archives. Additionally, neither study addressed the scalability of ChatLoS for deployment across more extensive archival collections or multiple datasets within the LoS project. ChatLoS's context awareness remained limited to a single dataset, restricting its ability

to draw meaningful historical connections across various datasets within the LoS collection with potentially interlinked relationships.

1.2.3 Objective III: *Evaluating and Optimizing Generative AI for Ethical and Scalable Archival Access*

As a continuation of my dissertation strategy's second prong, and to address the limitations identified from Objective II, as a first part of Objective III, I conducted Study 4 (submitted in May 2025) to optimize AI-driven archival analysis through Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG). The motivation to perform this study arises from the research that calls for data available to the Gen-AI RAG tools be "*AI-Ready* [16]" to mitigate the limitations of Gen-AI solutions due to occasional factual errors or lack of deep historical context. KGs can supply structured context (e.g. relationships between people, places, events in the archive) that the Gen-AI can draw upon, making responses more accurate and culturally contextual [4] [17]. Unlike the earlier single-dataset RAG implementation in Objective II, this study integrates multiple LoS datasets (e.g., Certificates of Freedom, Domestic Traffic Advertisements, and Manumissions) into a unified KG, allowing the chatbot to link records across collections and provide richer genealogical and historical insights. Connecting historically related documents reduces hallucinations, enhances contextual accuracy, and strengthens cross-dataset relationships by providing structured semantic representations and facilitating consistent data integration and interoperability. KG-RAG enhanced ChatLoS as v3 also enhanced the contextual richness of responses by structuring data as AI-ready knowledge representations and anchoring outputs with citations to scanned archival documents and human-readable query logs to enable *"explainability"* in these Gen-AI solutions. This study

addressed contradictions in previous Gen-AI implementations by aligning tool capabilities with archival principles of provenance and interpretability. To evaluate the real-world effectiveness of all these versions of ChatLoS, two IRB approved systematic user evaluations were conducted: one with the MSA Director and another with the Deputy Director and the results were triangulated from these user interviews. These evaluations were guided by an *AT*-grounded evaluation rubric, allowing for the comparative assessment of ChatLoS v1 (RAG based), v2 (AI Agent), and v3 (KG-RAG) alongside the current traditional access tools like the MSA online database. These studies illuminated how each version of ChatLoS mediated archival research practices, resolved usability and trust-related tensions, and enabled expansive learning through deeper data engagement.

The research questions answered in this part of Objective III are:

- RQ4A: How does the integration of Knowledge Graph-based Retrieval-Augmented Generation (KG-RAG) with LLMs enhance the analysis and accessibility of interconnected historical archives, specifically the LoS collections at the MSA?
- RQ4B: How do users across different roles evaluate evolving Gen-AI systems for archival access when assessed using a rubric grounded in Activity Theory?

As the second part of Objective III, in Study 5 (published, 2024), I applied a *three-step model-comparison methodology*, using *design science and empirical evaluation* to assess four leading LLMs—GPT-4o, Claude, Llama, and Gemini—across free, paid, and enterprise cloud-based environments. These models were evaluated along key dimensions including data privacy, customizability through guardrails, multi-user support, accuracy, cost-efficiency, and security. Results indicated that enterprise solutions like Azure OpenAI GPT-4o offer the most favorable balance for sensitive archival deployments. These findings are crucial for informing future

development and deployment strategies, particularly in public-sector or cultural heritage contexts with stringent ethical requirements. The research question answered by this study is:

- RQ5: Which LLM is best suited for analyzing 19th-century domestic trade advertisements while maintaining historical accuracy and ethical sensitivity?

Throughout these studies, the research methodology remains iterative and interlinked. Each chapter builds on the previous ones to advance its respective research objective. A combination of exploratory case studies, iterative design science, empirical evaluations, and AT-grounded user assessments ensures both the technical viability and practical relevance of the solutions. This mixed-methods approach provides a comprehensive understanding of how these Gen-AI systems transform the activity of archival research and offers direction for equitable and culturally respectful technological interventions [12]. A pictorial representation of the hierarchical organization of the five studies in the dissertation is shown in the Figure. 1.1. The dissertation research timeline roadmap is shown in the Figure. 1.2.

1.3 The Legacy of Slavery (LoS) Dataset Collection

Since its initiation in the fall of 2001, the Maryland State Archives (MSA)'s project, officially named the Study of the LoS [18] in 2005, has focused on uncovering the narratives of individuals who resisted enslavement in Maryland, USA. The project's original concept was to discover unknown 'heroes' of slave flight and resistance. 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

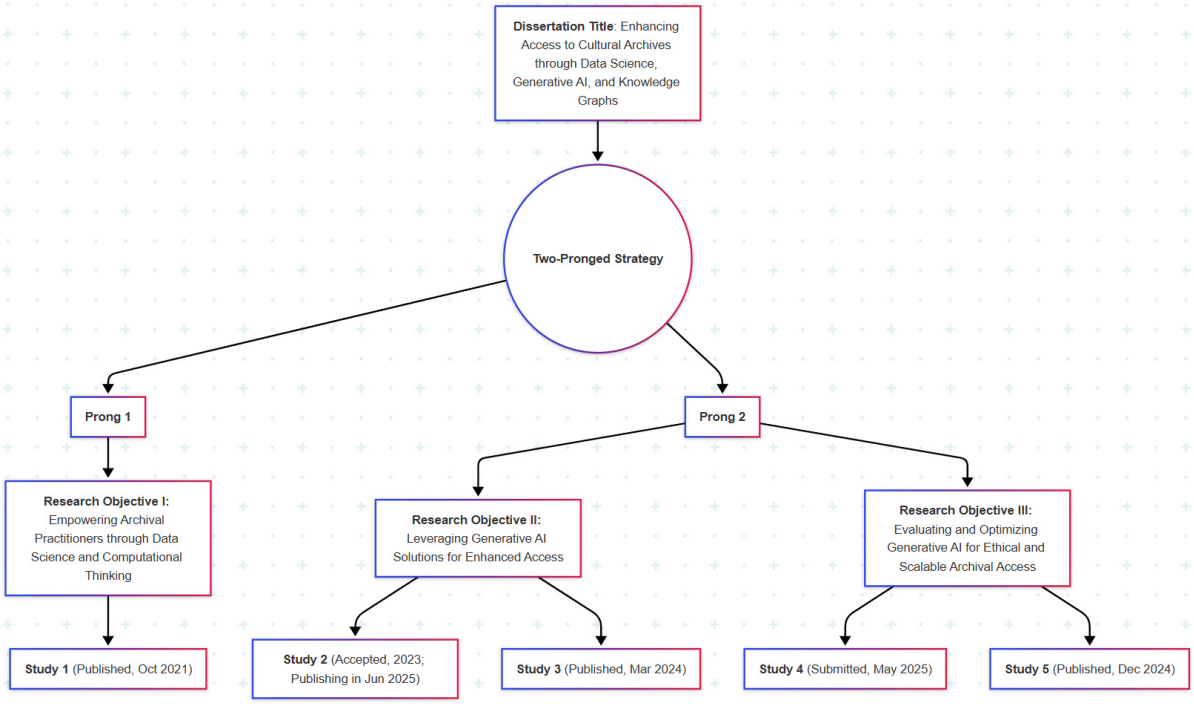


Figure 1.1: Pictorial Flow chart of the Dissertation Research

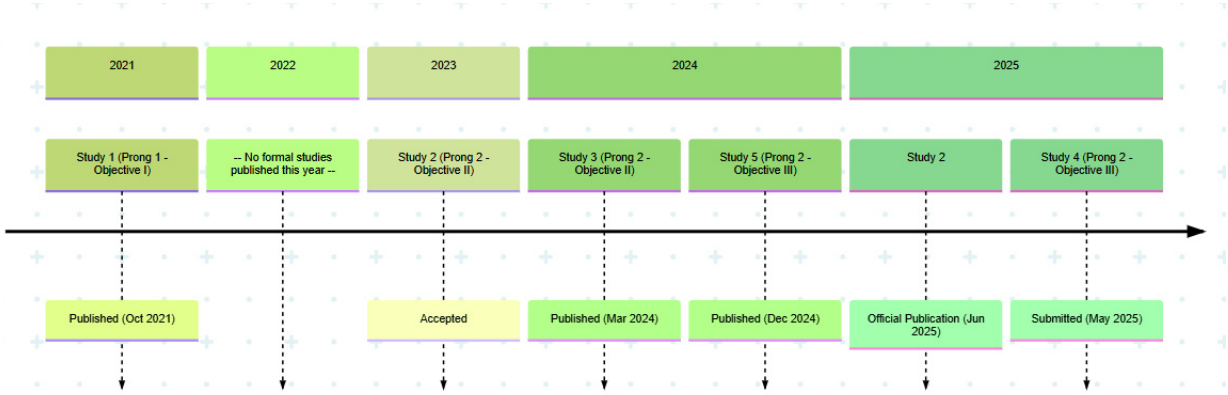


Figure 1.2: Dissertation Research Timeline Roadmap (2021 - 2025)

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 (CoF) 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 [19] [20]. Since 2019, I have been part of the MSA's LoS project trying to uncover and unravel the hidden stories and insights from these culturally rich and historically sensitive LoS dataset collections by collaborating with the MSA director and other multi-disciplinary researchers. This dissertation uses the datasets from this LoS collection. In Study 1, I used the CoF dataset, Studies 2, 3, and 5 used the DTA dataset, and Study 4 will use the CoF, DTA, and Manumissions datasets.

1.3.1 Certificate of Freedom (CoF)

A CoF [21] is a legal document issued from 1803 to 1865 in Maryland to African American Enslaved persons, who were required to record proof of their free or emancipated status in the county court. The certificate was issued based on information documented and provided by the former slaveholder and a witness. The CoFs were handwritten documents containing general biographic, demographic, and descriptive information about the enslaved person. In most cases, the data captured or documented for the CoF by the courts or court clerk followed a set of standards for documents across different counties which consisted of data features like county issuing the document, slave owner's first and last name, enslaved person's first and last name, gender, age, height, complexion, scars (for identification purposes), alias name, date of issue, witness name, prior status of the enslaved person, and a notes feature to enter comments/remarks. There are

23,655 records in the LoS collection CoF dataset—the issuance dates range from 1806 to 1864, covering sixteen Maryland counties and one Maryland city.

1.3.2 Domestic Traffic Ads Dataset (DTA)

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, and children. DTAs were a means of communicating to the general public the subscriber's desire to buy or sell a slave(s). Private slave dealers and agents could place ads, gentry needing domestic help, yeomen needing extra field hands, or a public sale of an estate by the orphan's court. These documents, cataloged, calculated, and geo-spaced may identify sales patterns during different periods of a year or era. The DTA dataset comprises digitally transcribed data of advertisements for enslaved individuals placed in Maryland newspapers over 40 years, 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. There are about 2100+ advertisement records found in MSA's digital database. Figure. 1.3 and 1.4 show a few examples of scanned images of the DTA ads.

1.4 "Activity Theory" as an Unifying User Evaluation Theoretical Framework

Activity Theory (AT) is a theoretical framework for analyzing human practices, viewing any goal-directed activity as a holistic system of interacting components [22]. Originating from Vygotsky and Leont'ev's work in cultural-historical psychology, AT was later expanded by

Valuable Negroes

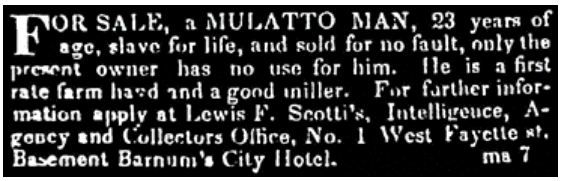
AT PRIVATE SALE.

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 1.3: 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 1.4: Sample 2 of the DTA scanned ad.

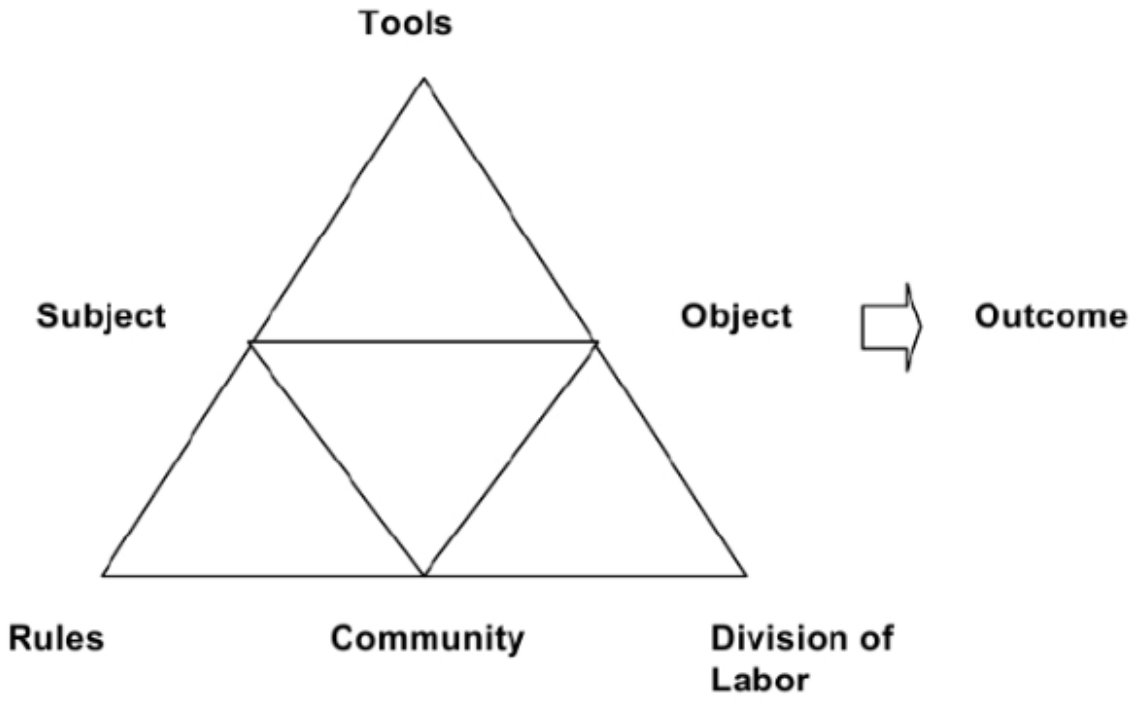


Figure 1.5: Pictorial representation of Activity Theory - Engeström's Triangular model [1]

Engeström to model complex, collective activities [22]. A core insight of AT is that humans do not act in isolation or in a vacuum of mere stimulus-response; rather, action is mediated by context, tools, and social relations. By focusing on the entire activity system (including people, artifacts, and social rules), AT provides a holistic, systemic lens for understanding technology use and sociotechnical change. This makes it especially suitable for research at the intersection of users, digital tools, and institutional practices – exactly the scenario of enhancing access to cultural archives. *Key Concepts:* In AT's classic second-generation model (Engeström's triangular representation [22]), an activity system comprises the following elements:

- **Subject:** The individual or group engaged in the activity, e.g. an archivist, researcher, genealogist, student or a patron using a tool.
- **Object:** The goal or problem-space that the activity is directed toward, which is transformed into an *Outcome*. For instance, the object could be analyzing archival data to gain historical insights, with the outcome being new knowledge or solutions.
- **Tools (Mediating Artifacts):** The material or symbolic means used by the subject to act on the object. Tools can be physical instruments, software, language, or even abstract artifacts like knowledge and skills. In this context, tools include Gen-AI powered conversational chat interfaces, which mediate the subject's interaction with archival data. Importantly, tools carry with them the history and culture of their development; they shape how the task is performed and can introduce new capabilities or constraints.
- **Community:** The broader social group or stakeholders involved in or affected by the activity. For example, the community in an archival research activity includes other archivists,

historians, genealogists, students, institutional authorities (like Maryland State Archives staff), and even descendant communities interested in the archives.

- **Rules:** The norms, regulations, and social conventions that influence how the activity is carried out. Rules can be formal policies (e.g. archival access rules, data privacy laws) or informal cultural expectations (e.g. ethical guidelines for Gen-AI). These rules mediate the relationship between the subject and the community, shaping acceptable behavior. In practice, rules might include metadata standards for cataloging such as archival descriptor records, protocols for using archival material, or guidelines for Gen-AI usage in sensitive historical contexts.

- **Division of Labor:** The way tasks and responsibilities are distributed among participants in the activity. This covers the roles of different people (or even Gen-AI agents) and the power dynamics or expertise they contribute. In an archival setting, division of labor might delineate what tasks are done by archivists (data curation, providing context) versus end-users or Gen-AI tools (searching, analysis). It mediates the relationship between the community and the object, determining who does what in pursuit of the goal.

A pictorial representation of the AT elements arranged together is shown in Figure. 1.5. All these elements interact as a *system*. AT emphasizes that changes or tensions in one element can ripple through the whole system. In AT, *tensions* and *contradictions* are seen as the driving forces of change and learning within an activity system. A contradiction refers to a deep-seated structural misalignment such as between the tools available and the user's goals, or between institutional rules and emerging practices that impedes smooth functioning. Tensions, on the other hand, are the observable symptoms of these contradictions, often manifesting as user frustration,

inefficiencies, or conflicting expectations. Contradictions are not failures but opportunities for transformation, and their resolution through redesigned mediations is central to the AT grounding of this research. Crucially, contradictions, historically accumulated tensions or mismatches either within an element or between elements of the activity system are the primary drivers of change and development. When a new tool is introduced or when objectives evolve, contradictions often emerge (for example, a new technology might conflict with old rules or skillsets). Rather than being purely negative, these contradictions can lead to innovation and learning: they catalyze adjustments in the activity, sometimes resulting in an expanded or transformed activity system that resolves the tension. Engeström calls this process expansive learning, wherein the object and motive of the activity are reconceptualized to accommodate a broader vision, leading to a qualitative shift in practice [22]. By applying AT, I have aimed to capture not only the efficacy of a tool, but also how that tool mediates practice, how it fits (or clashes) with users' objective and outcomes, and what new practices or learning emerge. In designing and evaluating user interactions in digital environments, AT provides a rigorous vocabulary to capture contextual factors. It ensures our focus remains on *activities and outcomes, not just on interface features*. In summary, AT provides a unifying theoretical framework for analyzing the systematic user evaluation studies that follow an evaluation rubric by linking the technical interventions (ChatLoS - a Gen-AI chat interface) to the human contexts of archival work and analysis. It allows me to frame each study as an intervention into an activity system, analyze the mediations introduced by the new tool, and understand the resulting changes (outcomes) in terms of resolving contradictions or creating new possibilities for the community. AT's application in sociotechnical contexts underscores that the success of an innovation like Gen-AI depends on the whole activity system's alignment. Contradictions often surface – e.g. between a Gen-AI tool's capabilities and existing

work rules or skill sets – but these can be sources of expansive learning. Identifying and addressing such contradictions is crucial for designing Gen-AI interventions that truly empower users rather than disrupt or alienate them. The details of this evaluation would be explained in detail as part of the research objective III.

1.5 Archival Description, Finding Aids and AI-Enhanced Contextualization

Archival description traditionally encompasses the creation and maintenance of *finding aids* guided by key archival principles such as provenance and original order. Finding aids are structured tools intended to facilitate discovery and access to archival materials, often involving detailed manual processes that contextualize records based on their source, content, and inherent relationships. This meticulous work remains largely manual and heavily reliant upon archivists' subject-matter expertise and understanding of historical contexts. This thesis acknowledges the significance of archival description and explicitly situates itself within the larger context of traditional archival practices. By leveraging data science techniques, generative AI, and knowledge graphs, this research aims to enhance and augment, rather than replace, these traditional archival methods. Specifically, this thesis explores how computational methods can complement manual archival description processes by:

- Automatically extracting and identifying digitized descriptive metadata that adhere to archival standards.
- Facilitating robust semantic linking and cross-referencing across archival collections, thereby enhancing the access to the digitized versions of the traditional archival finding aids.
- Preserving contextual integrity by clearly tracing provenance and original order through

digitally enriched pathways.

By incorporating these computational methodologies, this thesis proposes an evolution of access to the digital archives, enhancing their robustness and scalability while maintaining their archival description foundational principles.

1.6 Organization of the Dissertation

This dissertation is organized into six chapters, corresponding to the research structure. Below is an overview of each chapter and how they connect to form a coherent narrative:

- Chapter 1: Introduction – *(Current chapter).* Provides the overall framing of the background, research problem, research objectives, and methodology. It establishes why enhancing access to cultural archives is important and challenging, introduces the two-pronged approach (computational learning tools and Gen-AI applications), and outlines the studies undertaken. This chapter also defines key terms and concepts (such as CT, LLM, GPT, Gen-AI, and KGs) that will recur throughout the dissertation.

- Chapter 2: Literature Review – Reviews relevant literature and prior work in areas that underpin this research: the role of cultural archives in the digital age and digital humanities initiatives; the integration of data science and computational thinking in archival practice; the emergence of Gen-AI and LLMs in information systems; the use of KGs in cultural heritage; use of AT in sociotechnical systems and Gen-AI and ethical considerations of applying Gen-AI to sensitive historical data. By surveying these domains, Chapter 2 identifies the scholarly gap this dissertation addresses: the lack of comprehensive case studies, design and evaluation frameworks for marrying computational education and Gen-AI innovation in

archives [3]. The literature review thus provides the theoretical foundation and justification for my research approach. I should also note that this chapter has been organized to cover the literature review of the overarching research objectives and the need for this research. Wherever applicable, each study's specific prior work and literature review are organized under each study accordingly.

- Chapter 3: Objective I – Empowering Archival Practitioners through Data Science and Computational Thinking – Describes the first study in detail.
- Chapter 4: Objective II – Designing and Evaluating Gen-AI Solutions for Enhanced Access – Describes the Studies 2 and 3 in detail.
- Chapter 5: Objective III – Evaluating and Optimizing Gen-AI for Ethical and Scalable Archival Access – Describes Study 4 and 5 in detail with Study 4 describing a detailed systematic two user evaluation framework grounded on AT framework.
- Chapter 6: Future work and Conclusion – Details the plan for future work in this research area and concludes the thesis.

Chapter 2: Literature Review

2.1 Cultural Archives in the Digital Age and the Need for Enhanced Access

Cultural archives—repositories of historical documents, records, and artifacts significant to a community or society—serve as essential resources for scholarship and public memory. In the digital age, many archives have undergone or are undergoing digitization [23], leading to vast digital collections available for exploration. This transition has aligned archival studies with the field of digital humanities, where computational tools are employed to analyze humanities data. However, the promise of digital archives comes with the challenge of ensuring they are accessible and meaningful to users. Many digitized archives remain underutilized because they are essentially "*dark data*" – data preserved but not effectively unlocked for analysis [3]. The concept of "dark archives" in archival science traditionally refers to holdings that are not publicly accessible [3], but in a broader sense it highlights that simply having data in digital form does not guarantee its discoverability or usability. Still, in a broader sense it highlights that merely having data in digital form does not guarantee its discoverability or usability. Users often lack adequate tools or knowledge to navigate these large corpora. Culturally sensitive archives (such as those involving indigenous knowledge, slavery records, or personal data) pose additional layers of complexity; access must be balanced with context and care to avoid misinterpretation or harm. As Patton notes, AI and computational methods can "*open up access to manuscript collections in ways that OCR did*

for printed texts", indicating the potential for technology to improve accessibility [4] vastly. Yet, without guided approaches, there is a risk that these rich cultural datasets remain "underutilized" – a concern explicitly noted in my problem statement and echoed by other scholars. For example, Richard Marciano's afterword on Archives, Access and AI observes that while AI applications in archives are emerging, there is still "*a lack of compelling case studies*" demonstrating their full potential [3]. This points to a gap in technology and documented practice, where more examples are needed of how digital archives can be made accessible and engaging. Despite digitization efforts, significant barriers persist in accessing culturally sensitive materials. Ethical concerns revolve around the potential for algorithmic bias in automated processing [24]. The LoS collections present unique complexities due to inconsistent 19th-century record-keeping practices and terminology that reflects dehumanizing historical norms [25].

The importance of enhancing access to cultural archives goes beyond academia. It has implications for public history. Archives like the LoS or other African American history collections hold keys to understanding contributions of marginalized groups. Improved access can empower descendant communities and the general public to engage with their history directly. There is an increasing call for archives to be more open and interactive, aligning with movements in participatory heritage and community archives. The LoS collections exemplify this role, documenting the lived experiences of enslaved individuals through legal records, advertisements, and manumission papers. Recent scholarship emphasizes the ethical imperative to center marginalized voices in archival practice [26], particularly in postcolonial contexts where traditional archives often reflect colonial power structures. Digital platforms, if well-designed, can democratize who gets to analyze and interpret archival materials. In sum, literature in archival science and digital humanities converges on the idea that new computational approaches are needed to unlock

archives for broader use [3] [27]. These approaches must handle the scale of data, as many cultural collections now amount to "*big data*" in volume. They must incorporate the cultural context, ensuring that data-driven analyses do not strip records of their meaning or sensitivity [28].

2.2 Data Science and Computational Thinking in Archival Practice

To bridge the gap between archives and accessibility, one line of research and practice has focused on infusing data science and CT into archival work. Computational Thinking, a term popularized by Jeannette Wing (2006), refers to a problem-solving approach that draws on concepts fundamental to computer science, such as abstraction, decomposition, algorithms, and recursion [29]. Wing famously argued that "*computational thinking is a fundamental skill for everyone, not just computer scientists*", akin to reading and writing in its broad utility [29]. In archival science, this means training archivists, librarians, and humanities scholars to think about historical questions in ways that can be addressed with computational methods. There has been a clear trend towards integrating CT into archival education in recent years. For example, Underwood et al. (2019) and Buchanan et al. (2022) discuss how archival studies curricula are evolving to include computational methods, driven by the reality that *"the vast majority of records that will be acquired by archives moving forward are being created computationally"* [5] [30]. Born-digital and digitized records require skills in managing and analyzing data at scale, which traditional archival training did not emphasize.

Studies have highlighted an acute skills gap: a report by the Institute of Museum and Library Services (IMLS) titled "*Shifting to Data Savvy*" identified urgent needs for greater automation and data science skills in archival sciences and libraries [5]. As a result, initiatives like the *LEADING*

program and specialized projects have aimed to "*strengthen digital and computational literacy and training for future librarians and archivists*" [5]. The literature shows consensus that without upskilling in these areas, archival professionals will struggle to appraise, preserve, and provide access to modern collections effectively [10]. In response, Computational Archival Science (CAS) has emerged as a transdisciplinary domain. Marciano et al. (2018) formally articulated CAS as "*the application of computational methods and resources to large-scale records/archives processing, analysis, storage, long-term preservation, and access"*, with the twin aims of improving archival efficiency and understanding how technology changes archival work [30].

Several practical and pedagogical efforts in recent literature illustrate how data science can be brought into the archival realm. One approach is the development of frameworks mapping CT practices to archival tasks. [31] had identified categories of CT for education (data practices, modeling and simulation, etc.), and archival educators have explored analogous practices for archives [30]. For instance, "data practices" in archives might include collecting and cleaning archival datasets, while "modeling" might involve creating a representation of an archive's structure or network of entities. Educators explicitly draw these parallels to make computational steps intuitive for archival problem-solving.

In summary, the literature indicates that building computational thinking and data science capacity among archivists and researchers is a crucial step toward making archives more accessible and analyzable. This is the Prong 1 of my dissertation's approach. Efforts in this space address the human side of the equation: empowering people with the knowledge and tools to use data science on archives. They align with the philosophy that computational thinking should become as common as traditional literacy for those dealing with information resources [29]. However, literature also acknowledges limitations: not everyone will become a programmer or data scientist,

and even trained individuals can be overwhelmed by the size and complexity of archival data. Thus, while education and skill-building are necessary, they may not be sufficient on their own. This sets the stage for Prong 2 – introducing advanced AI tools that can further lower barriers and assist users in exploring archives. The interplay between these prongs is implicit in the literature: In [10], Marciano noted that solving archival challenges requires both "investing in people and technology", where people are trained to use technology responsibly and technology is designed to augment human capabilities [4].

2.3 Gen-AI and LLMs in Information Access and Archives

The past few years have witnessed remarkable advancements in AI, particularly in Gen-AI and LLMs. LLMs are pre-trained language models (often based on deep neural networks like the Transformer architecture [32]) trained on massive text corpora to predict and generate human-like text. OpenAI's GPT-3 [33] and its successor GPT-4 [34] are prominent examples, containing hundreds of billions of parameters and demonstrating unprecedented language understanding and generation capabilities [33] [35]. The arrival of LLM-based systems such as ChatGPT [33] has popularized conversational AI, showing that AI can engage in dialogue, answer complex questions, and produce coherent essays or code. In the context of information access, these models offer a unique possibility: allowing users to retrieve and synthesize information through natural language conversation, rather than through keyword searches or manual lookup. Early adopter domains have included customer service (chatbots replacing FAQ pages [36]) and healthcare (AI assistants summarizing medical literature [37] [38]), and now attention is turning to libraries, archives, and other knowledge repositories. Gen-AI represents an innovative frontier in machine learning,

providing a transformative tool for various sectors, including libraries and archives [39]. Gen-AI models can create new data instances resembling the original data, harnessing patterns within the training dataset [40]. 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. Autoregressive models, like the Transformer model [32] used in OpenAI's GPT models, have been monumental in generating text, offering potential applications in managing and curating textual archives [33].

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 [41]. Concerning the archives field, a recent example is the open-source *WARC-GPT* [42] 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. 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 the contextual information, and thus play a crucial role in cataloging, cross-referencing, and retrieving information efficiently [35]. This has the potential to make digital archives more accessible and user-friendly, promoting broader engagement with historical data. Moreover, LLMs' ability to generate context-appropriate text can be employed to create annotations, descriptions, or synopses of archival documents, enhancing metadata richness and reliability. Academic and professional literature is beginning to explore how LLMs can be applied to digital libraries and archives. One area of focus is using LLMs to improve search and discovery. Traditional search engines in archives rely on keyword matching

and metadata – which can fail when users' queries do not exactly match the terms in the records or when users are not sure how to formulate their query. By contrast, LLMs can interpret a question's intent and generate an answer by drawing on information spread across documents. Nguyen et al. (2024) propose a "smart search in digital archival systems" using a RAG approach [12]. In their framework, an LLM interprets a user's natural language query and a retrieval system fetches relevant documents from the archive, which the LLM then synthesizes into a concise answer. This method leverages the strengths of both information retrieval and AI: the precision of the former and the linguistic fluidity of the latter. Experiments have shown such systems can return more precise and relevant results for complex queries compared to conventional keyword search [12]. A key advantage is handling the "*nuances of user queries*" – LLMs can deal with questions that are phrased in everyday language or that require combining data points (e.g., "*Find all instances where a person traveled from Town A to Town B in these records*") which might stump a regular search interface [12].

Another line of research looks at LLMs for tasks like information extraction and summarization in archives. For instance, there is work on using LLMs to automatically identify and redact sensitive information (like Personally Identifiable Information) in archival documents [43]. Also, LLMs have been tested for transcribing and summarizing historical documents. Blevins (2024) notes how off-the-shelf LLMs can assist in transcribing handwritten letters and then explaining their content, something traditionally requiring bespoke tools or expert effort [44]. These examples indicate that Gen-AII is becoming versatile in handling various aspects of archival content, effectively acting as an intermediary between raw data and user-friendly information. Libraries and archives are cautiously experimenting with such AI. Nawaz and Saldeen (2020) reviewed AI chatbots in academic libraries, finding that these chatbots are increasingly adopted to assist users in

reference services and that natural conversation is a powerful mode for information delivery [13]. Users appreciate asking questions in plain language and getting immediate answers. This can significantly benefit those not skilled in database querying or who might not know which finding aid or catalog entry holds the answer. [45] showcases using GPT-4o to create data visualizations from prompts.

Despite these opportunities, the literature also strikes a note of caution. Gen-AI systems have well-documented issues: they can produce factually incorrect statements with a confident tone (a phenomenon often termed ”hallucination” in LLMs) [46]. This is a serious concern in archives, where factual accuracy is paramount. Additionally, LLMs lack inherent understanding of context that is not in their training data. For example, a general model like GPT-3 might not know specifics about a niche archival collection unless it was included in its input prompt, referenced, or fine-tuned. This has led to strategies like fine-tuning LLMs on archival data, performing special retrieval techniques, or using KG augmentations (as I pursue in my Study 3 and 5 respectively). Another challenge is ensuring that the AI can cite sources or explain how it derived an answer – something users in scholarly environments expect. Current LLMs do not natively provide citations, which means additional layers are needed to make them act transparently. The literature on explainable AI and human-AI interaction emphasizes that for tools like chatbots to be trusted in libraries/archives, they should ideally point to the documents or records that back their answers [4].

In conclusion, Gen-AI and LLMs promise to transform archival access (Prong 2 of my research). They offer solutions to the problem of user engagement: rather than learning SQL or browsing dozens of files, a researcher could have a conversation with the archive. This resonates with the broader trend of AI in GLAM (Galleries, Libraries, Archives, Museums) where professionals like Patton argue that *“GLAMs must embrace these emerging tools to remain*

relevant" [4]. My literature review suggests that the time is ripe to explore such applications; however, it also underscores the need to do so carefully, addressing accuracy, transparency, and alignment with archival principles. The Gen-AI approach should therefore be seen not as a replacement for human expertise or traditional tools, but as a powerful augmentation.

2.4 Role of Knowledge Graphs in Generative AI for Cultural Heritage

A knowledge graph (KG) is a structured representation of facts where entities (nodes) are connected by relationships (edges), often following ontologies or schemas. KGs have been used in cultural heritage contexts to interlink data across collections and provide rich context for items. For example, the Heritage Connector project built a KG linking museum collection records with external sources like Wikidata, which "*allowed visualization and exploration of collections in entirely new ways*" by highlighting connections between objects, people, and concepts that were not obvious before [4]. Such interlinking transforms isolated records into a network of knowledge, enabling users to traverse from one entity to related entities (e.g., an artist to their artworks across museums, or from a historical person to all archival records mentioning them). Another critical aspect of the accuracy and correctness of the responses by Gen-AI tools using unique retrieval mechanisms purely depends upon the domain knowledge of information that's shared with the LLMs. Many researchers and scientists believe that the data supplied to these LLMs should be prepared and pre-processed to be made *"AI-Ready"* as indicated by [16]. Greenberg (2024) highlights the importance of incorporating metadata and ontologies into AI-ready data frameworks and demonstrates how AI methods, including Gen-AI, can leverage metadata and ontologies to improve and validate data, suggesting that building KGs of datasets from a single corpora, like the

LoS, could enhance the data's AI readiness by providing structured semantic representations and facilitating consistent data integration and interoperability. KGs serve the purpose of getting the data "*AI-Ready*." Additionally, in digital archives, KGs can address the limitation of "*item-centric*" views. Traditional archive databases show search results as lists of items with metadata [17]. In contrast, a KG-driven interface can show how items relate (chronologically, geographically, by topic, etc.), supporting users in browsing related items and synthesizing narratives [17]. Khoo et al. (2024) demonstrated this by developing a visualization interface for archives based on a KG; users could click on a person in one record and immediately see other records connected to that person, or follow links to events and places related to them [17]. The user study indicated that this helped in forming a more holistic understanding of the content, as users were not confined to one record at a time [17]. KGs have emerged as a potential solution to provide LLMs with structured context and reasoning pathways beyond simple text matches. For example, [**?**]'s approach showed that incorporating an LLM-generated KG can substantially improve question-answering on private datasets, enabling the LLM to ground its answers in the graph and even provide provenance links to sources [**?**]. Similarly, in the cultural heritage domain, KGs have been used to interlink records and reveal hidden connections [**?**]. These successes suggest that a knowledge graph-enhanced RAG (KG-RAG) could address my use case's needs: by unifying multiple LoS datasets in a graph, we can help the LLM traverse cross-collection links (e.g. person-to-record relationships) and handle quantitative or constraint-based queries via cypher graph queries, while grounding answers in a network of verifiable facts.

For my research, the interest in KGs is two-fold. First, as a tool for enriching archival data: by extracting entities (people, places, dates) from archival records and connecting them, I create a semantic layer that AI can use. Several published studies have shown the use of KGs in their

problems in combination with LLMs to successfully get better results than with just the use of LLMs and the retrieval techniques [47] [48] [49] [50] [51]. Second, KGs are central to addressing the context deficit of LLMs. An LLM generates answers from patterns in text, but it doesn't "*know*" a fact in a robust way that it can cite or justify. If I integrate an LLM with a KG, it can retrieve relevant nodes or subgraphs to ground its responses. The literature on retrieval-augmented LLMs suggests they can significantly reduce hallucination and improve factual accuracy by anchoring generation on retrieved information [12].

Regarding cultural heritage, KGs also embody respect for provenance and multi-perspectiveness. Because they can store different types of relationships and even contradictory information, they are suited to representing complex historical realities (for example, an individual might be recorded under different names in different documents; a KG can connect those names to one identity with a "same as" relationship). By using KGs, archivists can ensure that an AI referencing the KG has access to these nuances, rather than relying on a homogenized model. One challenge noted in literature is building the graph in the first place – it can be labor-intensive to curate, or error-prone if fully automated. However, projects have succeeded in semi-automated graph construction from archival texts [17]. There are also domain-specific knowledge bases (like people directories and gazetteers of historical places) that can be incorporated. Overall, KGs are a promising solution for digital heritage systems to move beyond basic search [17] and produce optimized Gen-AI tools. They complement AI by providing structured knowledge that AI can leverage.

In summary, my system builds upon these lines of research by combining a KG of historical LoS datasets with an LLM in a RAG framework. To my knowledge, this is one of the first applications of KG-RAG in this context.

2.5 Employing Activity Theory for Systematic User Evaluation

Archival research and digital library use are inherently socio-technical activities, making "Activity Theory (AT)" an apt lens to study and design archival interfaces. Researchers in information science have highlighted AT's potential to provide a holistic context for information practices. Wilson (2006) [52], for example, argued that studying information-seeking and use through AT allows deeper understanding of context than traditional user behavior models, by discovering contradictions that affect information practices. In an archival setting, this means *I can examine not just how a user searches a catalog, but why certain searches are difficult* – perhaps due to a misfit between the archive's metadata schema and the user's mental model (a contradiction between tool and object), or due to institutional rules that restrict access to certain records (a rules vs. object tension). AT can also serve as an *"overarching paradigm"* to bridge different problem areas in information studies – for instance, linking issues of information seeking (user perspective) with issues of information organization/retrieval (system perspective). By treating the entire interaction as one activity system, we avoid siloed solutions and instead seek interventions that address systemic tensions. In the broader realm of human–computer interaction (HCI) and user experience, AT has been a foundational framework for moving beyond interface mechanics to understanding contextual user interaction. From a user interaction standpoint, AT prompts designers to ask: What is the real activity the user is trying to accomplish, and how does the design support or hinder it? This goes beyond usability at the micro level and into the realm of usefulness and integration in the user's life or work. From a user interaction standpoint, AT prompts designers to ask: What is the real activity the user is trying to accomplish, and how does the design support or hinder it? This goes beyond usability at the micro level and into the realm

of usefulness and integration in the user's life or work. Good design must accommodate varying contexts. If a contradiction exists – say the interface is optimized for one type of user's activity but not the other – AT would highlight that as a design issue.

AT has a rich history of use in studying sociotechnical systems [53] [54], where technology and human social factors are deeply intertwined. It is fundamentally a sociotechnical framework: rather than analyzing technology in isolation, AT considers how technological tools become part of human activity systems, mediating work and reshaping social practice. This approach aligns with broader sociotechnical perspectives that emphasize the co-evolution of social and technical elements in any system of work or organization [22] [55]. Notably, AT has been applied to contexts of AI in workplaces and organizations. For example, Allen et al. (2022) [55] use AT to investigate how AI systems become "tools" within work activity and how employees perceive AI's role in their jobs. In their study, AI systems (e.g. algorithms, automation tools) are conceptualized as mediating artifacts that workers use to achieve organizational objectives. By structuring interviews and analysis around AT components (subject, tools, object, rules, community, division of labor), they identified key conflicts and contradictions arising from AI adoption. A major finding was that employees recognized AI's potential to enhance efficiency, but also felt tensions – for instance, new AI tools introduced disturbances or uncertainties in established work practices and norms. These tensions echo what AT literature calls *knotworking* or *disturbances* – misalignments that "*interfere with the realization of individuals' and communities' goals*". By highlighting such contradictions, AT-based analysis in sociotechnical systems reveals where an intervention (like AI) may require adjustments in rules (e.g. new policies), redefinition of roles (division of labor between humans and AI), or additional training (subjects acquiring new skills) to fully integrate the technology. This perspective is valuable for ensuring ethical and effective AI adoption.

Moreover, AT's emphasis on learning and development is particularly relevant when introducing AI into existing social systems. As Jarzabkowski (2003) [56] noted, AT is *"essentially a learning theory"* that helps explain how processes evolve. In the context of AI, this can mean observing how workers learn to collaborate with AI tools or how organizational practices expand to accommodate AI capabilities. Indeed, recent work has explicitly used AT to frame human–AI collaboration and future-of-work questions. For instance, a 2023 study by Nah et al. [57] deploys AT to explore generative AI's impact on work practices and poses research questions about how AI and humans can collaborate effectively, how AI might shift roles, and what new rules or safeguards are needed. By situating AI within the activity system, researchers can ask: *Who is the subject when AI tools take on agentic roles? How is the division of labor reconfigured when an AI performs tasks previously done by a human? What rules (ethical guidelines, oversight mechanisms) must evolve to govern this new partnership?* AT provides a structured way to analyze these questions, ensuring we account for both technical and social dimensions in equal measure.

In modern digital environments, especially those incorporating AI, the user experience often involves a conversation or interaction loop with the system. Here, AT can frame the Gen-AI tool or system as not just a passive tool but an actor-like tool that the user interacts with. Some have even considered Gen-AI as a quasi-community member or at least a very advanced tool in the activity system (though not an independent subject since it lacks its own motive). Regardless, the mediated nature of the interaction means issues like trust, transparency, and control become central – these can be seen as rules governing the user-AI interaction. Recent studies of human-AI interaction through AT suggest examining how rules (e.g. user expectations of explanations, organizational policies on Gen-AI use) and division of labor (what the Gen-AI tool does vs. the human does) need to be negotiated for a smooth user experience. If the Gen-AI tool takes too

much autonomy, users may feel a loss of agency (a contradiction between the subject's desired control and the tool's operation). If the Gen-AI tool is too passive, it may not add value (failing to resolve the initial contradiction it was meant to address). To sum up, in designing and evaluating user interactions in digital environments, AT provides a rigorous vocabulary to capture contextual factors. It ensures our focus remains on activities and outcomes, not just on interface features. Throughout this dissertation where the new Gen-AI tool- ChatLoS versions are dealt with, user studies and design reflections are informed by AT: I consider the motive behind user behaviors, the contextual constraints they face, and the systemic effects of introducing new interactions (like natural language querying or multi-step AI-guided analysis). This approach leads to more robust insights than a narrow usability focus, and it guides me in refining the tools so they truly empower users in their archival research activities. This style of user evaluation involving Gen-AI tools in an activity system like the MSA's LoS hasn't been conducted before which is identified as a research gap. In Study 4, user studies and design reflections are informed by AT: I consider the motive behind user behaviors, the contextual constraints they face, and the systemic effects of introducing new interactions through Gen-AI tools like ChatLoS. This approach leads to more robust insights than a narrow usability focus, and it guides me in refining the tools so they truly empower users in their archival research activities.

2.6 Ethical and Societal Considerations of AI in Archives

Any application of AI to cultural archives must grapple with a set of ethical considerations. Historical archives are not just datasets; they are the raw materials of people's history and identity, sometimes containing painful or sensitive information (such as records of enslaved individuals,

war crimes, personal letters, etc.). Introducing AI into this domain raises questions about bias, accuracy, transparency, and the impact on both users and the communities represented in the data.

One primary concern identified in the AI ethics literature is the bias in AI models. LLMs trained on Internet-scale data are known to pick up and amplify biases in their training text [9]. Bender et al. (2021) caution that LLMs "*encode and reinforce hegemonic biases*", often reflecting dominant viewpoints and potentially marginalizing others [9]. In an archival context, this could manifest in an AI system that, for example, downplays the agency of historical subjects from oppressed groups or uses inappropriate language when summarizing records about them. Additionally, if the archival training data is skewed (many archives traditionally emphasized records of elites over everyday people), an AI might reinforce that skew unless countermeasures are in place. The concept of "*algorithmic bias*" thus extends to historical analysis: I must ensure the AI tools do not perpetuate historical injustices or erasures. Strategies to mitigate bias include thorough evaluation (as done in Studies 2, 3, and 4), incorporating diverse perspectives (via KGs in Study 4 or community input as future work), and implementing filters or checks for sensitive content (also explored in Study 5) [9].

Another ethical issue is accuracy and truthfulness. Archives are often used to establish facts (for legal cases, genealogies, etc.), so an AI that hallucinates or provides incorrect information can have serious repercussions. The literature notes that hallucination in LLMs is a persistent problem where models "*produce outputs that are coherent... but factually incorrect or nonsensical*" [46]. In my domain, a hallucination might be an AI-generated archival quote or statistic that is entirely fabricated. This is unacceptable for scholarly or public history uses. Therefore, researchers emphasize the need for transparency and verification in AI-assisted archival research. [4] suggests that GLAM institutions use AI in a way that they can "*document their decision-making processes,*

data selection, and tool usage to mitigate biases" and maintain trust [4]. This means, for example, logging which records were retrieved to answer a query and perhaps showing those to the user as evidence. This is identified in my research as part of the future work.

Finally, there is the question of impact on professionals and communities. The literature generally frames AI as a replacement tool that frees professionals to focus on higher-level tasks [4]. However, to ensure this, archivists and librarians need to be involved in developing and deploying AI tools – hence the significance of Prong 1 in my study (so they understand what the AI does and can curate and correct it). From the user community side, ethical use of AI means ensuring these tools genuinely serve users' needs and do not create new forms of exclusion. For example, if an AI interface is only in English, it might exclude non-English-speaking community members from engaging with archives of their heritage; thus, multilingual support could be seen as an ethical imperative for inclusive design [12]. The literature underscores that responsible AI for archives must incorporate fairness, accountability, transparency, and ethics-by-design. My dissertation reflects these priorities. Moreover, the future work I envision aligns with ethical scholarship: involving community stakeholders to guide AI development ensures that the resulting tools are not just technically sound, but also culturally respectful and aligned with the values of those represented in the archives. By reviewing these considerations here, I establish the guiding principles that inform my research's design decisions and evaluative criteria.

2.7 Summary and Research Gaps

This literature review has examined the interdisciplinary landscape encompassing archival science, data science education, Gen-AI, KGs, and AI ethics. The following key points have

emerged: Cultural archives are increasingly digital and large-scale, offering rich opportunities for computational analysis but facing challenges in accessibility and user engagement [3] [27]. Traditional methods alone are insufficient to tackle the volume and complexity of these collections.

- Computational thinking and data skills are now essential for modern archival work. There is a documented skills gap among information professionals, and integrating CT into archival education is a strategic priority [5].

- Gen-AI and LLMs represent a frontier for enabling natural language access to archives, potentially transforming user experience by allowing conversational querying and automated synthesis of information [12] [13]. However, these systems must be adapted to the domain and come with caveats regarding accuracy and bias [9] [46].

- KGs offer a means to inject domain knowledge and context into AI systems and provide users with novel ways to explore archival relationships [4] [17]. They could act as a bridge between unstructured archival data and structured queries/knowledge.

- Application of AT based systematic user evaluation in designing and assessing use of Gen-AI tools like ChatLoS in a cultural archives space.

- Ethical considerations are paramount: any solution must mitigate AI biases, prevent misinformation, respect privacy, and involve stakeholders to ensure the tools align with human values and social good [4] [9].

Through this review, I identify a clear research gap at the intersection of these areas. [3] pointed out the lack of compelling case studies combining AI and archives [3] – this dissertation seeks to fill that void by providing several detailed case studies. Specifically, no comprehensive

research follows a project from teaching computational skills on an archival collection all the way to building an AI chatbot for that collection and evaluating one against the other. My work therefore contributes novel insights into how educating users and augmenting systems with AI can work in tandem to enhance archival access. The literature also suggests that experiments with AI in archives are in early stages, often focusing on technical feasibility. This dissertation pushes the boundary by critically assessing user interaction and effectiveness of Gen-AI tools versus traditional methods in a real-world archival scenario, thereby addressing an important gap in systematic evaluation research. In conclusion, the reviewed literature informs my research questions and methodology, reinforcing the importance of a dual strategy: empowering the people and improving the technology. The subsequent chapters will build on these foundations, each contributing to filling the gaps identified.

Chapter 3: Objective 1 – Empowering Archival Practitioners through Data Science and Computational Thinking

3.1 Introduction

Archives and libraries today hold vast digital collections but remain challenged by ever-growing data and the need to equip students and researchers with computational skills. As highlighted in the literature review (e.g., [10] [31]), many archival professionals and humanities scholars have limited training in data science and programming. This gap hampers effective use of large-scale digital archives such as the LoS collections. In [58], the authors stress on the importance of the disruption taken place in the Archival world with the emergence of groundbreaking computing technologies that are now used in creation, sharing and storage of artifacts, thus making it even more necessary for aspiring archivists to learn and be equipped with the understanding of these new technologies. In [30], Underwood et al., argue that it is vital for students aspiring to work with archives to understand the computational terms and develop their computational skills to meet patrons' needs better, help with data analysis, and address any issues or concerns that would arise. As part of the Research Objective I, in this Study 1, I address these gaps mentioned above as a first prong of the overall research strategy by designing *interactive Digital Computational Notebooks (iDCNs)*, a set of cloud-hosted learning tools that blend narrative

text, code, and real archival datasets to teach basic data science in an archival context. This study was presented, peer-reviewed and published as a full paper at the International Symposium on Grids and Clouds, Taiwan in October 2021 [11]. The dataset I chose for this study is the CoF dataset from the LoS, which records key demographic and historical details of formerly enslaved individuals in Maryland between approximately 1806 and 1864. The overarching goals of this study are:

- To develop hands-on learning modules that guide non-technical users in data-cleaning, exploration, and visualization of an archival dataset.
- To integrate CT practices [31] such as data manipulation, modeling, and systems thinking into archival education.
- Finally to assess the potential of iDCNs to enhance archival data literacy among students and archival professionals.

The iDCNs were created using an open-source cloud-based web application tool called *Jupyter Notebooks* (JNs),[1] with the computational programming language *Python.*[2] These iDCNs get their name also because they are tailored to teach computer programming through a step-by-step data science driven analytical treatment of a dataset collection for analysis and visualization purposes. Such a case study has not been conducted yet on a culturally rich dataset like the CoF. These iDCNs would form a new learning environment which could become part of courses taught for non-STEM higher education students across the country, including archival sciences. To evaluate the acceptance of this novel educational tool among the students, and educators, after the iDCNs

[1] Jupyter Notebook - https://jupyter-notebook.readthedocs.io/en/stable/notebook.html.
[2] Python Documentation - https://www.Python.org/doc/.

were created and tested, a user survey was distributed to current students and educators from these programs and the responses were studied and understood for future improvements to this unique learning tool. The rest of this study is divided into the following sections: Literature Review, Research Problem and Questions, Approach, Results, Discussion, Future Work and Conclusion.

3.2 Literature Review

A literature review identifies several research studies that use JNs as a teaching tool to teach programming in higher education. In [59], Davies et al., created a set of JNs for bioscience and informatics education. The paper details the anatomy of the JNs with explanations and pictorial representations on how to set up the infrastructure and the environment for performing basic computational commands using them. The authors convey the benefits of using JNs as a user-friendly collaboration tool, encouraging reproducibility and allowing the sharing of the modules, and explaining how educators could also use them as an assessment tool. The study also performs case studies to teach basic programming and uses coding commands in tutorial lessons, modified to suit bioscience education's domain. Unlike my research, their project does not seem to conduct a data-driven analysis of a real dataset. However, it publishes a survey of experiences gathered from the students as part of a 6-week course that teaches them the fundamentals of programming using JNs. In another study [60], Reades conducts an evaluative research on how JNs performed in the teaching of programming to undergraduate students from geography, another STEM background subject. This study explains the creation of three JN modules called 'geocomputational' modules and evaluated the JNs based on factors such as minimal complexity, maximal flexibility, interactivity, utility, and maintainability. The authors

published a detailed comparison of pros and cons of using JNs for teaching programming to their undergraduate students. A similar study was conducted by [61] to prepare teaching materials for electronics undergraduates through JNs. The authors created modules to teach the subject *Digital Signal Processing* to the students using these modules. The authors used the *Python* programming language to teach these STEM background students. The paper talks about the growth of JNs as an increasingly accepted educational tool and how the supporting software libraries available for *Python* language makes it an attractive option for students and educators alike to incline towards these open-source resources.

There is one notable archival study [62] which uses JNs in the Archives domain, a non-STEM background. The authors highlight the ability of JNs to "*see exactly the lines of code that have produced the data, aid transparency and supply the necessary contextual information for meaningful browsing.*" They discuss the development of JNs with another software to enable archivists to explore, investigate and analyse the metadata in the Netherlands Institute for Sound and Vision (NISV) archive. The challenges faced by the authors in the development of these JNs are discussed. My study differs from this work in how it introduces JNs to non-STEM students and its step-by-step data science educational focus. The iDCNs created by my project not only use the robust infrastructure already offered by the JN project, but also follow a set of proven practices in performing a contextual data science driven approach on a culturally rich dataset collection. This is an intentional decision to empower the scholars in archival domains with the much-needed CT capabilities necessary to perform computational operations on the archived datasets, thereby enhancing access to them.

In prior collaborative work [20], I conducted preliminary data science driven approaches to the MSA's dataset collections through a detailed study on how to derive "*data background*"

from the contextual analysis of data with collaborations with experts from different disciplines. I used several tools, some of which were open source like *Open Refine,*[3] *R* [4] for data exploration and manipulation, and other industrial tools like *Tableau,*[5] and *Neo4j* [6] for data visualization purposes. I documented detailed results from these analyses of MSA datasets, which also included the cross-collection study on trying to link data between the two datasets. This current study deals with a data-driven approach of one of these MSA datasets but in a different infrastructure. The *JupyterLab* [7] environment powered with *Python* programming language's support for software libraries allows me to perform most if not all of these operations in a homogenous environment unlike the prior study where different tools were used for various data operations. This was one of the important reasons for choosing JNs for developing iDCNs, as switching between different tools could quickly distract beginners in programming from non-STEM backgrounds.

Weintrop et al., in [31] developed on the views of [63] and presented their framework for applying CT practices to mathematics and science education. The four practices identified in the study are: data, modeling & simulation, computational problem solving, and systems thinking. The authors formulated these practices from detailed experimentation and external reviews. As mathematics and science education have been increasingly intertwined with the push of computing, these practices were created to enhance instruction and learning [31]. The authors presented their views and discussed a framework for applying CT practices to mathematics and science problems in K-12 schools. The practices aligned well with the objective of this study, and I have attempted to emulate them and document how they applied to the development of iDCNs. I needed to follow

[3]Open Refine - https://openrefine.org/documentation.html.
[4]R - https://www.r-project.org/other-docs.html.
[5]Tableau Documentation - https://help.tableau.com/current/pro/desktop/en-us/gettingstarted_overview.htm.
[6]Neo4j Documentation - https://neo4j.com/docs/.
[7]JupyterLab Documentation - https://jupyterlab.readthedocs.io/en/stable/.

a well-established best practices framework to make developing these iDCNs easy for me and keep them logically arranged.

3.3 Research Problems and Questions

There is increasing demand for educators and students to stay abreast of learning and teaching these new technologies. In addition, it would be vital to understand the computational concepts behind the generation and dissemination of these digital artefacts. To equip themselves with this skill and to address the increasing demand for distance learning, course content and lesson plans appropriate to suffice these requirements need to be created. To solve these problems in my project, I formulated the following research questions:

- RQ1A: Can a novel set of cloud-based iDCNs be developed with content used as a learning tool to teach computing for higher education students from non computational backgrounds?
- RQ1B: Can the MSA's digitized CoF dataset collection be used to create the computational content as a case study through a step-by-step contextual data science driven analytical treatment?
- RQ1C: Can the CT practices introduced in [31] by Weintrop et al., be extended and applied to non-STEM backgrounds in creating the computational content?
- RQ1D: Can the resulting novel learning environment be useful for the Library and Archival Science educators and students? How would they react to such changes to their pedagogical learning environment?

3.4 Methodology & Results

To address the research questions, I divided the study into two main parts: (1) an exploratory case study leading to the creation of a set of iDCNs using an open source tool, and JNs following CT practices from Weintrop's Taxonomy [31], and (2) a user survey to gather feedback from the end users of this educational tool (students and educators with Library and Archival Science backgrounds).

3.4.1 Creation of iDCNs - An exploratory case study

To address RQ1A and RQ1B, I used the CoF dataset, to build the iDCNs. The idea to use the CoF dataset collection was inspired from the long-standing efforts by the MSA staff on several projects, including the Faces of Freedom Project [64]. MSA's Projects, in addition to digitizing the physical documents like CoF, have gone as far as turning data into people through stories and documenting African American enslaved people's lives and experiences as described in [65]. In the Faces of Freedom Project [64], the MSA approached a criminal forensic detective to use the text description identified from the CoF dataset collection to create a visual image of an enslaved, later freed, person named Lot Bell. This unique rendition of textual content transformed into a digital image inspired me to use this dataset to unravel stories from the hidden dataset through data science and analysis. Papadakis et al., in [66] used a similar exploratory case study research method to perform research on the effect of using ScratchJr to teach basic programming concepts to pre-kindergarten students at a school in Greece. The exploratory case study method was well suited for this part because it allowed me to focus on the historically and culturally rich dataset features or columns available from this specific dataset. As these data features hold tremendous

historical value and possess human sensitivity, I worked closely with historical subject-matter experts and historical resources before manipulating the data.

3.4.1.1 Why the *Jupyter Notebook* (JN)?

Jupyter Notebook (JN) is an open source web application, created in 2014, which enables software developers and researchers to design, create, collaborate and publish their creative work in the form of narrative text like stories coexisting with computational coding snippets that perform specific tasks along with pictures and other media files. This web application is an outcome of *Project Jupyter*,[8], a non-profit open-source project. In addition to the central product, the Project community also has developed and implemented other products, namely *JupyterLab*, which is an integrated development environment. The *JupyterLab* provides a user interface that allows developers to access resources like datasets, files, and use them across multiple digital JNs created in the environment. I used *JupyterLab* to develop these notebook modules in this study. Fully developed and tested JNs could be implemented as narrative HTML markdown files. In [59], Davies et al., have explained the anatomy of the JNs and they provide several training resources [67] for beginners to advanced users. An important feature of JNs is the ability to use them as a live editor to modify elements like Text and use different Markdown styling options to make the content appear as paragraphs with headings, indentations, bulleting and other text formatting. In [68], Smith details the usage of Markdown in developing JNs.

In [67], the authors discuss the benefits of using JNs in teaching and learning pedagogical processes. The authors indicate that incorporating JNs benefits the course from increased participation, engagement, and real-world understanding of the subject matter. They also point out

[8]Project Jupyter - https://jupyter.org/

that students could benefit from CT, active learning from wherever they are and whatever time they want to, as the resources are free, open source, and cloud-based. Instructors and educators could leverage these unique digital resources in several ways, notably as a learning material, as demonstrative lectures, to perform online laboratory exercises, and to conduct exit ticket quizzes. The authors also point out that the JNs are becoming popular in STEM areas and in Digital Humanities, Social Sciences, Writing, Music, and Introduction to Programming courses. JNs allow developers to include coding commands from many computing programming languages like *Python*, *R*, however, I have used *Python* as it had supporting libraries for running a full suite of data analytical methods like exploration, manipulation, analyzing and visualization of data. In particular, it provides libraries that allow developers to create interactive visualizations for creating networks from data, georeference areas by counties in the USA, create word clouds from textual data, and plot interactive charts and graphs on the cleaned data.

3.4.2 Following CT Framework

To answer RQ1C, the approach is to adhere to the CT practices proposed by Weintrop et al., as shown in Figure 3.1 and capture the list of CT practices that were followed during the development of the iDCNs in this study, and justifying how these steps relate to the practices, so the results could be extrapolated for further adoption of these CT practices in other areas to solve complex problems.

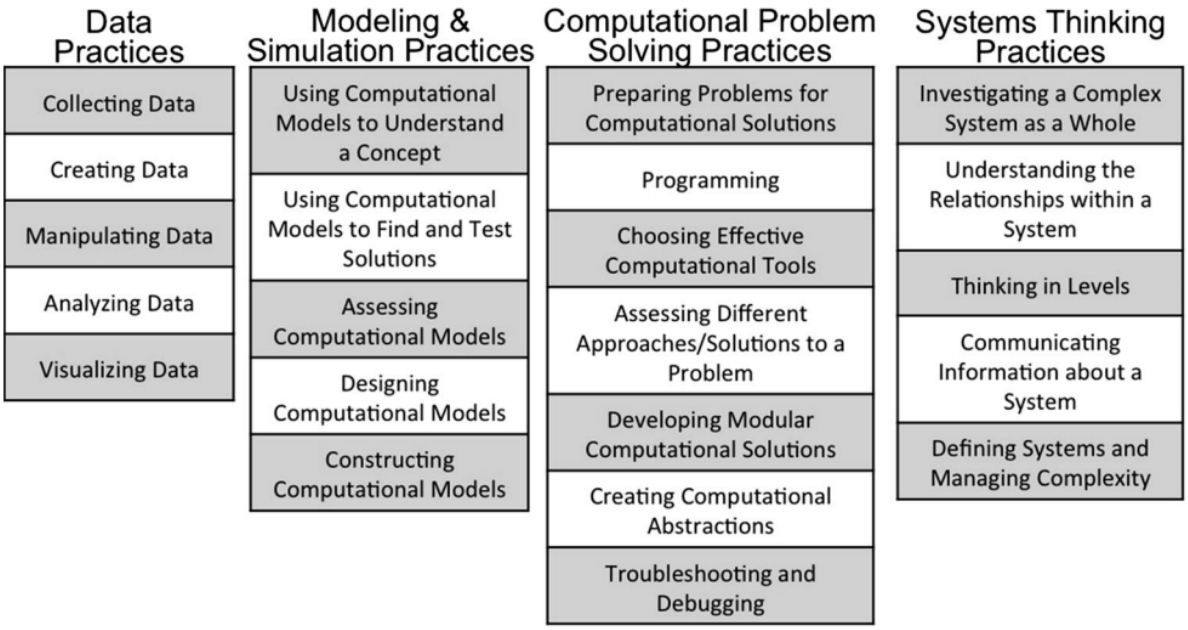


Figure 3.1: Computational Thinking (CT) Practices Taxonomy

3.4.3 User Survey

To address RQ1D, I performed an IRB-approved user survey study with graduate students and educators collaborating in a non-STEM Library and Information studies course at two higher education institutions in the USA. In [69] Sun and Oza proved that using a User Survey research method effectively gauges the understanding and issues in using a collaborative online system among 260 users. As these new cloud-based iDCNs would also be collaborative in nature and accessed online, I decided to use this method to gather feedback on the tool. The results are documented under the User Survey Results section.

3.5 Results

We had two outcomes from the project's two main parts: the iDCN learning modules and the user survey results.

3.5.1 Development of iDCNs

We arranged the new iDCNs into four learning modules as follows:

- *Index Notebook*
- *Contextual Data Analysis and Manipulation - Part 1 Module*
- *Contextual Data Analysis and Manipulation - Part 2 Module*
- *Contextual Data Visualization Module*

Using *JupyterLab* as the development environment, and *GitHub,* [9] a cloud-based version control tool, I was able to update the version control files. Once merged with the master branch, the changes were instantly available. They were picked up at *https://cases.umd.edu*[10] which hosts a public cloud-based infrastructure for demonstration, sharing, and collaboration of these learning modules.

3.5.1.1 Index Notebook

As a starting point to this novel learning environment, the index notebook module was designed in such a way that it conveys two things. One, through narrative text introduces the

[9] GitHub Documentation - https://docs.github.com/en

[10] Legacy Of Slavery iDCN Project - https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb.

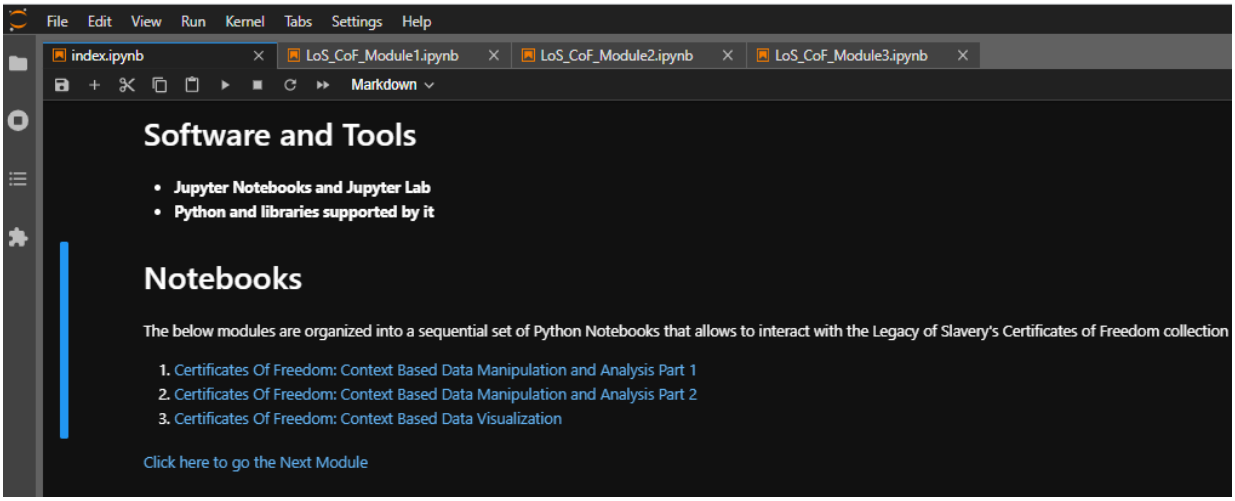


Figure 3.2: Screenshot showing a sample of Markdown edited text with Hyperlinks

LoS Project at MSA, from which the CoF dataset was utilized for this study. Another topic is to show the various formatting options available on these iDCNs using Markdown text editing such as Title, first, second and third level headings, creating a paragraph structure with bullet points, highlighting, bolding and italicizing abilities, options to provide hyperlinks to external resources, attaching images in line with the text, and other basic text editing features. Figure. 3.2 shows a picture of the index module's screenshot. One could click the next module or choose a desired link from the available modules as shown in Figure. 3.2. This module's theoretical text and pictorial representation was designed not to overwhelm non-STEM users with computational coding commands up-front, but to provide a preface to the upcoming modules.

3.5.1.2 Contextual Data Analysis and Manipulation - Part 1 Module

Continuing from the index module, I created the data science modules starting with the Part 1 module. According to the CT framework, *data practices* include data creation, collection, manipulation, analysis and visualization. As the CoF dataset was digitally created from the physical scanned documents and data was collected by the MSA in their LoS database, the first

two components of the *data practices* were already satisfied before this project. Using traditional database access tools, the MSA's LoS database was accessed to extract the CoF dataset in a commonly used comma-delimited file format called *CSV*.[11] From [31], data manipulation activities include sorting, filtering, cleaning, normalizing datasets, and data analysis pertains to looking for patterns or anomalies, defining rules to categorize data, identifying trends and correlations. I created the iDCN module showing how to perform these data operations on the CoF dataset using the *Python* programming language. Also, the activities involved with these two CT practices appear to be closely related as manipulation operations result in analysis which could lead to further manipulation and analysis. Hence, I decided to combine them in this module. By mastering these two practices, Weintrop claims that non-STEM students could work with "*big data*", thereby equipping them to bring any large datasets into a desired set up or configuration and help them make confident claims on the analysis they make.

As the Introduction section discusses, the CoF dataset has several culturally rich features. I chose to operate on four of these: CoF Issue Date, Enslaved person's Prior Status before issuance of the CoF, Enslaved person's Height, and Age. This decision was made for data manipulation and analysis due to the features' categorical and quantitative nature. As coding commands start to appear from this step, human interpretable comments were added to these commands to help the students understand their high-level overview and functionality. The iDCN module was split into two parts to keep the data operations demonstrable and modular. Part 1 processed CoF Issue Date, Enslaved person's Prior Status features, and Part 2 took care of Enslaved person's Height, and Age features. Figure. 3.3 shows a screenshot image of the Part 1 module notebook which shows code commands to import software libraries accessed by the *Python* programming language to create

[11] CSV reading in Python - https://docs.python.org/3/library/csv.html.

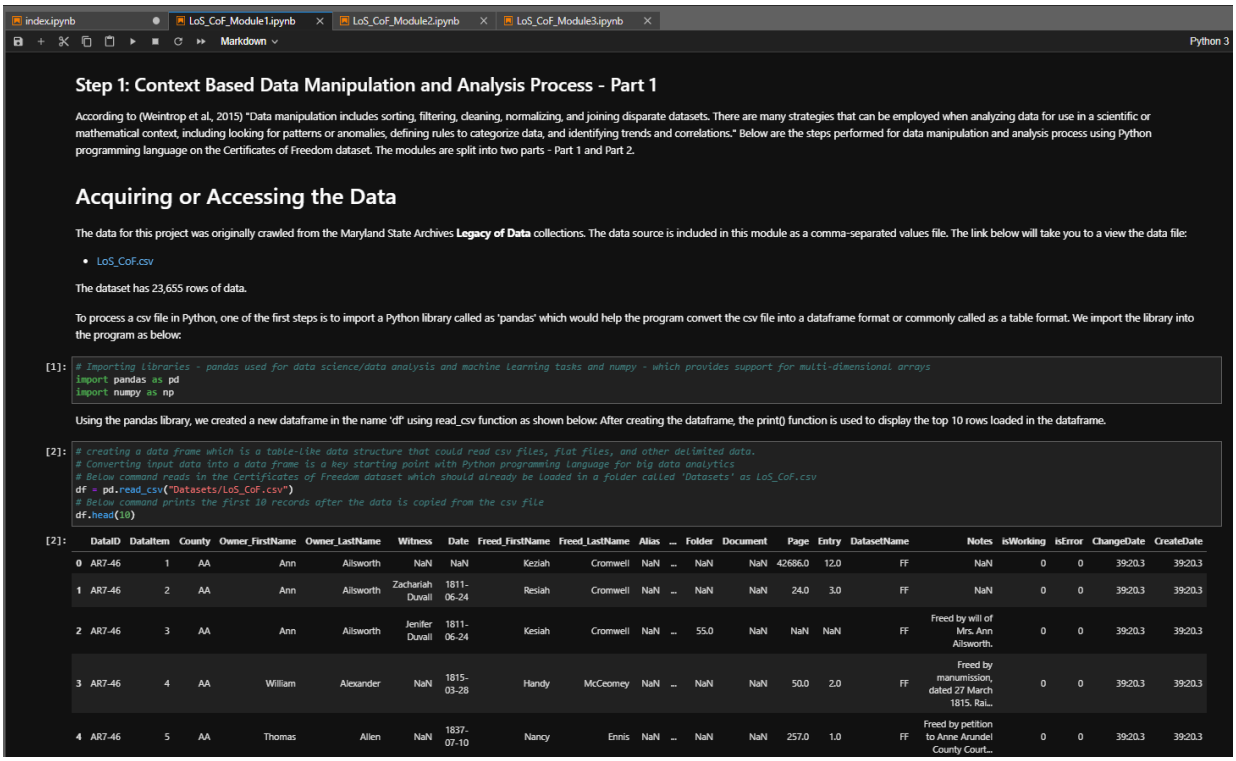


Figure 3.3: Screenshot showing Part 1 module with *Python* code snippet to import csv and read first few records

dynamic data structures that would be used further for data operation. The module begins with importing the CoF dataset uploaded to the project's Dataset folder as shown in Figure. 3.3. The dataset is then stored in *Python* programming memory into data structures called *'data frames'* similar to database tables. Further commands are run to display the first 10 rows of the data frame stored, then the focus shifts to the CoF Issue Date feature. Several data manipulations are run to first filter the unique formats and values, then to clean the date values into a standard format and substitute incorrect date values to a code, 'NaT', which could be parsed in the next step to isolate the erroneous records for further analysis or to perform corrections at the source data by sharing it with the MSA. Upon analysing the resulting values of error records, I could point out interesting patterns and anomalies with the data collected by the MSA for this date feature. In Figure. 3.4, of the 657 erroneous date records, two with a length of 6 digits were filtered for further analysis.

```
# Below command displays the specific records that was identified as erroneously entered. The inner command 'df[
# in-built function to filter the records that match with the supplied criteria and the outer 'df[]' displays th
df[df['Date'].astype(str).str.strip()=="184006"]
```

	DataID	DataItem	County	Owner_FirstName	Owner_LastName	Witness	Date	Freed_FirstName	Freed_LastName	Alias
23307	AR7-46	23310	BA	Geo	Gillingham	NaN	184006	Jeremiah W.	Brown	Jerry
23308	AR7-46	23311	BA	NaN	NaN	NaN	184006	Rachael	Brown	NaN

2 rows × 29 columns

In two of the instances, as seen below, the day of issue has not been found to be legible or visible, hence the MSA transcriber may

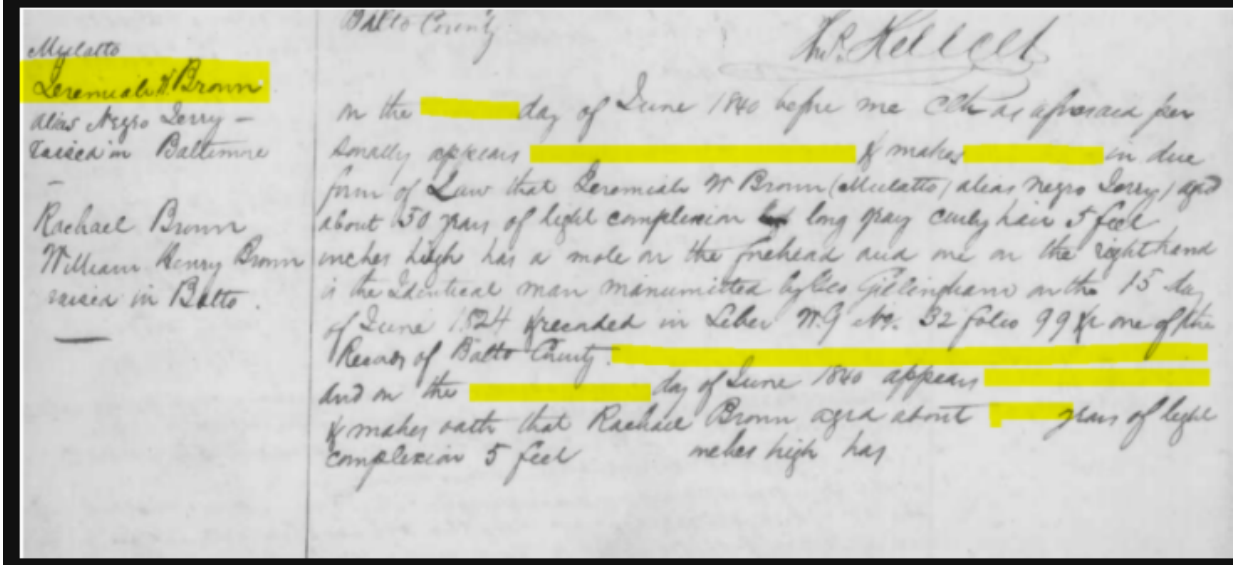


Figure 3.4: Screenshot showing data manipulation of erroneous date records from the CoF dataset

This resulted in the display of two rows from the dataset as shown in Figure. 3.4. Using this information, MSA's database was searched to find the scanned CoF record for the enslaved person named Jeremiah W. Brown, which showed that, on the physical recorded CoF, the day was not legibly visible as highlighted with blanks. This is a classic example of data manipulation and analysis steps that could help students quickly identify the root cause of problems with the source records. Similar interesting patterns were uncovered when working with the Enslaved's Prior Status feature, which is a categorical data. There were different formats of data capture like 'born free', 'free born', 'Born free', etc which could be grouped into a single value, however, values like 'Descendant of a white woman' as shown need to be fixed with the help of historical subject matter

experts since it possesses contextual information. To resolve this problem, I consulted historians who provided valuable insights that the person should be considered 'Free' by being born to a white woman. This shows the importance of collaborating with historians to perform specific key data manipulation and analysis tasks on the dataset collections of the field. The module follows with further code commands. After these data operations were completed, the in-memory data frame was saved to an output *CSV* file which could be used in the next module, Part 2.

3.5.1.3 Contextual Data Analysis and Manipulation - Part 2 Module

The CoF dataset's Height and Age features were worked on for the Part 2 module, whose input was the output dataset saved with new features from Part 1. Similar to the two features from Part 1, the Height feature had some anomalies. These features also required contextual historical data, so I consulted subject matter experts. They were able to provide their research results surrounding these features referring me to a study by Margo and Steckel in 1982 [70], who performed an analysis of the height and age from the Enslaved Manifest data of around 50,000+ enslaved people shipped between 1811 and 1861 to ports like Baltimore, Richmond and other cities from the Port of Savannah in USA. According to this study, the average height of enslaved people was around 67 inches. In the same survey where another set of Enslaved People's appraisal records showed the maximum height was found to be around 72 inches. This essential collaborative insight led me to identify the outliers above a height of 80 inches and below a height of 5 inches. I found four entries with invalid values as shown in Figure. 3.5. On performing a detailed analysis of looking up scanned documents of the records related to these error records, a result appeared as shown in Figure. 3.6 where for the enslaved person Milly Farmer, the date was

```
# code to show bad records
df.loc[(df["Height_Inches"]>80)|(df["Height_Inches"]<5),['DataItem','Height','Height_Inches']]
```

	DataItem	Height	Height_Inches
5032	5034	4' 44.75"	92.75
5625	5627	5'2 1.4"	81.40
6197	6199	5'85."	145.00
15694	15696	9'.75"	108.75

Figure 3.5: Screenshot showing erroneous values of Height feature after data manipulation operation

physically recorded as four feet eleven and three quarter inches however the data was digitally recorded as 4' 44.75" indicating an error. These entries were captured as having the incorrect value 'NaN' and shared with the MSA for source correction. It should be noted that in all the data manipulation tasks where the source data has to be changed, the modified data was stored as a new feature without touching the source data in the feature. This also teaches the students working on such projects to store the source data during such manipulation procedures. This module finishes by saving its results into an output CSV file for the next module, contextual data visualization.

3.5.1.4 Contextual Data Visualization Module

Weintrop [31] asserts that applying data visualization practice to the computational treatments helps students to effectively convey stories and insights gleaned from the previous data manipulation and analysis tasks. This module's key objective was to create meaningful context-based visualizations that highlight the findings as stories. This iDCN module starts by showing how to import specific unique software libraries like *plotly*,[12] *bokeh*, [13] *networkx*, [14] *wordcloud*, [15] which are necessary to

[12]Plotly Documentation - https://plotly.com/python/.
[13]Bokeh Documentation - https://docs.bokeh.org/en/latest/index.html
[14]NetworkX Documentation - https://networkx.org/documentation/stable/index.html.
[15]Wordcloud Python - https://pypi.org/project/wordcloud/.

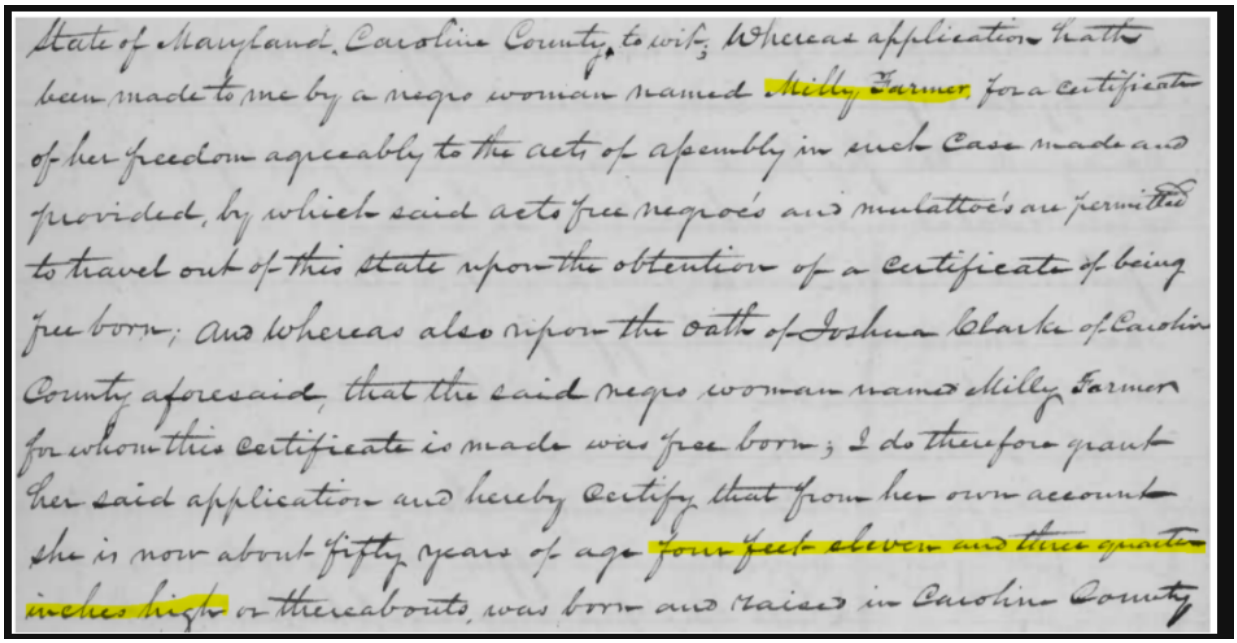

State of Maryland, Caroline County, to wit; Whereas application hath been made to me by a negro woman named Milly Farmer for a certificate of her freedom agreeably to the acts of assembly in such case made and provided, by which said acts free negroes and mulattoes are permitted to travel out of this state upon the obtention of a certificate of being free born; and whereas also upon the oath of Joshua Clarke of Caroline County aforesaid, that the said negro woman named Milly Farmer for whom this certificate is made was free born; I do therefore grant her said application and hereby certify that from her own account she is now about fifty years of age four feet eleven and three quarter inches high or thereabouts, was born and raised in Caroline County

Figure 3.6: Scanned CoF document of Milly Farmer highlighted with the height recorded at source

perform the visualization tasks. This adds to the benefit of creating dynamically changing graphs, charts, geomaps, and network graphs. In Figure. 3.7 and Figure. 3.8, *plotly*'s iplot package was used to create basic histograms of the newly created CoF features Enslaved's Age and Height. It could be seen that most enslaved people were between the ages of 20 and 40, and within the height range of 60 - 75 inches, which matches the study by Margo and Steckel [70]. A plot between gender and the number of CoF's issued by year showed an interesting historical contextual finding as shown in Figure. 3.10. As can be seen, there was a spike in the number of CoFs issued in 1832. By applying contextual historical analysis, historians identified that two key historical events are believed to have occurred between 1831 and 1832 according to Maryland's state history as described in [71] on page 29. Although I cannot make cause-effect relationships between these two factual findings, this kind of insight would not have been possible without the application of data science techniques followed in this project. These findings show how non-STEM students and educators can use the iDCN modules and help them identify patterns and make necessary

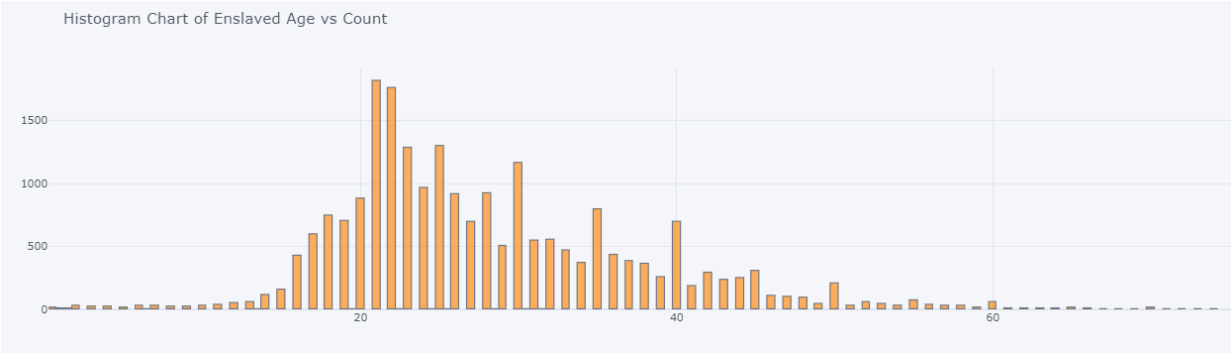


Figure 3.7: Screenshot showing a histogram chart between Age feature and the number of CoFs issued

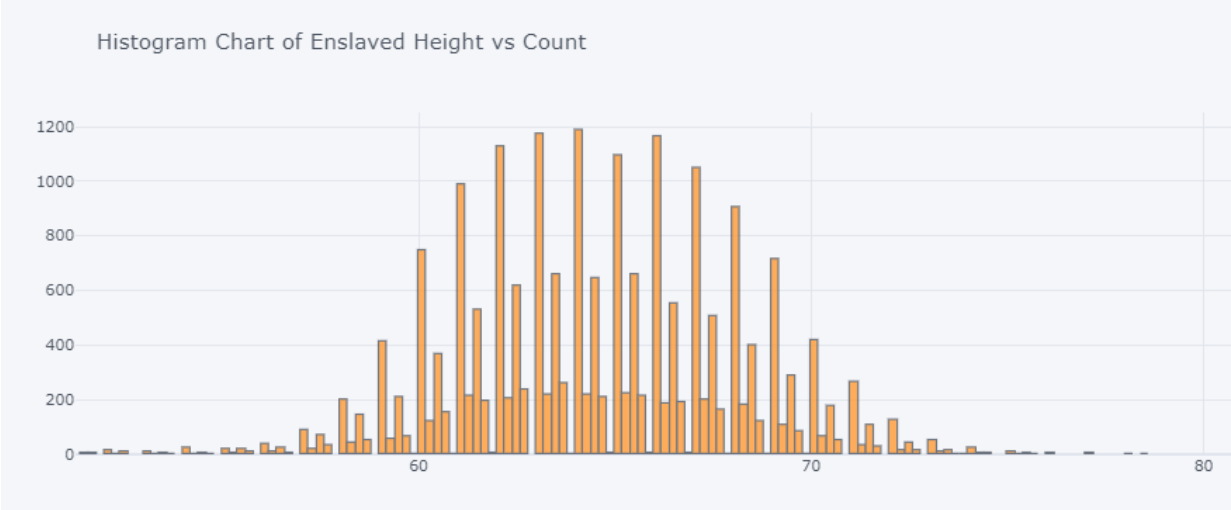


Figure 3.8: Screenshot showing a histogram chart between Height feature and the number of CoFs issued

connections that may not have been possible without a data-driven approach. Another unique visualization of geo maps was plotted using the *plotly* express's choropleth map box technique, which uses the FIPS code, a unique code given to each county in the USA, to plot the desired features on the map of the USA. In Figure. 3.11, an interactive map has been plotted that shows the color of the counties based on the weight of the counts of CoF's issued for them. This module also touches on the Natural Language Processing abilities of the wordcloud library supported by *Python*. From the CoF's Notes features, which included the handwritten text by the digital data transcribers at the MSA, I showed how to code a command to create a word cloud that results in a

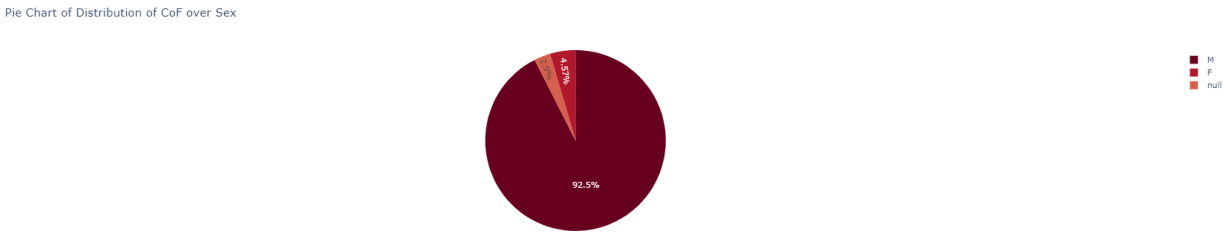


Figure 3.9: Screenshot showing a pie chart of percentage of CoFs issued by gender

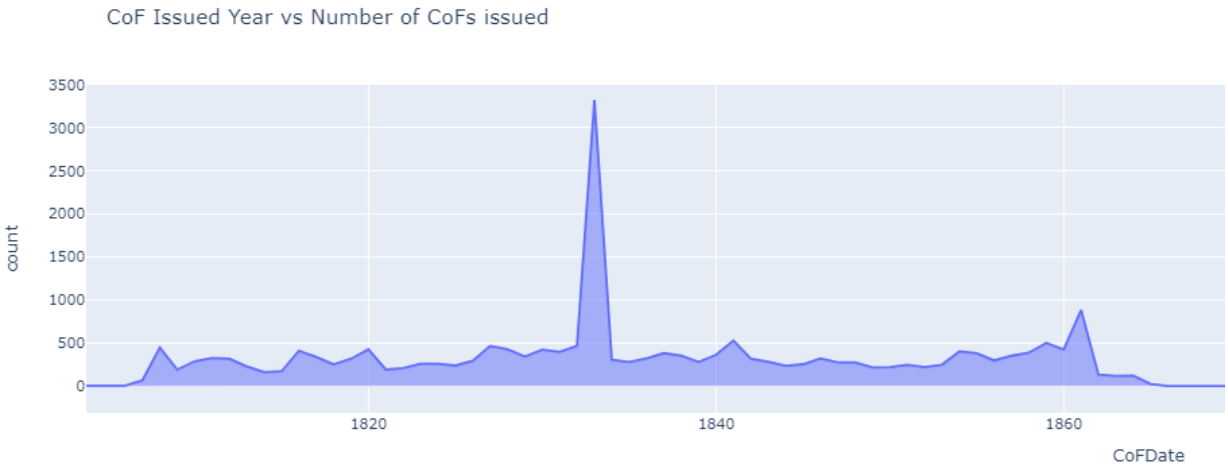


Figure 3.10: Visualization showing number of CoFs issued by Year from the CoF dataset

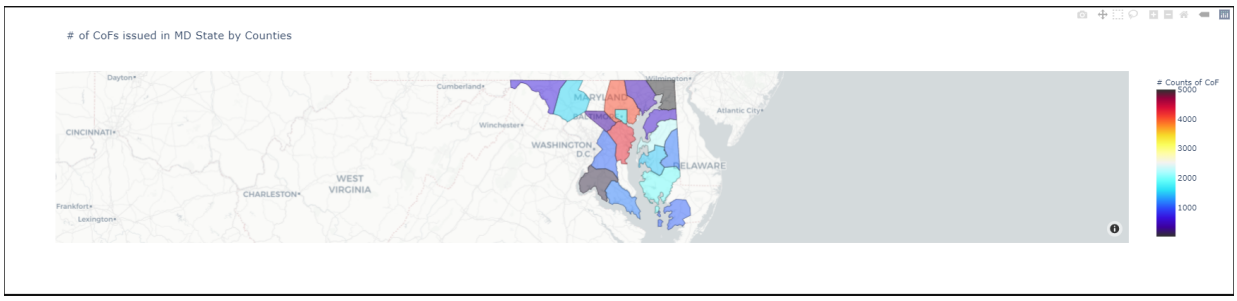


Figure 3.11: Screenshot showing a geomap diagram using FIPS County codes in MD state of the USA

graphical representation of the most commonly used words. This type of visualization could be used as a starting point to research the dataset's text-oriented features further. These four iDCN modules together form a coherently arranged learning tool that non-STEM educators and students could harness to understand digitized data appropriate to their fields better.

3.5.2 User Survey Results and Analysis

To conduct the user survey with the graduate students and educators with non-STEM backgrounds, the University of Maryland's Institutional Review Board (IRB) was contacted and presented with a package of the elements of the survey, including providing a CONSENT form, which was to be signed by survey participants at the beginning of the survey study. The IRB members reviewed the package thoroughly for human-subject research determination and approved the survey for distribution. The survey was designed to be anonymous and no personal information was captured. The survey had three sections: a CONSENT form at the front, a Yes or No question requesting users to answer if they could review the iDCNs web page hosted publicly, and a set of seven questions grouped on a Likert scale as given below. The Likert scale questions were not mandatory, and the students could skip any questions if they were uncomfortable answering them. These were on a 7-level Likert scale: Strongly disagree, Disagree, Somewhat disagree, Neutral, Somewhat agree, Agree, and Strongly Agree.

- Q1: Were the iDCN modules arranged logically and coherently?
- Q2: Does the idea of making text, pictures, visualizations along with computer programming code co-exist together, help make the iDCNs easy to understand?
- Q3: Do the iDCNs modules look appealing or inspire learning about the chosen dataset

collection?

- Q4: Does learning from the iDCNs provide the same aesthetic value or more or less as studying from a paperback book / traditional book?
- Q5: Do the entire iDCNs modules follow the CT framework introduced and identified in the index module?
- Q6: Were these iDCNs educational?
- Q7: Can these iDCNs be used for further course development with other dataset collections?

The survey was distributed to two groups of students and educators, most of whom I believe were being introduced to this learning tool for the first time. It was sent to an overall count of n=37 participants of which 4 were educators and 33 were students. Of the 37 survey recruits, 31 were able to complete the survey at a response rate of 83.7%. 2 of them were unable to review the iDCNs and four were unable to complete the entire survey. Of those 31 who completed the survey, only one could not answer all seven close-ended Likert scale questions. From Figure. 3.12, which shows a chart of responses to these seven questions, the results were analyzed holistically for overall feedback and key inferences were made. Although the sample size is not large enough to make generalizations from the survey evaluation, the survey results capture first impressions and provide significant insights into the possible adoption of this novel educational tool for non-STEM educational areas. Survey responses were overwhelmingly positive on all seven questions, with some stronger reservations on questions Q3 and Q4, which probe on the look and feel side: contrasting iDCNs to traditional books, and asking about their appeal. This is valuable feedback, as one can be taken aback initially by experiencing what may look like a “wall of text and code

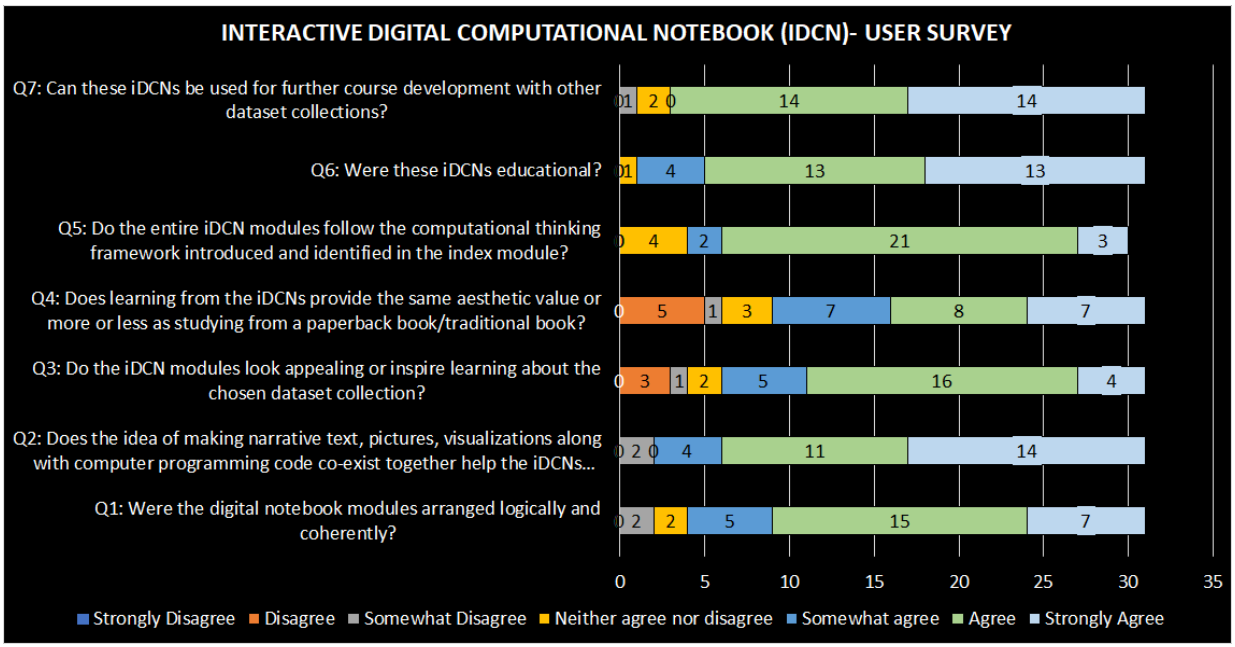


Figure 3.12: User Survey Results of Likert scale questions for the iDCNs, (n=31)

boxes." This suggests that particular care be given to telling an interactive computational story and its visual composition and representation. It was particularly encouraging to find highly positive feedback for questions Q6 and Q7, which indicate that the iDCNs were educational and that such courses could be developed for other dataset collections. This shows that I achieved the primary objective of the first prong in developing the iDCNs as a learning tool for promoting CT practices and encouraging the non-STEM students to adopt computational methods. One of the students with a historical background who was on the list of survey recruits sent a direct email to me with comments, as quoted below. I was unsure if the student completed the survey due to the anonymity of the data capture, however, these comments indicated positive feedback.

> "I think that some of the graphs that have already been created here do a good job of demonstrating the power and flexibility of these tools. The ability to deny or substantiate all sorts of historical claims using hard, quantitative data is of MASSIVE value to historians in the field. We've never been able to do this before on this scale,

much less make it transferable to other scholars."

3.6 Discussion

Though the objective for this study was achieved and the RQs were answered, it is understood that introducing a novel learning platform like this for traditional learners could be a challenge initially. To bring back the point made by Levander and Decherney in [72], on how internet accessibility is a central issue with regards to underserved populations, the need to get online to access these iDCNs might pose a significant challenge to reach all groups of students and educators. This challenge is particularly true for any online learning platform and needs to be addressed at an overarching higher educational level. Although the modular nature of the iDCNs that were developed provides a coherent way to navigate within the course content, the impression, as seen from some of the survey responses for Q1, indicates that this area could also be improved. This shows the limitations of working with an open-source tool designed and developed originally for purposes other than as a learning tool.

3.7 Future Work and Conclusion

In this study, I demonstrate designing and developing a set of iDCN modules and walk through a step-by-step process of performing contextual data science-based analytics and visualization on the CoF dataset. These iDCNs were designed based on an established CT framework. A user survey evaluated the resulting modules and showed positive feedback with potential to expand to other such datasets. This shows that the research objective I was successfully handling for the first prong strategy of this dissertation. As future work, this infrastructure could be developed to harness

the power of integrating with more software libraries to create iDCNs that could explore datasets on topics like Machine Learning (ML) concepts and advanced Natural Language Processing (NLP) processes. By creating, sharing, and making these open-source learning platforms available online for free, these resources help students, educators, and researchers worldwide leverage these and benefit from them. The iDCN modules have been published at the *https://cases.umd.edu.* [16] The source code is also published to *GitHub.* [17]

[16]Legacy of Slavery - https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb.
[17]Legacy Of Slavery GitHub project - https://github.com/cases-umd/Legacy-of-Slavery.

Chapter 4: Objective II – Designing and Evaluating Gen-AI Solutions for Enhanced Access

While Objective I demonstrated that iDCNs can empower archival students and researchers with computational skills, the reality is that not everyone will become a programmer or data scientist, nor will all trained individuals be able to manually process the vast and complex datasets found in cultural archives. Education and skill-building are necessary, but they may not be sufficient to bridge the accessibility gap fully. This recognition sets the stage for Prong 2—integrating advanced Gen-AI tools to lower barriers further and enable users to interact with archives more intuitively, natural language-driven. The Research Objective II forms the first part of this second-prong approach towards leveraging, evaluating and optimizing Gen-AI (via unique retrieval systems, Gen-AI agents) to create innovative Gen-AI applications (such as interactive conversational platforms aka chatbots [6]) for enhancing access with digital archives. This objective is made up of two studies.

4.1 Designing and Developing ChatLoS, an Gen-AI Powered Conversational chatbot

4.1.1 Introduction

Gen-AI offer an innovative interface to archival data that can democratize access, allowing students, researchers, and the public to discover and learn from archives in a conversational, context-aware way, while maintaining the ethical safeguards necessary for trust in AI-generated archival narratives. This study, peer-reviewed in 2023 and to-be published in 2025 at [73], adopts an exploratory case study and design science approach to build one such Gen-AI chatbot, named *"ChatLoS,"* leveraging OpenAI's GPTs and using a specific mechanism called Retrieval Augmented Generation (RAG). ChatLoS enables natural language querying of LoS datasets, specifically DTA, through iterative prototyping and real-world user testing, removing the need for database schema knowledge. This study, however, was approached with a need for a high degree of cultural sensitivity and awareness of the ethical implications involved. Considering the sensitive nature of the DTA dataset, it is paramount to treat the data and the resulting analysis with utmost respect, ensuring the enslaved individuals' experiences and identities are neither trivialized nor exploited. The study examines performance metrics such as user satisfaction, accuracy, and ethical risks like hallucinations.

4.1.2 Background

For this study, I used the DTA dataset from the LoS collection. After performing preliminary data exploration and a cleaning process, I 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, I realized that the project's real benefit could be vastly enhanced if the text from the advertisements was available to be fed as input to the GPT model. To achieve this, I used an OCR tool *ABBYYFineReader* [74] to extract the text for each of these 764 ads, and the entire dataset was augmented with this new feature. The objective in using this DTA dataset was to explore if Gen-AI solutions using the LLMs can provide an interactive chat interface to reveal hidden insights in the ad patterns or the behavior of the slave owners or enslaved people during this period. For example, *did more sales occur during December than certain other months because enslaved people are sometimes bought and sold as gifts?* Although this may be a horrid concept to the modern sensibilities, remember that enslaved people were considered commodities, as valuable in many instances then as a prized horse or a new car today. *Did the south westward migration which followed the 1820 Missouri Compromise provoke more purchases of enslaved Blacks to plow new lands?* Answers to such questions and illustrations of which skills, physical traits (age, gender, build, complexion), and departure sites were most often advertised may reveal preferences to the slave trade that vary from a general belief that any young African male was always in highest demand. The MSA's director asked some of the above questions due to the domain knowledge expertise and close association with the cultural archives. To find answers to these questions using non-AI tools, the director needs to employ staff with the computational skills required to perform these analyses, or the director should be equipped with the skills necessary. With Gen-AI tools introduced through this study, the MSA director could chat with the conversational chatbot to ask questions in natural language without having to perform complex and advanced data analysis on the semi-structured dataset such as the DTA. This advancement quickly enhances access to the

underlying culturally rich archives datasets such as the DTA, and to realize stories that have not been unearthed before.

4.1.3 Research Question

To achieve the objective set for this study, I came up with a main research question below:

- RQ2: What are the potential benefits and challenges associated with utilizing LLMs, such as OpenAI‘s GPT, to develop a tailored chatbot capable of exploring data from high cultural context datasets and unveiling previously undiscovered relationships and patterns within this data?

4.1.4 Methodology and Approach

The approach for this study follows an exploratory case study and design science methodology. By creating a unique *"ChatLoS"*, which acts as the virtual archaeologist aka the DTA Help Desk for any archival enthusiasts, this study involves the exploration of the DTA dataset as a case study, as well as a design science product due to the conversational interface that‘s interactive with the end users. The idea behind this Gen-AI chatbot was to use the trimmed exploratory DTA dataset with 764 records as a CSV file and allow users to chat against this dataset without knowing the underlying data layout or column names, but with the contextual understanding of what‘s in the dataset. As noted earlier, the DTA dataset was augmented with additional features to supplement the structured metadata of this CSV file. As part of this process, other features on this dataset were also prepped to be AI-ready. Upon completion of the data cleaning process, the CSV dataset was imported into a custom software project[1] that used the OpenAI GPTs via OpenAI API, the

[1] https://github.com/rgnanase/los-gpt/blob/main/api.py

Python language, the LangChain[2] libraries, and open source web application projects to process the CSV dataset and create an interactive user interface. A key feature of this project is employing a specific mechanism called RAG, explained below.

4.1.4.1 ChatLoS with RAG mechanism

For this study, I implemented a widely used technique for creating Gen-AI conversational interfaces called RAG [75] in the custom software project to develop *ChatLoS*. In simple terms, the project is like a recipe for a computer program that helps us understand the dataset collection of DTA about the slave trade in Maryland. These ads could be considered a gigantic book too big to read simultaneously. Firstly, this program reads the whole book and then divides it into smaller, more readable parts named ”chunks”, like chapters in a novel. Secondly, after it breaks down the big book into chapters, it gives each word in the chapter a unique label called ”Tokenization” for easy reference. Thirdly, a special program translates these tokenized words into a unique computer-friendly language, called vector embeddings. The program then saves these translated tokens in a vector database. Then, the retrieval system comes into play. This technique allows the chatbot to work with the pre-trained LLM on limiting what information it uses to respond 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. I have provided the steps shown in Figure. 4.1 to understand how this works. Here‘s a simple breakdown of the image using the DTA dataset as an example:

[2]https://python.langchain.com/docs/get_started/introduction.html

1. *Query*: Imagine an end user using ChatLoS to ask a question about the history or specific details 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 vector embeddings database—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.

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 explicitly 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

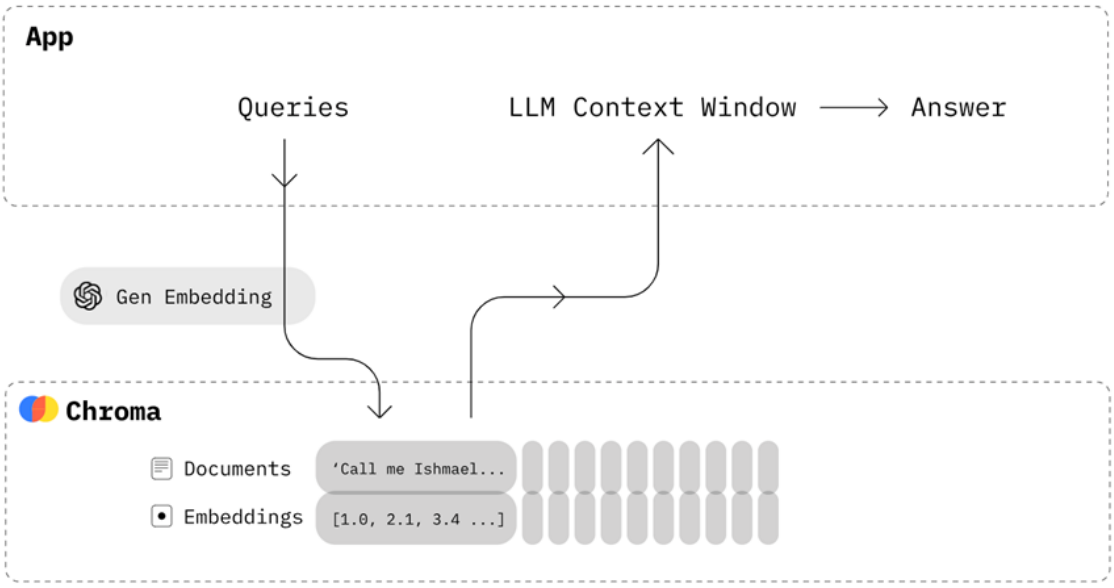


Figure 4.1: Retrieval Augmented Generation [76]

from these ChatLoS-type conversations are discussed in the following section. The sections below provide a detailed step-by-step account of the script features.

4.1.4.2 Reading and Tokenizing the Dataset

The ‘start_openai_db‘ function begins by opening and reading the dataset file. The text content of the file is then tokenized using the ‘tiktoken‘[3] library from OpenAI, which provides an approximate count of the tokens present in the text. Tokens, special computer-friendly atomic units of processing in language models like GPT, are used here to identify, count, and manage the pieces of text being worked with.

[3]https://pypistats.org/packages/tiktoken

4.1.4.3 Text Chunking

The text data is then divided into manageable ”chunks” using the ‘RecursiveCharacterTextSplitter‘ from the ”langchain” library. The process of chunking is a pragmatic necessity when working with large datasets, as it allows data to be broken into smaller, more manageable pieces that are easier for the Gen-AI model to process and analyze. The chunks are created with a specified size (350 tokens in this instance) and overlap to ensure continuity and coherence in the processing. Each chunk is also tagged with metadata — a unique identifier (its index in this case) for reference and tracking purposes.

4.1.4.4 Creating Vector Embeddings

The langchain library‘s functionalities are further leveraged to create vector embeddings from the chunks. Embeddings are representations of the chunks in a high-dimensional vector space, enabling the capturing and encoding of semantic information from the text in a manner useful for machine learning models. These embeddings are stored in a ‘Chroma‘[4] vector database, a powerful tool for managing and querying vector data.

4.1.4.5 Constructing the Conversational Retrieval Chain

The ‘ConversationalRetrievalChain‘ is constructed from the language model, vector database, and a memory buffer. This chain, powered by the GPT model, forms the core of the conversational Gen-AI system. It uses the vector database to retrieve relevant responses and the memory buffer to store conversation history. The chain enables the Gen-AI to maintain context throughout the

[4]https://python.langchain.com/docs/modules/data_connection/vectorstores/integrations/chroma

interaction, resulting in more coherent and context-aware responses.

4.1.4.6 Interaction with the ChatLoS User Interface

Finally, a POST endpoint named ‘send_openai‘ is set up, which accepts text input, processes it through the Gen-AI model, and returns the AI‘s response. When accessed with the appropriate text input, this endpoint triggers the Gen-AI model to generate a response, marking the culmination of the data preparation and model setup process.

In conclusion, the preparation of the DTA dataset, its transformation into tokens, the subsequent chunking process, the construction of the retrieval chain, and the eventual conversation with the Gen-AI model (the chatbot) constitute a meticulously orchestrated setup of data and AI, unlocking unprecedented insights into the DTA dataset. A sample conversation using the chatbot from this step is shown in Figure. 4.2 and 4.3 below. The responses from this version of the chatbot to the question on ”how many advertisements” it has stored in its context imply that it could only store the textual data for a few of the 700+ advertisements that were chunked and retrieved based on the semantic search against the vector database. Any responses to the questions would be limited to the text data found in these few advertisements. However, a unique advantage with the ”chunking” process is that depending on the question, the model would select other advertisements to be placed in its memory to curate an appropriate response.

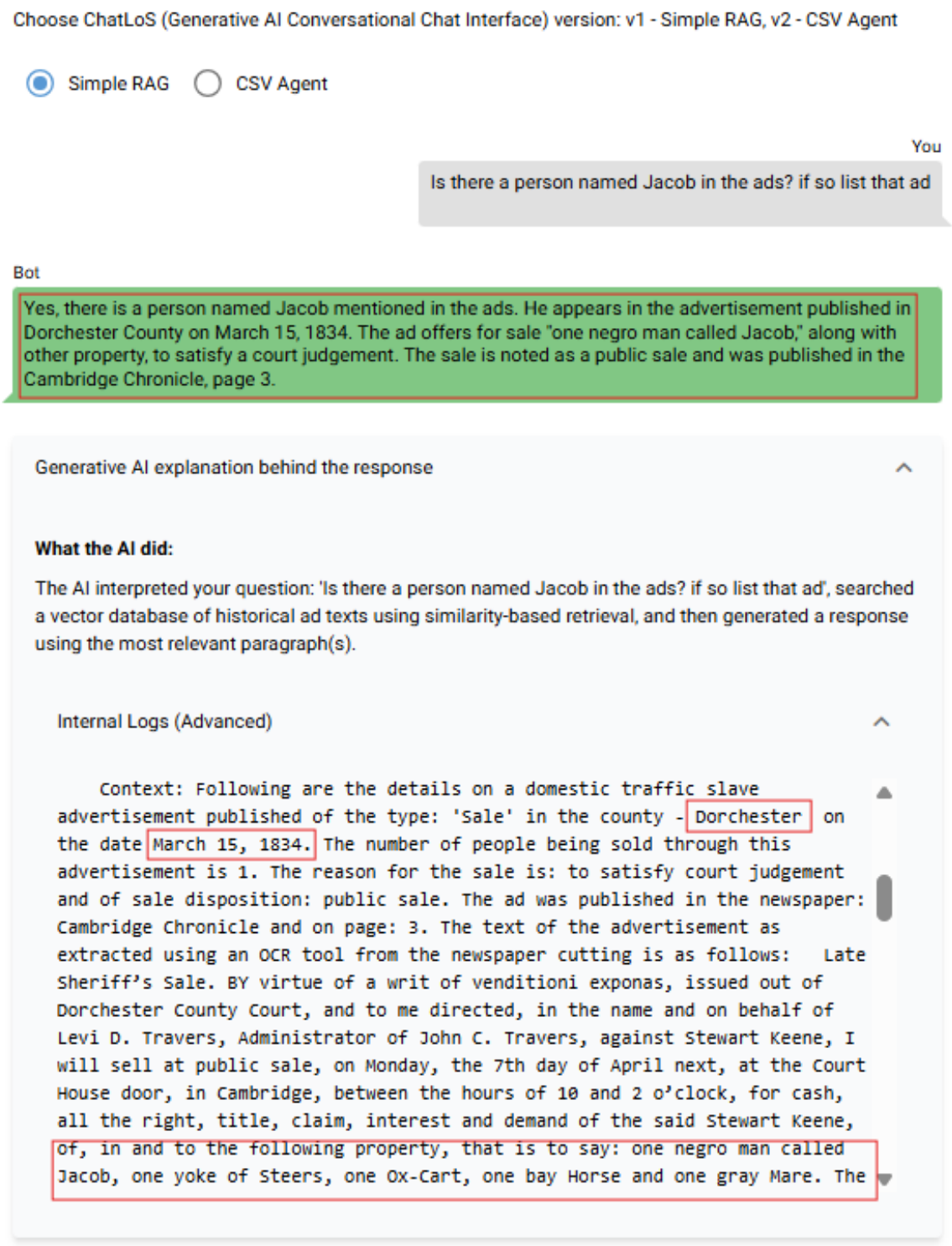


Figure 4.2: Simple RAG based ChatLoS v1 response for a sample name search question

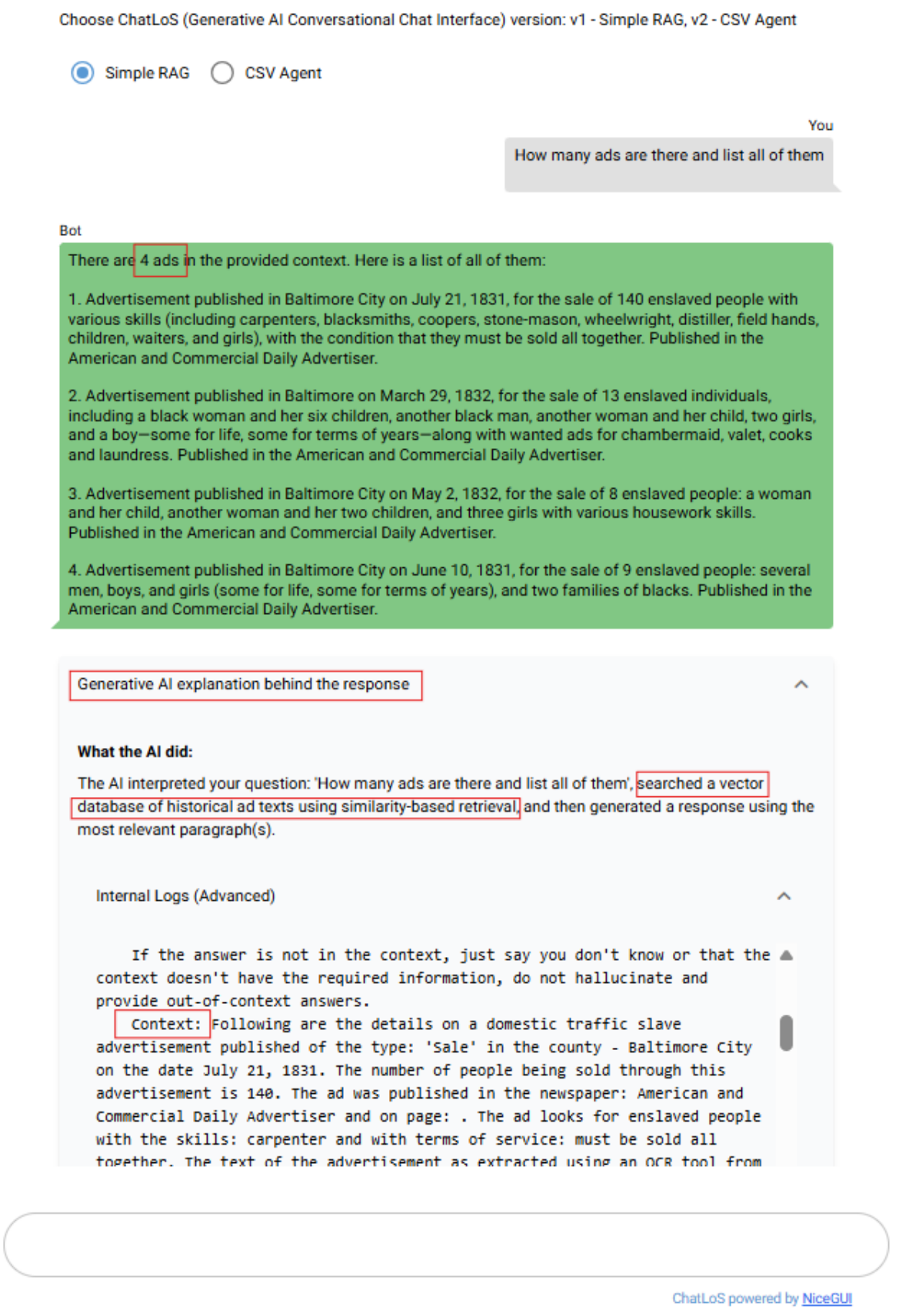


Figure 4.3: Simple RAG based ChatLoS v1 responding to a sample aggregate question

4.1.5 Results and Discussion

4.1.5.1 Prompting in Gen-AI

Prompting involves feeding a user-generated input to a language model, which then infers a completion [77]. For instance, if the input is ”4 + 3 = ”, the LLM would likely respond with ”7”. Using prompts to extract valuable data from a model is called ”Prompting”. This technique does not require a sizeable offline training set or offline access to a model and feels intuitive for engineers and non-engineers. The process starts by identifying a problem that could be solved using prompting. These are problems to which well created prompts might elicit the desired behavior from the language model. Creating these ”prompts” requires some basic knowledge and careful consideration of factors like being assertive rather than defensive, and being concise rather than repetitive. Working with a historical and cultural context such as the DTA, creating prompts becomes a vital step, even though it could be considered a trivial task in the grand scheme of things.

4.1.5.2 ChatLoS as a Contextually-aware Gen-AI Chatbot

Another significant expectation from a tool like ChatLoS that uses a culturally sensitive dataset such as the DTA is to be contextually aware of the data and respond to questions accordingly. The whole point of enhancing ChatLoS with RAG is to use its ability to understand the common aspects of a language like ”English” to build on a specific knowledge domain. In this case, the reference knowledge was performed on the OCR text column of the DTA dataset records, and the ChatLoS was expected to limit the responses to the context learned from this text. Below are the

results on how ChatLoS performed. In Figure. 4.4 and 4.5, the simple RAG based ChatLoS is seen to respond that it could not find references to the people asked about who are commonly known and prominent figures in American History. This indicates that the ChatLoS is tuned to the domain knowledge it learned on the 764 rows of DTA.

4.1.5.3 ”Explainability” in ChatLoS v1 (Simple RAG)

A key design feature implemented in ChatLoS v1 (the Simple RAG-based version) is the inclusion of an *explanation layer* alongside its generated responses. This feature was introduced to promote transparency, trust, and user interpretability—particularly crucial when deploying Gen-AI on culturally sensitive collections like the LoS datasets. In this version, the system provides not only an answer to the user's question but also an explicit **explanation of the retrieval and generation process**. The user interface displays a collapsible panel labeled *" Gen-AI explanation behind the response"*, which details:

- How the model interpreted the question.
- That it searched a vector database using similarity-based retrieval.
- A breakdown of the prompt, context, and instructions used to generate the answer.

For instance, when the question *"Who is George Washington?"* is asked, the chatbot explicitly states: *"The context does not contain any information about George Washington"*, and the explanation panel logs that no matching entries were found in the context window. It further reveals a system-level instruction: *"If the answer is not in the context, just say you don't know or that the context doesn't have the required information; do not hallucinate or provide out-of-context*

answers." This directive reinforces ethical safeguards and prevents speculative or misleading completions.

This explainability feature addresses several critical concerns:

1. Mitigation of Hallucinations: By disclosing that the model's outputs are based only on retrieved context, the system reduces the risk of hallucinated responses—especially important when handling marginalized histories and records of enslaved individuals. For example, in response to an irrelevant question about George Washington, the model correctly admits that no context is available, rather than fabricating an association.

2. Ethical Transparency: Gen-AI systems, particularly those interacting with historical archives, face significant scrutiny regarding trust and verifiability. Showing internal logs and search contexts allows researchers and users to evaluate the provenance of the model's conclusions.

3. Cultural Sensitivity: The LoS collections contain sensitive narratives. In this context, the ability to trace responses back to semantically retrieved segments—such as a sale advertisement mentioning an individual named *Jacob* (as illustrated in Figure. 4.2) supports respectful and grounded engagement with the material. These responses include metadata such as ad date, location, and document source (e.g., Cambridge Chronicle), aligning with archival norms for source citation and authenticity.

4. Empowering Accountability: The design of ChatLoS v1 encourages archivists, researchers, and even general users to understand how the system arrives at an answer. In the example shown in Figure. 4.3, the model not only lists the number of ads retrieved but also details the publication

dates, content summaries, and source newspapers, ensuring that outputs can be independently validated.

In summary, the explainability layer in ChatLoS v1 fosters ethical deployment by reinforcing non-hallucinatory behavior, improving trust in results, and offering a self-documenting trail of interpretability. While this version does not yet support aggregation or cross-collection reasoning (limitations later addressed in the ChatLoS v2 and v3), it sets a strong ethical foundation for generative archival interfaces by embedding transparency directly into the interaction loop.

4.1.6 Limitations & Future work

Despite the significant promises of Gen-AI and LLMs, it is critical to address ethical considerations, especially the risk of generating synthetic data that may mislead or misinform users [78]. Therefore, while Gen-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 [79]. Therefore, the potential of LLMs' ability to manage and make sense of digital archives is immense, and their deployment needs to be monitored for ethical application and potential bias in their outputs. With the contextually aware ChatLoS version, as noted earlier, and with the RAG enhanced mechanism, there are significant issues when performing advanced data analysis on the DTA dataset. This is because ChatLoS only retrieves specific blocks of information stored as chunks, and this method isn't conducive to performing aggregate data analysis. This approach is valid if one wants to search for a specific record from the DTA dataset

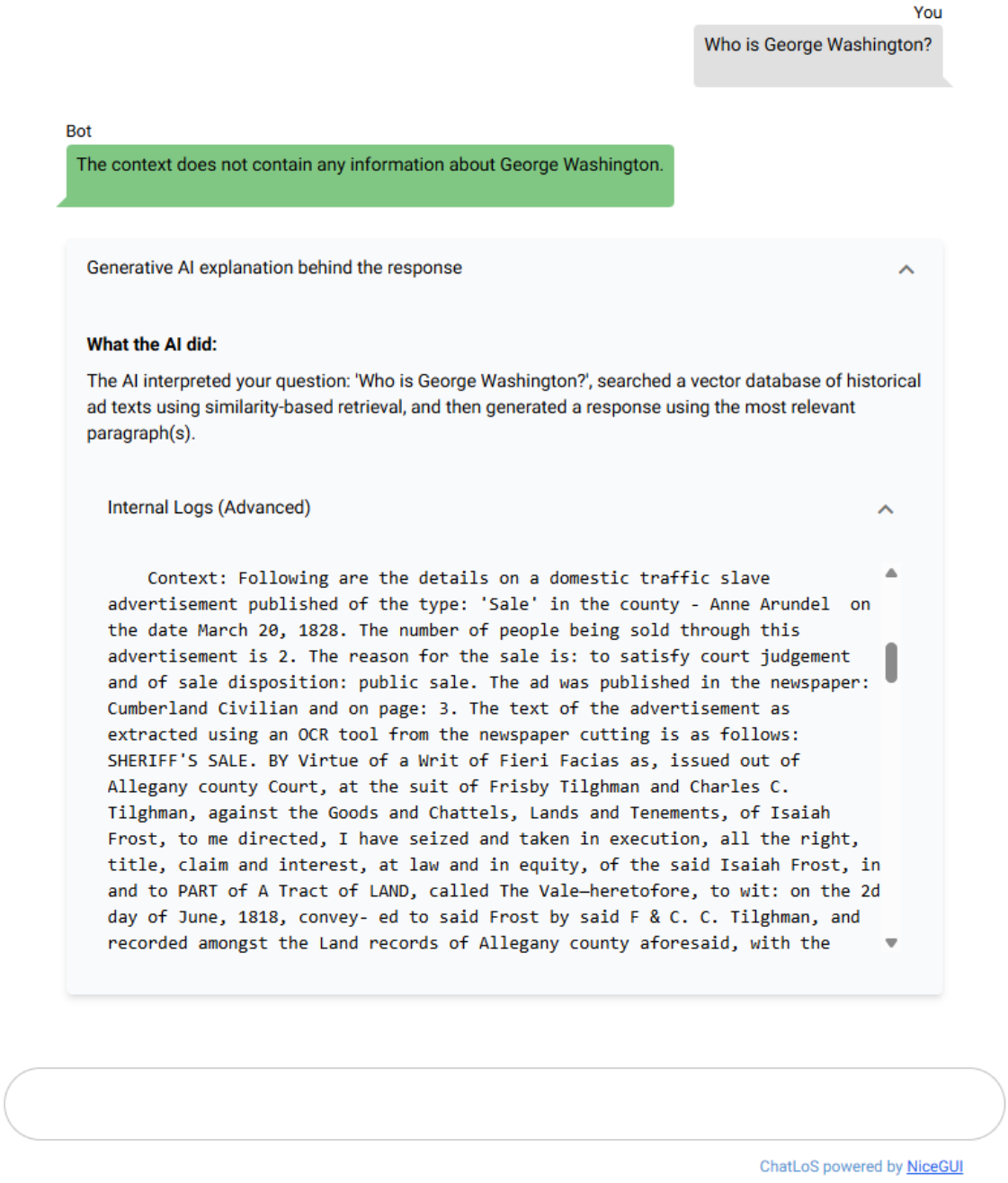


Figure 4.4: Simple RAG based ChatLoS v1 showing prompt and retrieved context

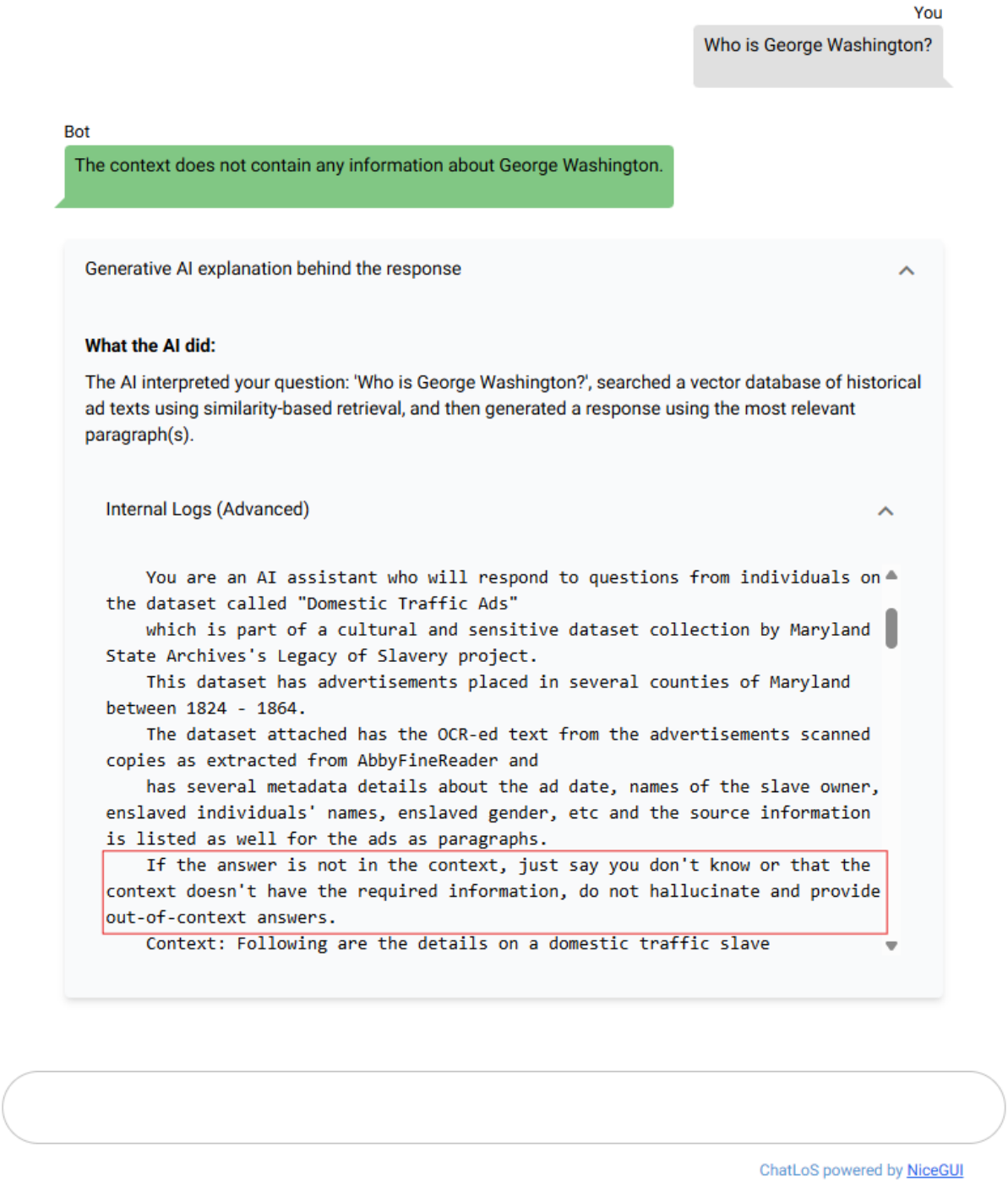


Figure 4.5: Simple RAG based ChatLoS v1 responding to out-of-context questions

based on specific keywords and to limit the knowledge of the responses to a particular domain. To address this challenge, I conducted the Study 4 which employs an unique Gen-AI agentic approach to perform advanced data analysis on the DTA dataset.

4.1.7 Conclusion

In conclusion, to answer RQ2, OpenAI‘s GPT LLM was used to create ChatLoS. This contextually aware conversational chatbot contained knowledge based on the RAG enhancement using the MSA‘s DTA dataset. The RAG enhanced chatbot showed how it employs a chunking process that breaks down large amounts of text data into manageable pieces for more efficient querying and processing by the language model to learn appropriate context. This enables the chatbot with a short-term memory buffer to store user inputs and system outputs for a particular chat conversation. This method optimizes performance and resource allocation by not simultaneously overwhelming the language model with excessive data, thereby failing with errors. The chunking process is critical for maintaining the coherence and consistency of a conversation, mainly when the conversation involves large volumes of data. It allows the language model to work in a manageable context, improving the accuracy and precision of its outputs for questions with specific record searches in mind. Despite its limitations, this is a powerful tool for the users of ChatLoS as it abstracts the users from the underlying data schema and allows them to chat in natural language freely. This is a paradigm shift in accessing the cultural archives data. With further enhancements in the upcoming studies, I will show how the Gen-AI space is conducive to such use cases.

4.2 Enhancing ChatLoS for Advanced Aggregate Data Analysis using Gen-AI Agents

4.2.1 Introduction

Study 3, peer reviewed and published in March 2024 [14] forms the second part of the Research Objective II. This study focuses solely on enhancing ChatLoS, developed in Study 2, to achieve capabilities to respond to DTA dataset-specific data aggregation analytical questions in natural language statements without knowing the details of the metadata or columns. Upon successful enhancement, I also perform an empirical comparative study to assess how the ChatLoS data analysis responses compare to results from traditional data analysis using a commonly used tool such as *Tableau* [80]. I also provide a comprehensive metrics analysis comparing the results of both studies. 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* [81], 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 my efforts in addressing the task above as part of this project's objective. The rest of the paper is divided into the following sections: Research Methodology & Questions, Approach, Results & Findings, Future Work, and Conclusion.

4.2.2 Research Methodology & Questions

I have chosen the empirical comparative study methodology for this project using the DTA as the case study dataset. To address this study's specific objective mentioned above, I formulated the following research question:

- RQ3: 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?

4.2.3 Approach

To address the RQ3, I enhanced the ChatLoS designed and developed in Study 2.

4.2.3.1 ChatLoS using Langchain's CSV Agents for Advanced Data Analysis

To address RQ3, I 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' [82] 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 Gen-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 respond with accurate results. A simple example of a chat conversation is shown in Figure. 4.6; let's understand how ChatLoS aggregates the number

of advertisements placed in a specific county named “Harford” in Maryland. Figure. **??** 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 column 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 RQ3, a new Tableau workbook was created using the DTA dataset, and each sheet was made to perform data aggregation analysis and visualization accordingly. The results of the ChatLoS responses and the comparison between them and Tableau's non-AI traditional data analysis results are performed and discussed in the next section.

4.2.4 Results & Findings

This section details the results obtained from the CSV Gen-AI Agent based ChatLoS v2 to address RQ3 and the comparative study results with traditional data analysis and visualization tools.

4.2.4.1 Explainability in ChatLoS v2 (CSV-Agentic AI)

ChatLoS v2 builds upon the limitations of the original simple RAG interface by integrating LangChain's `create_csv_agent` function, enabling the chatbot to interact directly with structured CSV datasets through dynamic code execution. A key strength of this architecture lies in its embedded explainability mechanism—rendered as an automatically generated log of intermediate computation steps. This capability is not only technically transparent but also ethically meaningful, especially when deployed over historically sensitive datasets like LoS's DTA dataset. Each answer generated by ChatLoS v2 is accompanied by an *Internal Logs (Advanced)* panel. The explainability in ChatLoS v2 demonstrates transparency through live introspection of executable code fragments, Python interpreter traces, and dataset-specific reasoning paths. For example, when asked:

"Were any ads placed on Christmas Day, if so, how many and what are the years?"

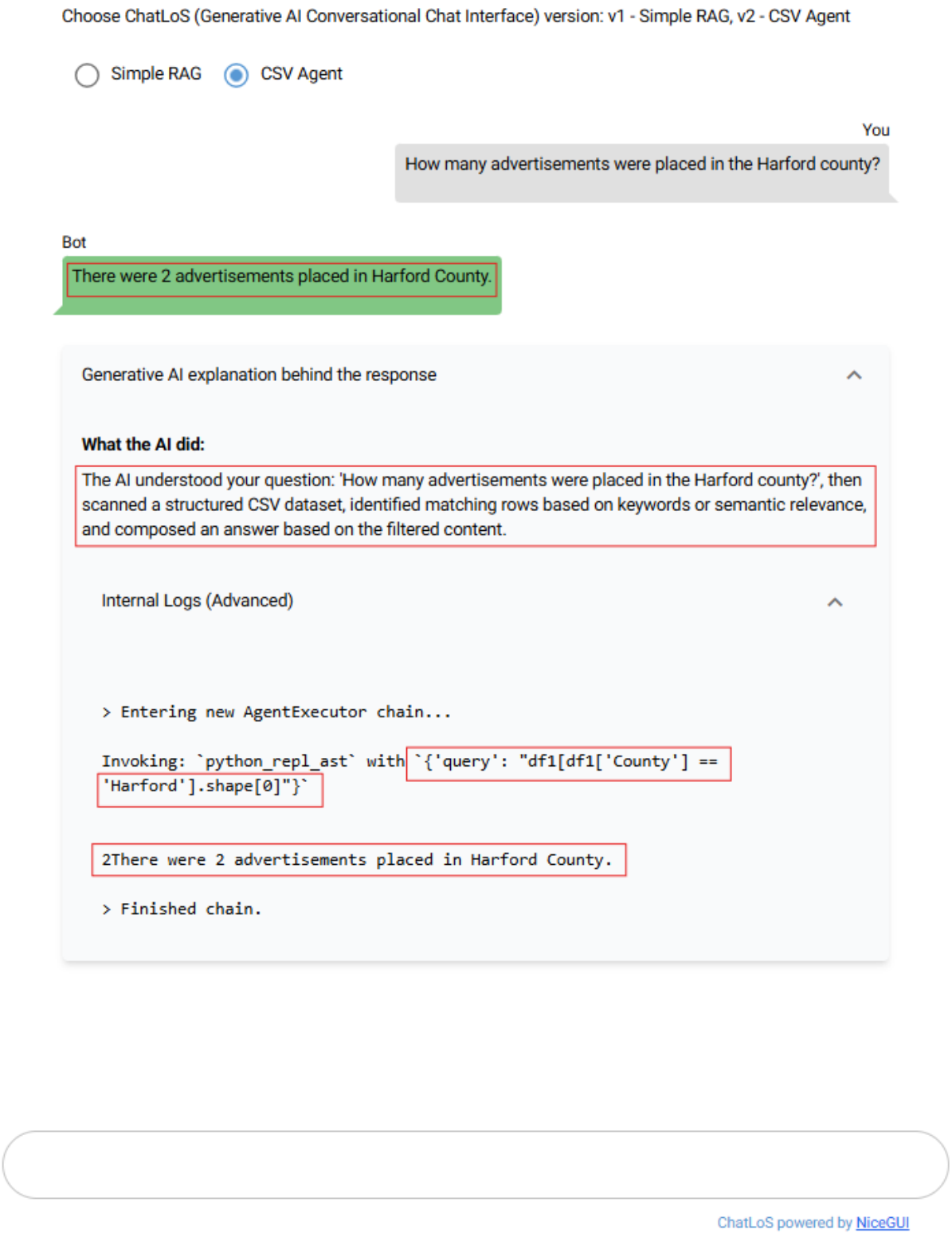


Figure 4.6: CSV Agent based ChatLoS v2 responding to a data aggregation question

the system responds not only with the correct number of ads but shows a real-time invocation of the underlying logic: `df1['_Ad_Date'].str.contains('December 25').sum()` along with the filtered rows (see Figure. 4.13). This makes the entire reasoning traceable and reproducible by users or external auditors.

1. Executable Logic = Verifiability: By revealing the exact code used for analysis (e.g., Pandas operations like `value_counts()`, `shape[0]`, or filtered selections), ChatLoS v2 elevates transparency from a conceptual level to a technical artifact. Whether answering questions about the number of advertisements in *Harford County* (Figure 4.6) or visualizing bar charts (Figure. 4.19), users are empowered to verify the underlying data operations without needing to trust a black-box model. This contributes to both scientific reproducibility and digital provenance.

2. Constraints and Honest Boundaries: In complex scenarios such as tracing a person across datasets—e.g., from Domestic Traffic Ads to Manumissions to Certificates of Freedom, ChatLoS v2 honestly surfaces its inability to perform multi-dataset reasoning (Figure.4.23). Instead of hallucinating links, it provides a structured explanation: limitations of matching by first names alone, need for contextual triangulation, and potential for name ambiguity. This reinforces ethical use by maintaining epistemic humility when evidence is insufficient.

3. Building End-User Trust: The consistent design of showing *"What the Gen-AI did"* aligns well with ethical guidelines in human-centered AI. Users are treated as collaborators rather than passive recipients. Particularly in educational or archival research contexts, this design choice builds confidence among users—historians, students, genealogists—who may otherwise distrust opaque Gen-AI systems. Additionally, these explanations can serve as pedagogical tools, teaching

users how structured data queries are formed. In summary, explainability in ChatLoS v2 is achieved through direct reflection of execution traces, making each answer auditable and grounded in structured computation. This creates a high degree of transparency, prevents hallucination, and fosters trust—all crucial factors for deploying Gen-AI tools in cultural heritage and archival domains.

4.2.4.2 ChatLoS v2 Test Results & Metrics Comparison with Tableau results

To evaluate ChatLoS with advanced data analysis capabilities using Gen-AI Agents, the bot was asked six 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, I 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. 4.22. Individual test case results are captured below with a Pass or Fail against the tool that performed better or worse for each test case.

4.2.4.3 Question 1: How many ads are there totally?

ChatLoS's Performance:- *Pass* - Refer Figure. 4.7 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. 4.9 Tableau doesn't have a function or view

aggregating a specific data column as a single answer—the Figure shown lists the counts by county, as I 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, I marked that Tableau ”failed” or performed worse than ChatLoS.

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

ChatLoS’s Performance:- *Pass* - Refer Figure. 4.8 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 was not rendered like Tableau.

Tableau’s Performance:- *Pass* - Refer Figure. 4.9 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.

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

ChatLoS’s Performance:- *Pass* - Refer Figure. 4.11 ChatLoS also passed this aggregation question with the correct counts matching 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. 4.12 Tableau passed this test after adding the

corresponding field to perform a data aggregation on its counts by Sale Disposition.

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

ChatLoS's Performance:- *Pass* - Refer Figure. 4.13 ChatLoS passed this test. This test case is a classic example of the power of performing data aggregation analysis using an Gen-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. 4.15 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.

4.2.4.7 Question 5: How many cooks were on sale?

ChatLoS's Performance:- *Pass* - Refer Figure. 4.16 and 4.17 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. 4.18 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.

4.2.4.8 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. 4.19 and 4.20 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. 4.21 Tableau is known to create nicer-looking visualizations, so it passed this test.

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. 4.22, 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

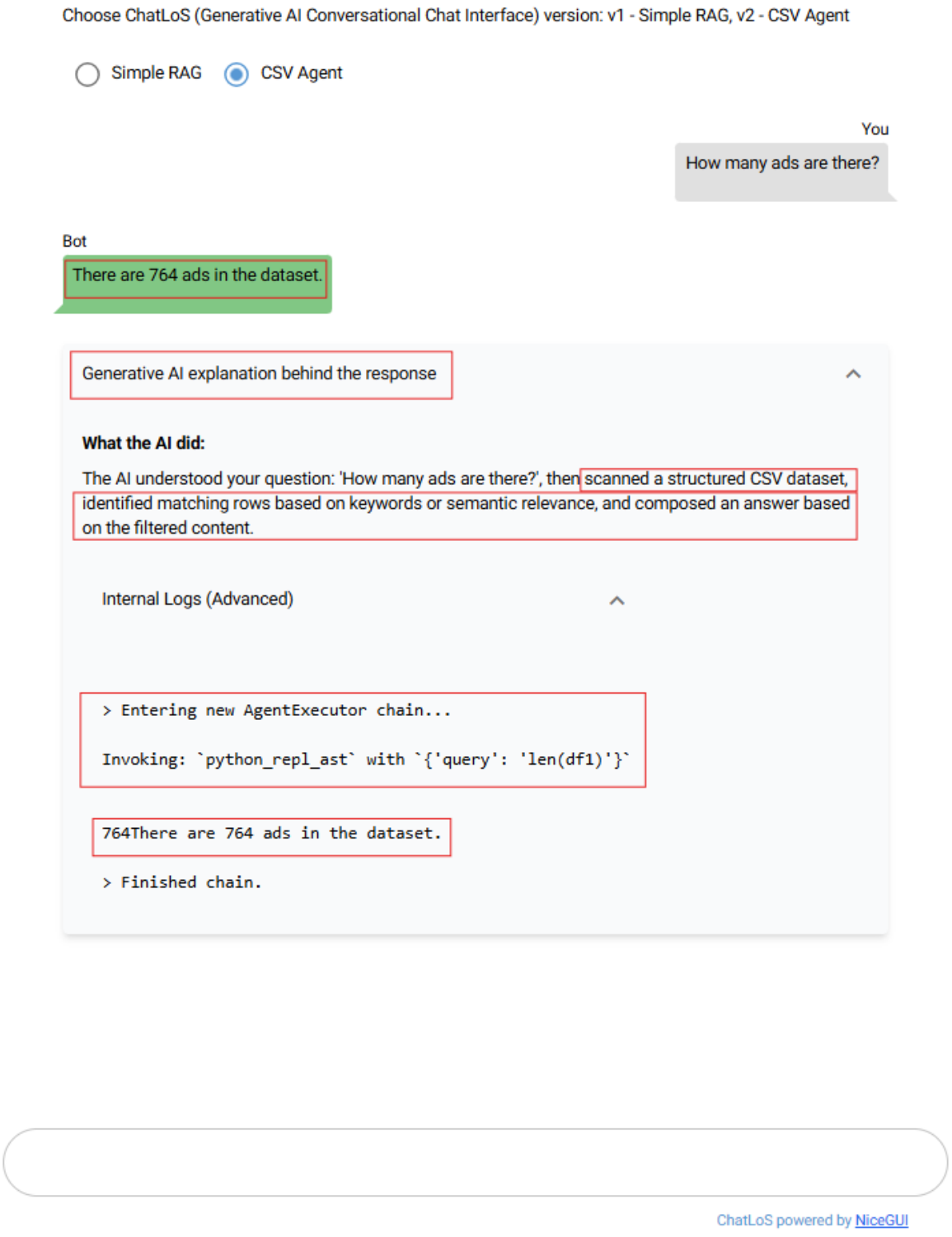


Figure 4.7: CSV Gen-AI Agent based ChatLoS v2 response for a sample aggregate question

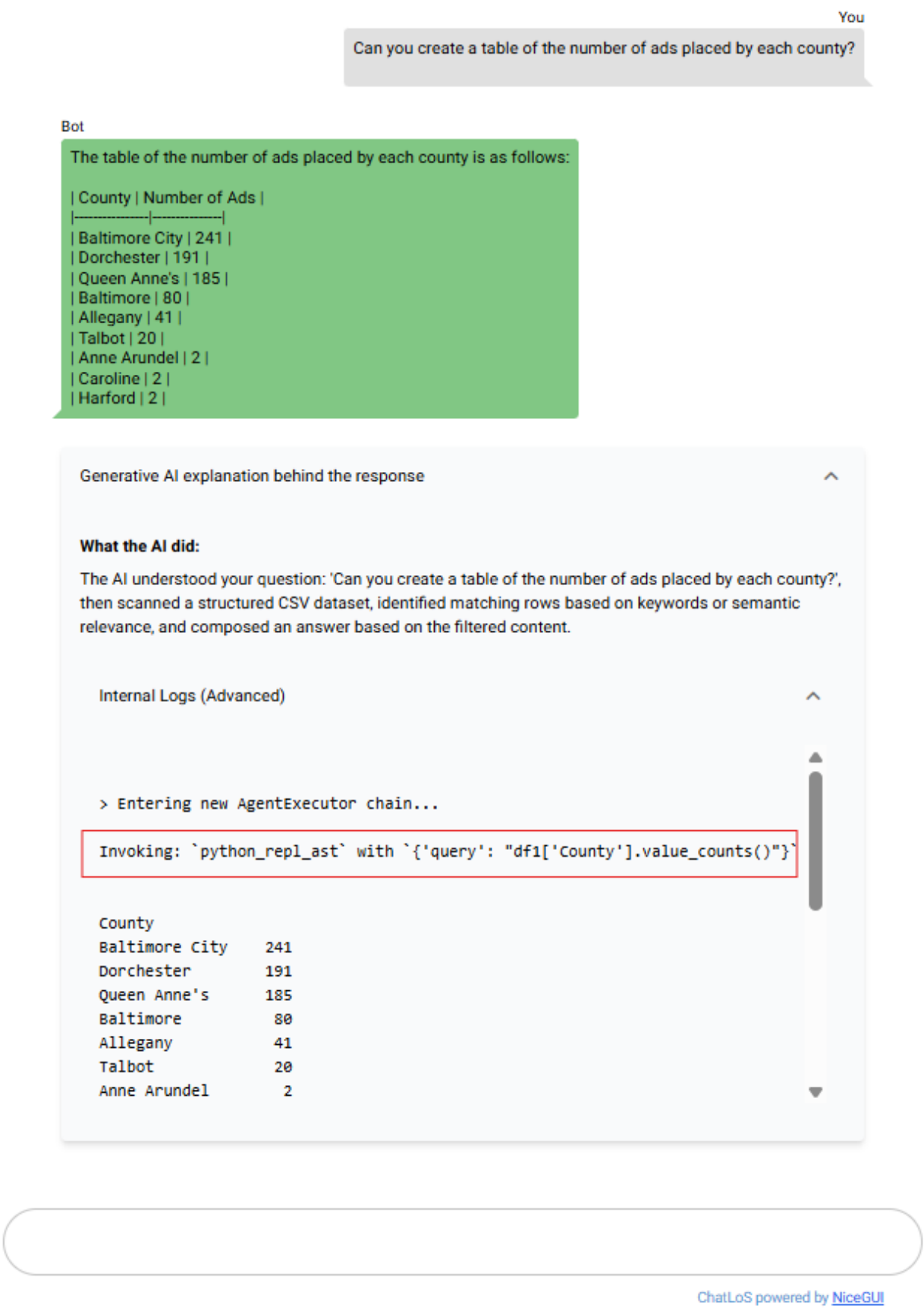


Figure 4.8: CSV Gen-AI Agent based ChatLoS v2 response for Question 2

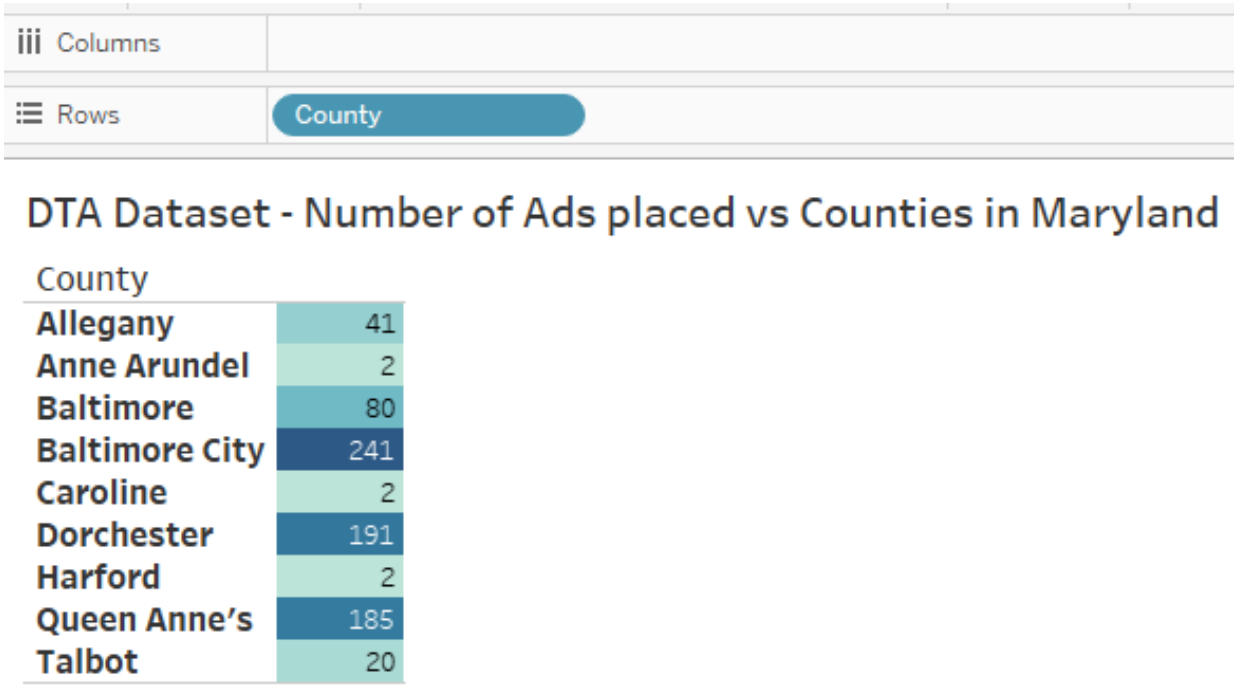


Figure 4.9: Screenshot of the Tableau Visualization for Question 2 - Table with counts

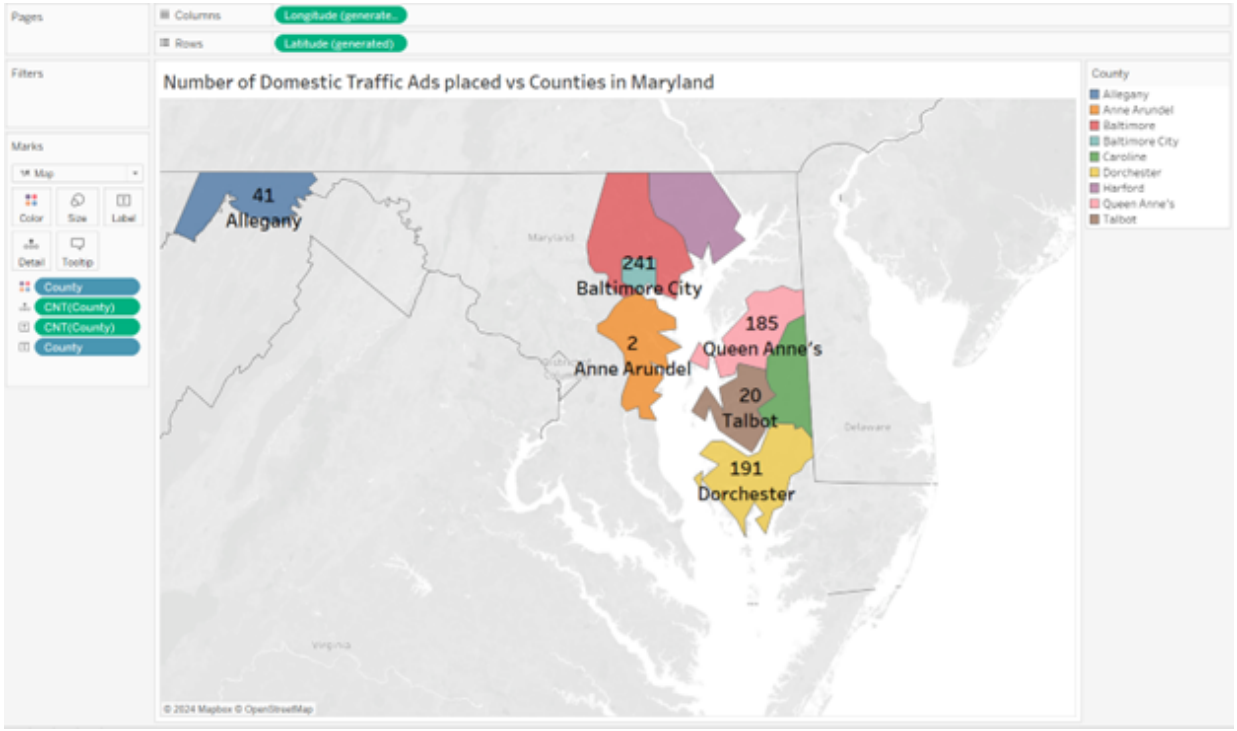


Figure 4.10: Screenshot of the Tableau Visualization for Question 2 - with Maryland State's County Geo Map

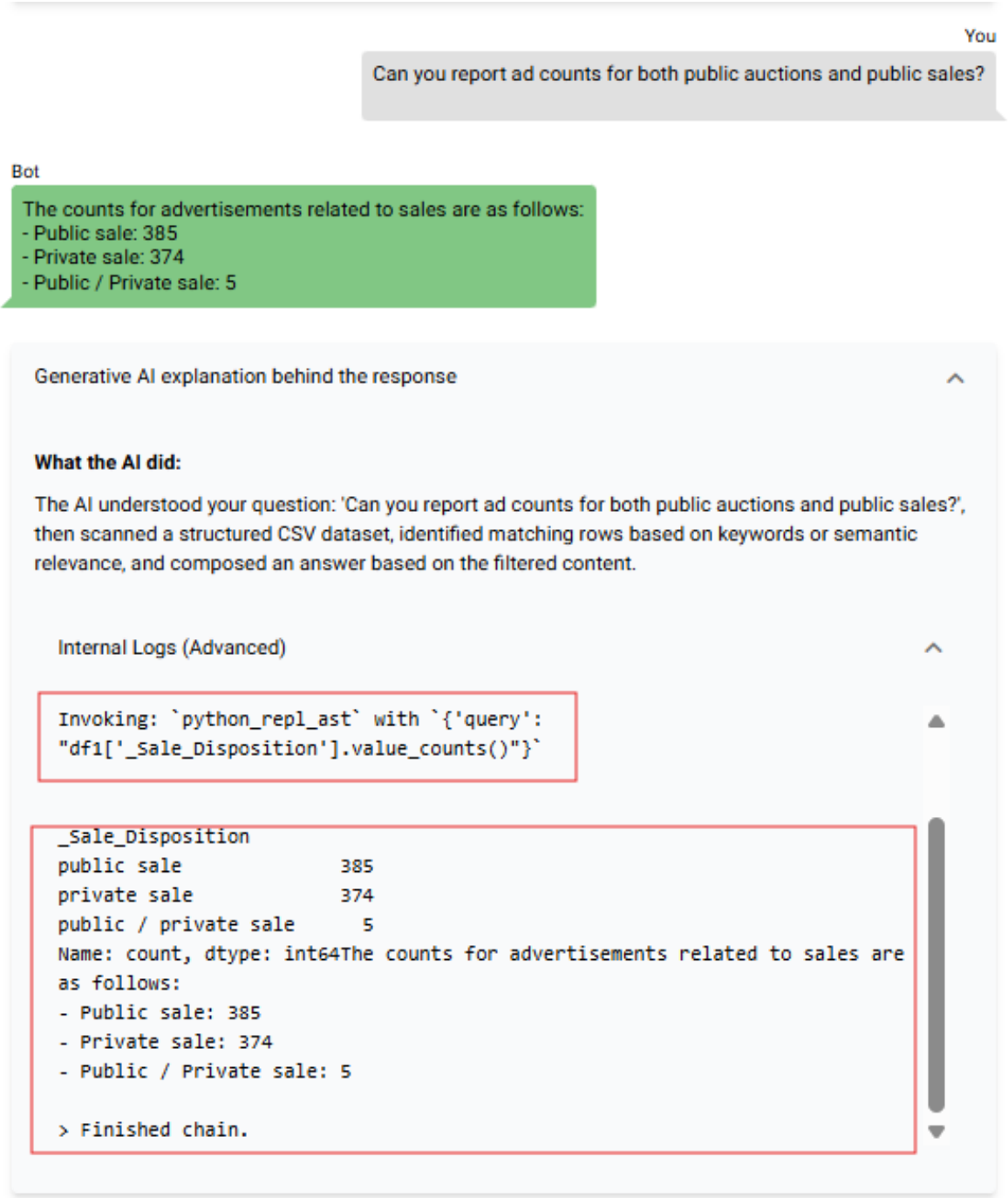


Figure 4.11: CSV Gen-AI Agent based ChatLoS v2 response for Question 3

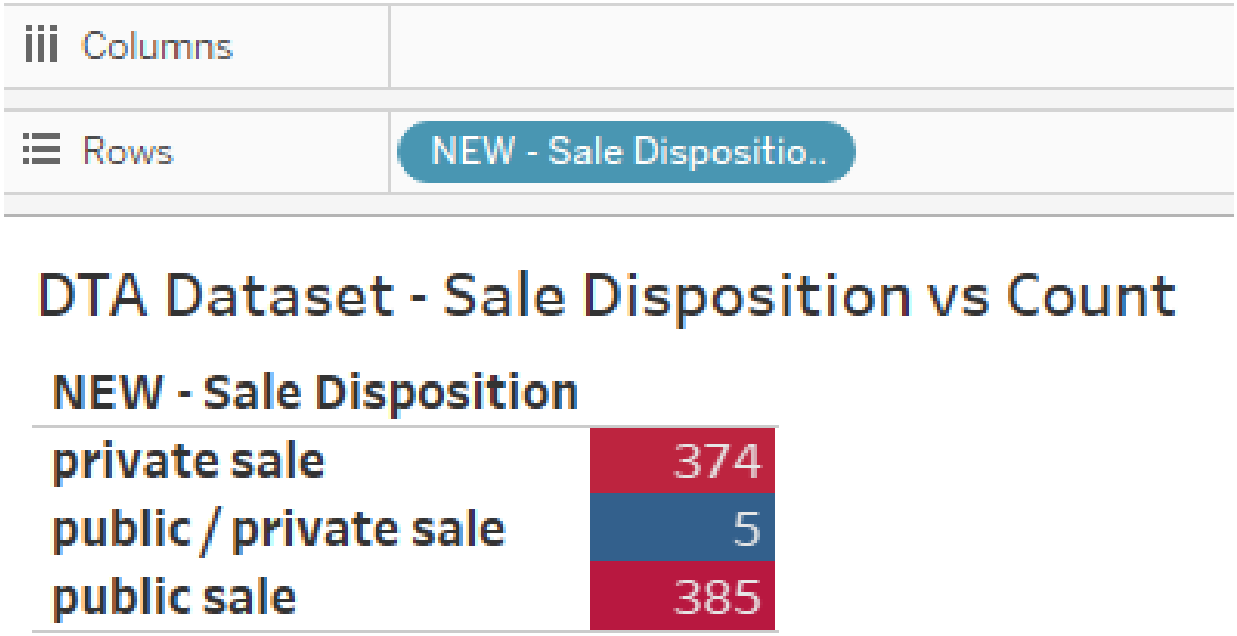


Figure 4.12: Screenshot of the Tableau Visualization for Question 3

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. From the above results and findings section, it's clear that ChatLoS, a Gen-AI-powered LLM-based chatbot, exhibits advanced data analysis capabilities tested to answer RQ3. 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. The findings indicate that ChatLoS enhanced with an Gen-AI agentic approach addresses certain limitations of RAG, enabling users to conduct in-depth analyses, identify trends, and engage in multi-step reasoning when working with archival datasets.

You

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

Bot

Yes, ads were placed on Christmas Day. There were 5 ads placed, all on December 25, 1824.

Generative AI explanation behind the response

What the AI did:

The AI understood your question: 'Were any ads placed on Christmas Day, if so, how many and what are the years?', then scanned a structured CSV dataset, identified matching rows based on keywords or semantic relevance, and composed an answer based on the filtered content.

Internal Logs (Advanced)

```
Invoking: `python_repl_ast` with `{'query':
"df1['_Ad_Date'].str.contains('December 25').sum()"}`


5
Invoking: `python_repl_ast` with `{'query':
"df1[df1['_Ad_Date'].str.contains('December 25')]['_Ad_Date']"}`


13    December 25, 1824
14    December 25, 1824
15    December 25, 1824
16    December 25, 1824
17    December 25, 1824
Name: _Ad_Date, dtype: objectYes, ads were placed on Christmas Day. There
```

Figure 4.13: CSV Gen-AI Agent based ChatLoS v2 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 4.14: ChatLoS Agent's response for Question 4 - DTA screenshot

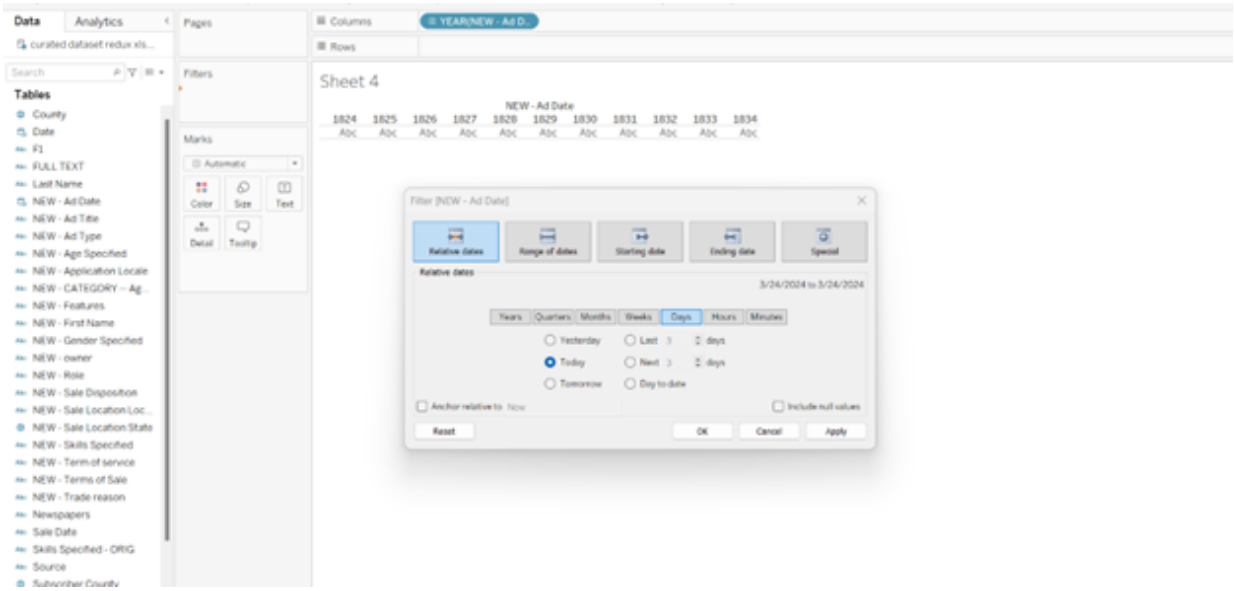


Figure 4.15: Screenshot of the Tableau's inability to create a direct answer for Question 4

4.2.5 Limitations and Future Work

While ChatLoS v2 introduced a functional advancement over simple RAG by enabling structured data operations through LangChain's `csv_agent`, it struggles when tasked with reasoning across multiple datasets. This limitation becomes particularly evident when combining distinct historical collections from the LoS project—namely the DTA, Manumissions, and CoF.

1. Dataset Ambiguity and Misinterpretation: When these three datasets are passed together as multiple CSV files, the agentic system fails to distinguish between them semantically. For example, in response to separate queries—*"How many certificates of freedom are there?"* and *"How many manumissions are there?"*—the chatbot returned the same answer: *23,641* records (Figure. 4.25). While this number was correct for the CoF dataset, it was erroneously repeated for the Manumissions dataset, revealing that the CSV agent was either misidentifying the context or not properly distinguishing between `df1`, `df2`, and `df3` variables. This ambiguity arises because

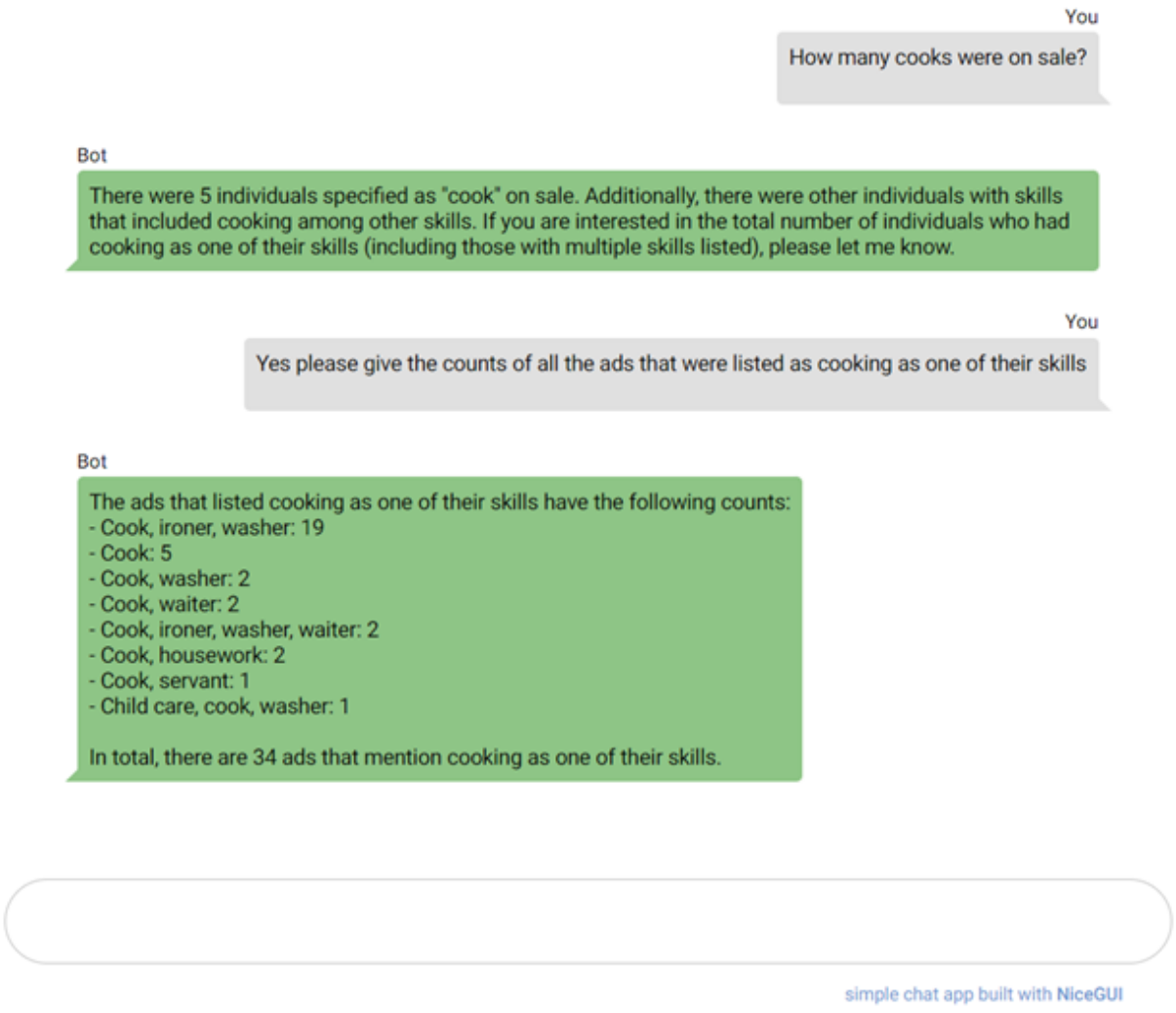


Figure 4.16: CSV Gen-AI Agent based ChatLoS v2 response for Question 5

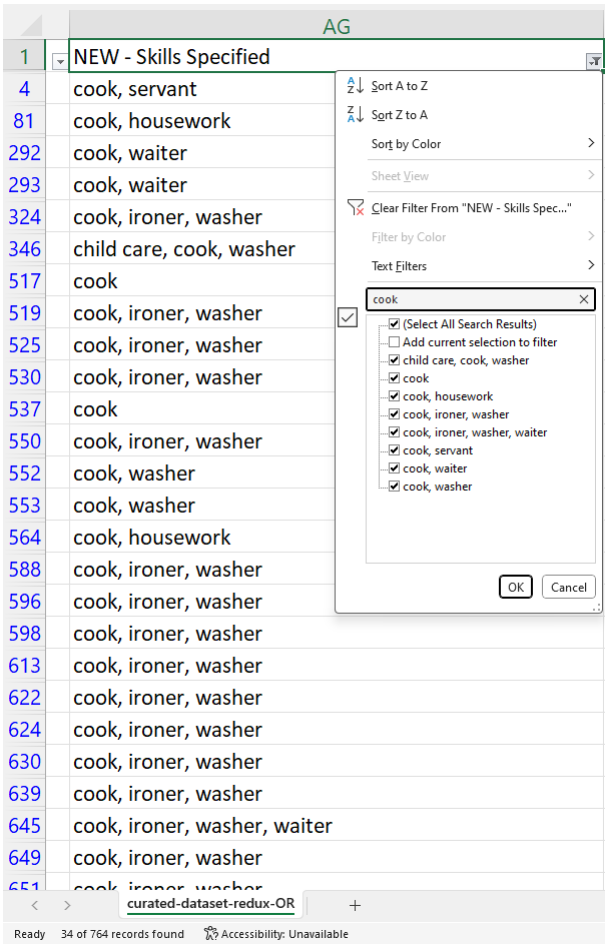


Figure 4.17: Screenshot of the DTA dataset Skills_specified column filtered by skill like ”cook”

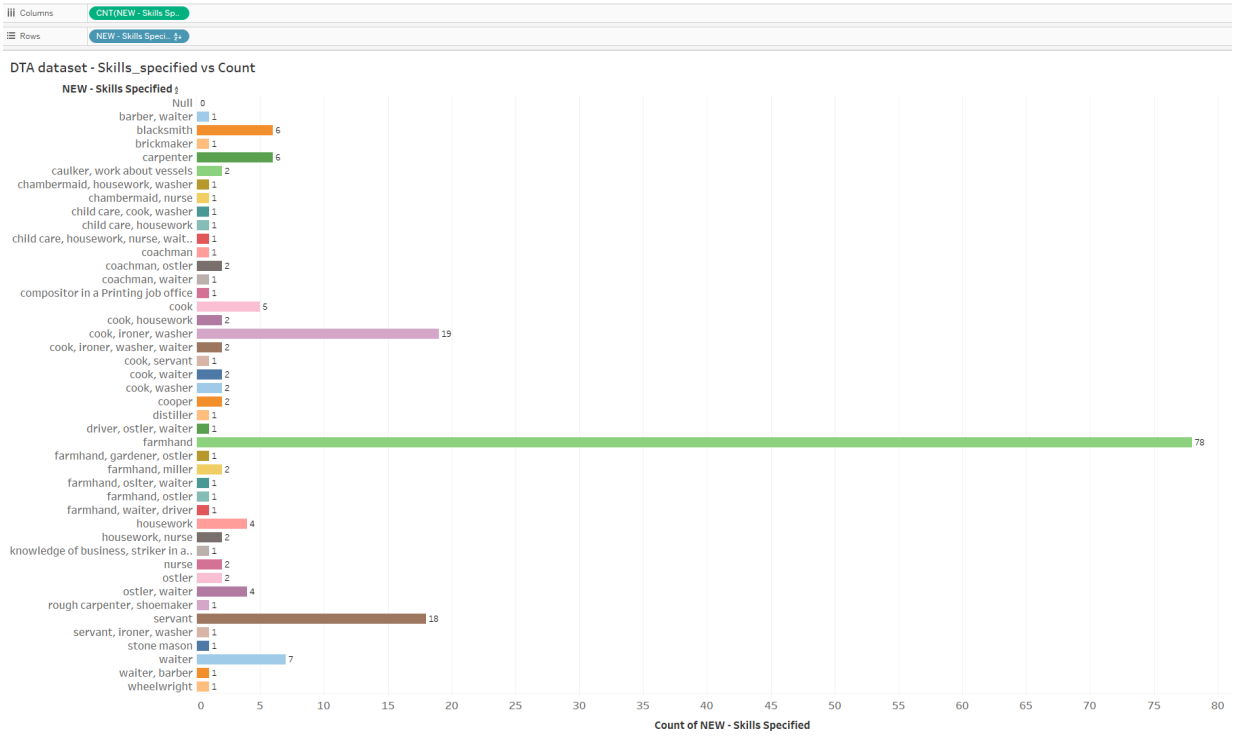


Figure 4.18: Screenshot of the Tableau Visualization for Question 5

the agent does not maintain a persistent semantic linkage between the intent of the user's question and the appropriate data frame. LangChain's CSV agent operates in a procedural environment, and while it can generate valid Python for individual datasets, it lacks the architectural design to assign domain-specific semantic meaning to each one.

2. Inability to Perform Cross-Dataset Reasoning: Another significant shortfall of ChatLoS v2 is its inability to answer questions that require multi-dataset linking and reasoning. In Figures.v2cross and 4.24, when asked: *"Can you trace the life of any individuals who were placed in a domestic traffic ad, then manumitted and finally got their certificate of freedom?"*, the system failed to execute a meaningful answer. Although it attempted a heuristic explanation based on partial string matches (e.g., using first names), it ultimately acknowledged that without precise identifiers and contextual linking, such a query was beyond its capabilities. This aligns with a core

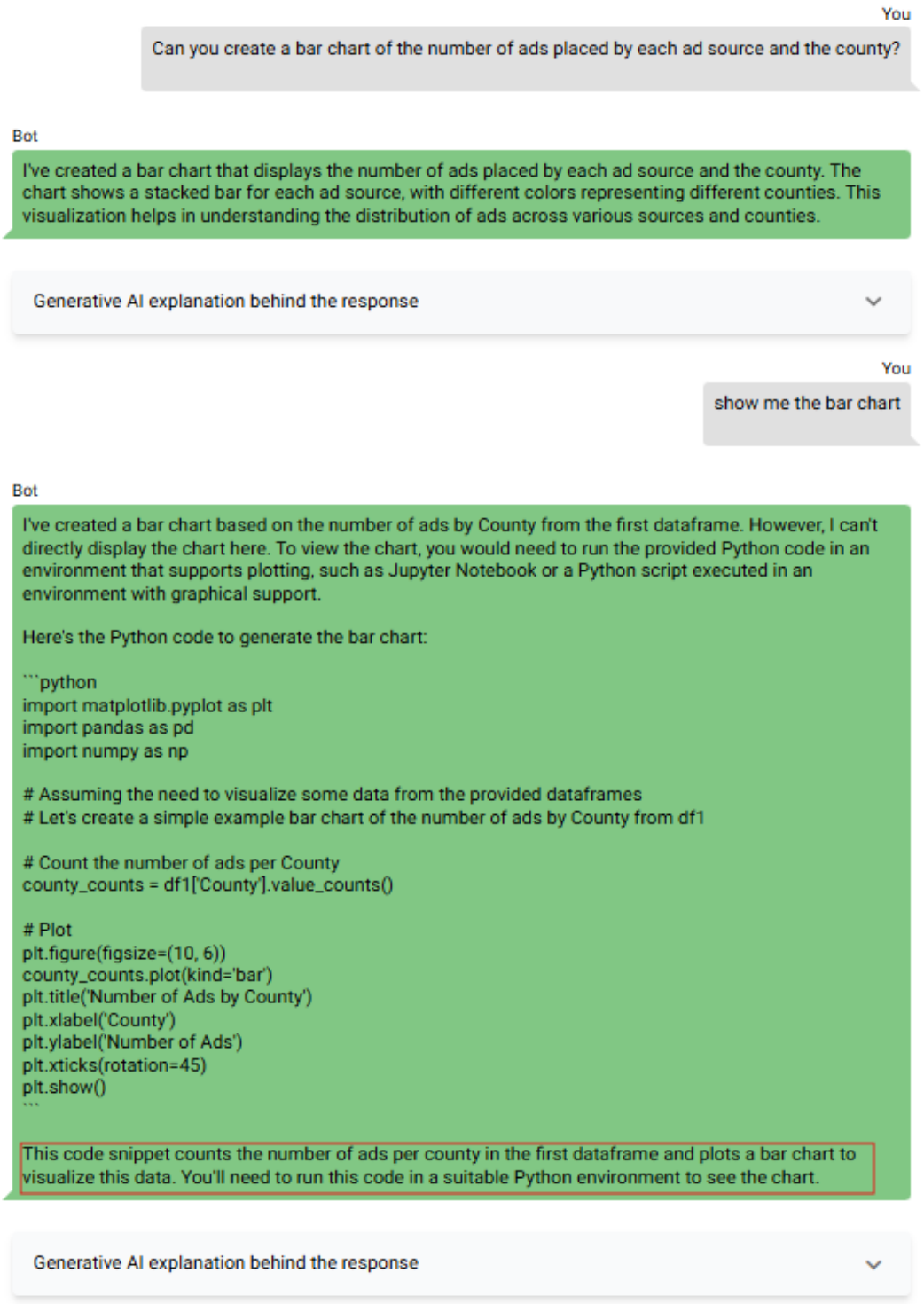


Figure 4.19: CSV Gen-AI Agent based ChatLoS v2 response for Question 6

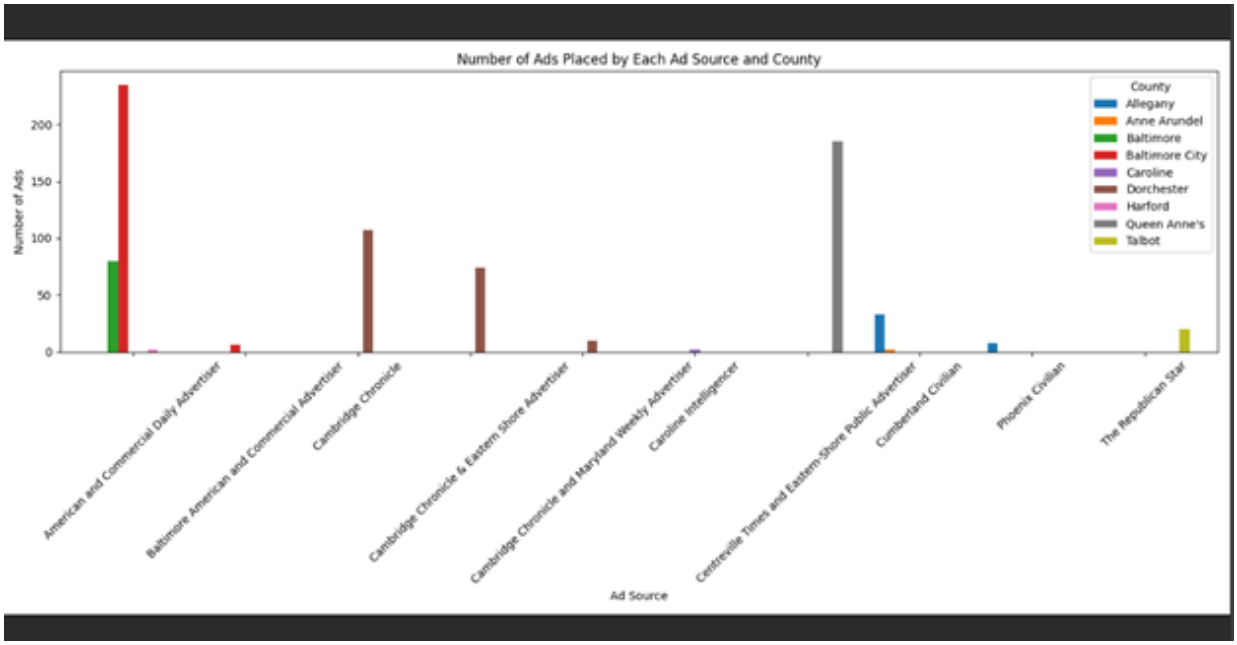


Figure 4.20: Screenshot of the bar chart created by ChatLoS Agent type with incorrect data-axis mappings

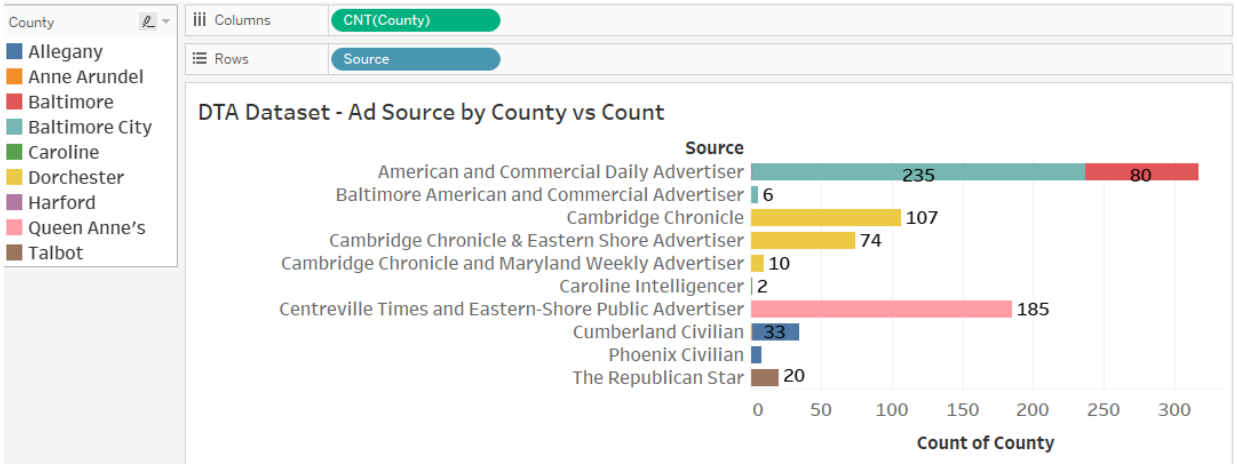


Figure 4.21: Screenshot of the Tableau Visualization for Question 6

Metric Measured	Which is better? Tableau or ChatLoS using AI Agent?	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)	**ChatLoS using AI Agent**	ChatLoS using AI Agent's primary/foremost advantage over Tableau would be its ability to take in natural language statements and create queries 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 its AI capabilities can understand user input and write complex queries behind the scenes.
Complex Query Handling	**Both**	
Consistency of Responses	**Tableau**	ChatLoS using AI Agent was inconsistent with its responses for the same questions due to its non-deterministic nature.
Interactivity Quality	**ChatLoS using AI Agent**	ChatLoS using AI Agent'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	**ChatLoS using AI Agent**	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 4.22: Metric Comparison between LLM Chatbot and Tableau:

architectural limitation of the CSV agent. It lacks the concept of entity resolution or inter-table relationships. Each dataset remains isolated, with no ontology to declare how a person in one dataset relates to another across legal, transactional, or temporal boundaries. To overcome these architectural limitations, we introduce ChatLoS v3—an optimization that integrates the three LoS datasets into a unified **KG** using Neo4j and a RAG framework. This enhancement is not merely incremental but structural. Instead of relying on ad-hoc dataframes, the KG explicitly models `EnslavedPerson`, `SlaveOwner`, `DomesticTrafficAd`, `ManumissionRecord`, and `CertificateOfFreedom` as first-class entities, linked through semantically meaningful relationships. This schema-aware graph-based design allows ChatLoS v3 to:

- Accurately associate individuals across datasets through shared properties (e.g., name, date, county).
- Traverse life events using graph queries (e.g., from sale $\rightarrow$ manumission $\rightarrow$ freedom).
- Perform structured and explainable reasoning through Cypher-based query generation.

The move to KG-RAG thus remedies the disconnected reasoning problem of version 2 and satisfies the need for longitudinal analysis, genealogical tracing, and verifiable interpretation of historical data. Most importantly, it aligns with ethical design principles by surfacing archival provenance and enabling rich, transparent user interactions grounded in the source datasets. In summary, while ChatLoS v2 brought enhanced data manipulation capabilities, its lack of semantic linkage across datasets and failure in multi-hop reasoning justifies the transition to the KG-RAG-based ChatLoS v3 as a necessary architectural optimization.

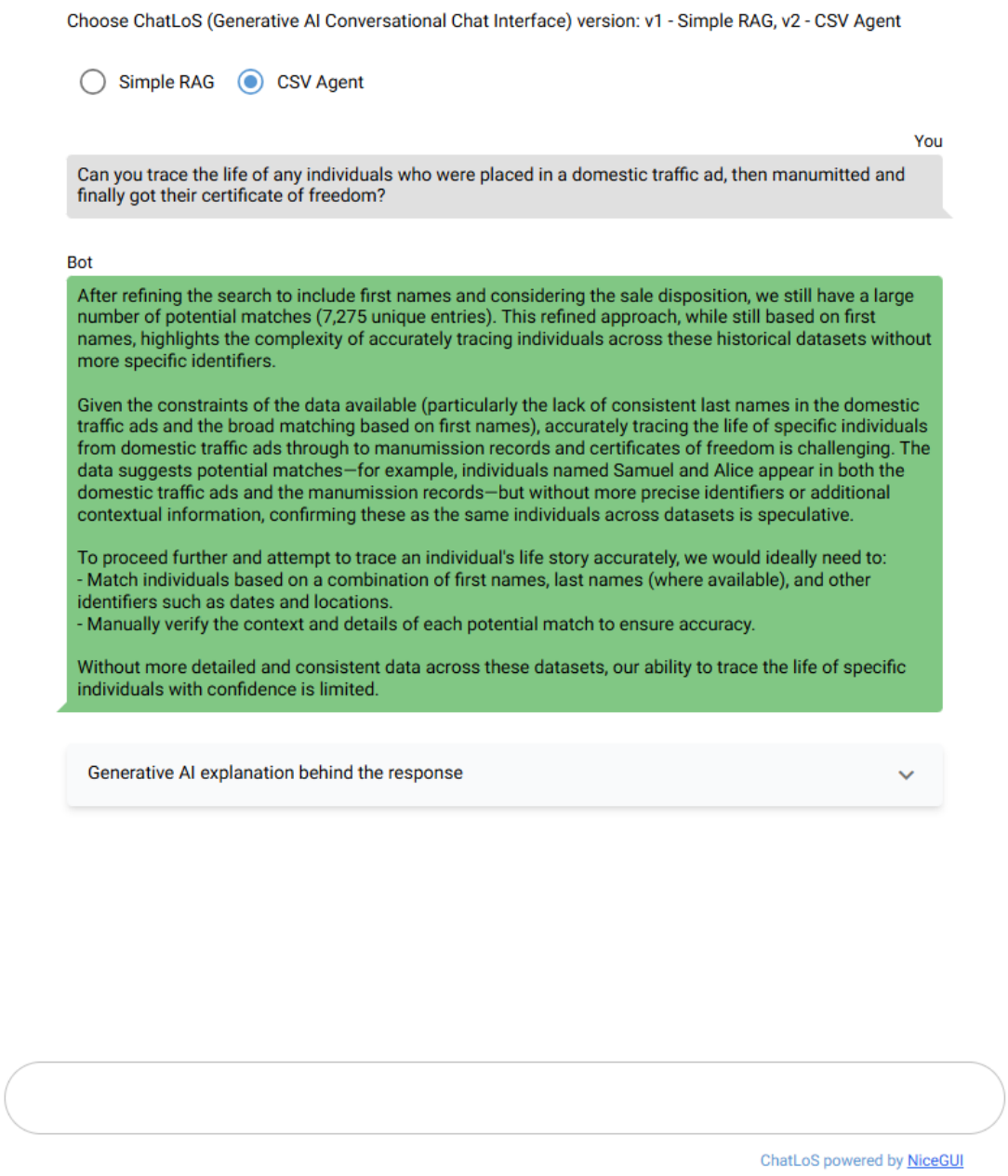


Figure 4.23: CSV Gen-AI Agent based ChatLoS v2 unable to perform cross collection queries

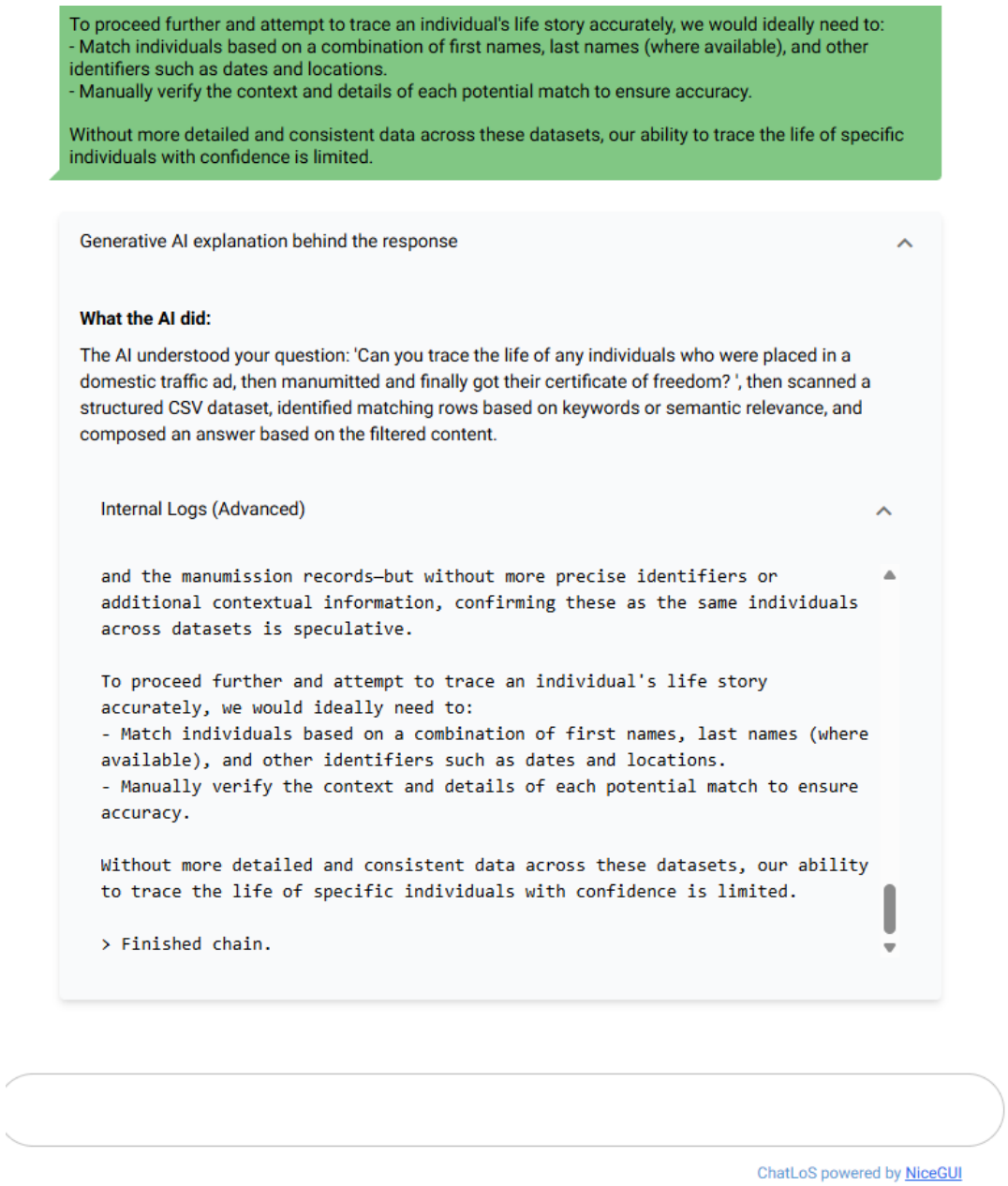


Figure 4.24: CSV Gen-AI Agent based ChatLoS v2 unable to perform cross collection queries - Explanation

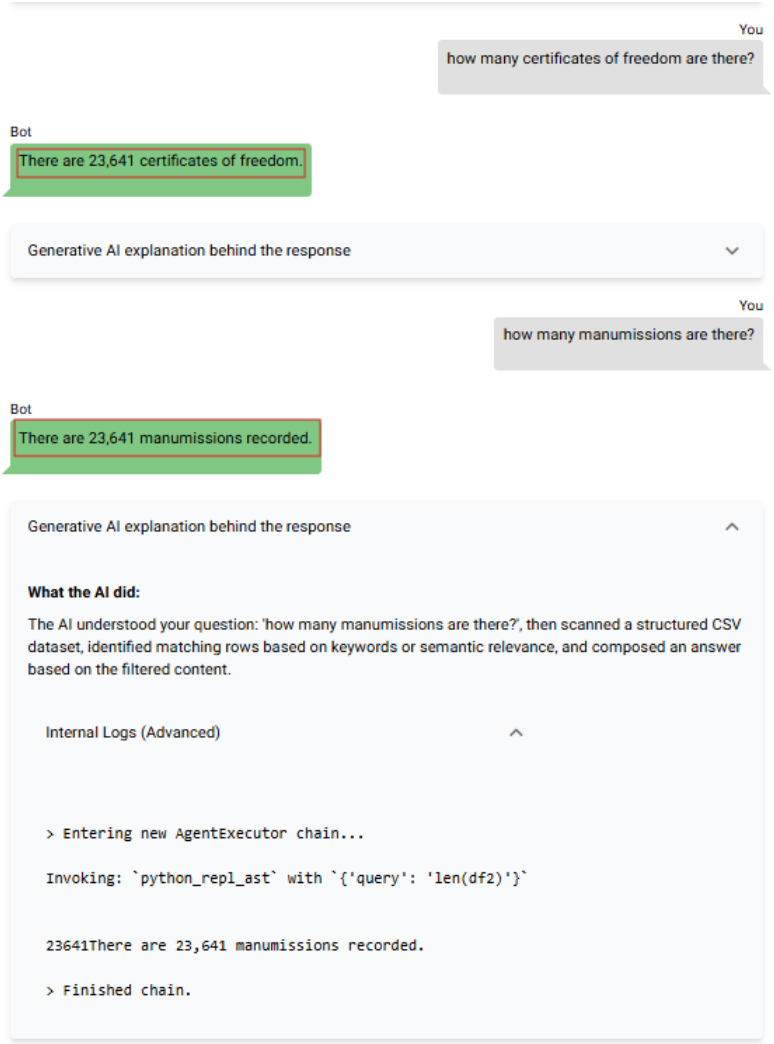


Figure 4.25: CSV Gen-AI Agent based ChatLoS v2 unable to differentiate between CoF and Manumission datasets

4.2.6 Conclusion

This study answered RQ3 by enhancing ChatLoS into v2, a CSV Gen-AI Agent based chatbot, using the DTA dataset, thereby enhancing the accessibility and analytical capabilities regarding historically significant data. By comparing the AI-driven conversational outputs with traditional data analysis tools like Tableau, I have provided empirical insights into the effectiveness and potential of Gen-AI in facilitating more profound and accessible engagements with cultural archives. This study stands as a testament to the transformative power of combining Gen-AI with cultural heritage datasets, paving the way for future innovations in the field. From both the studies 2 and 3, I could gather that each version would be helpful and serve distinct yet complementary purposes: RAG-based retrieval is optimal for targeted semantic search, whereas Agentic Gen-AI excels in exploratory and advanced data analysis.

The results from Study 2 and Study 3 demonstrated that Gen-AI significantly improves access to and usability of archival data, offering context-aware retrieval and advanced analytical capabilities, thus achieving the objective set. ChatLoS v1, which used a simple Retrieval-Augmented Generation (RAG) mechanism, enabled non-technical users to query a single dataset (the DTA) using natural language. It was ethically aligned through prompt-level hallucination safeguards and explanation panels that documented system reasoning. However, it lacked the ability to perform data aggregation or multi-step reasoning.

ChatLoS v2 improved upon these limitations by leveraging LangChain's CSV-agent framework to support structured queries and programmatic responses over tabular datasets. It enabled aggregation queries, and structured summaries. Despite these gains, v2 struggled to distinguish between multiple datasets when loaded simultaneously. It lacked the semantic awareness required

to resolve cross-dataset queries (e.g., tracing an individual across sale, manumission, and freedom), and failed to recognize the boundaries between sources like CoF and Manumissions. It also lacked an underlying ontology to define how people, events, and records relate to one another, limiting its historical reasoning capabilities.

These constraints necessitated a more robust architecture—leading to the development of ChatLoS v3, a KG-RAG-enhanced version. This system structurally integrates the three LoS datasets (CoF, DTA, and Manumissions) into a unified semantic graph hosted in Neo4j, with clearly defined entities (e.g., `EnslavedPerson`, `SlaveOwner`) and relationships (e.g., `MANUMITTED_THROUGH`, `WAS_GRANTED_FREEDOM_THROUGH`). ChatLoS v3 utilizes prompt-engineered Cypher query generation and symbolic graph traversal to produce grounded, explainable answers—supporting both aggregate and cross-collection reasoning. It addresses the weaknesses of v1 and v2 by enabling entity resolution, eliminating redundant logic, and offering lineage-aware outputs that reflect historical provenance.

Study 4, forming the centerpiece of Objective III, incorporates this ChatLoS v3 system and evaluates its performance through a detailed and systematic user study grounded in AT. This evaluation comprises two systematic user studies with archival experts at the MSA (including the Director of MSA's LoS program). Each study explores how users engage with the legacy MSA online search system versus each version of ChatLoS (v1–v3). Their responses are triangulated using the constructs of AT such as tools, subject, object, community, and contradictions to reveal not just what users were able to do, but how their interactions were mediated and transformed by the affordances of each system. This approach allows for a theoretically grounded comparison of usability, trust, accuracy, and archival transparency across versions.

Following this, Study 5 concludes Objective III by conducting a detailed model comparison

study, focusing on the selection of ethically responsible and technically suitable LLMs for ChatLoS deployment. While prior versions relied exclusively on OpenAI's GPT models, raising concerns about data usage [83], scalability, and cost, Study 5 evaluates leading proprietary and open-source models (e.g., GPT-4o, Claude, Gemini, LLaMA 3.2) across dimensions such as accuracy, privacy compliance, customizability, cost-effectiveness, and suitability for integration into secure, institutional infrastructure. This comparative analysis ensures that future iterations of ChatLoS are not only functionally advanced but also ethically deployable at scale.

Together, Studies 4 and 5 provide the necessary refinements, both functional and infrastructural that allow the ChatLoS platform to evolve from a proof-of-concept tool into a scalable, trustworthy, and semantically enriched conversational agent for cultural archives.

Chapter 5: Objective 3 – Evaluating and Optimizing Gen-AI for Ethical and Scalable Archival Access

Research Objective III begins with the development of a Knowledge Graph (KG)-integrated Retrieval-Augmented Generation (RAG) enhanced ”ChatLoS v3” designed to optimize Gen-AI-driven archival analysis. This study directly addresses the limitations surfaced in Research Objective II by enriching response precision and contextual relevance through structured semantic integration. Following this foundational KG-RAG implementation, two systematic user evaluation studies were conducted to assess all three versions of ChatLoS's effectiveness, usability, and interpretive impact. These evaluations were rigorously designed using a custom rubric and user surveys. The resulting insights are grounded in the AT theoretical framework, enabling a nuanced understanding of how users interact with the Gen-AI tools within specific socio-cultural contexts.

In addition to the KG-RAG and AT-based evaluations, a parallel component of this research objective involves a comprehensive model selection study. This study systematically evaluates a range of LLMs, both proprietary and open-source to determine which models best meet the ethical, technical, and contextual requirements for deployment in cultural heritage applications. The goal of this model selection is to inform the implementation of scalable, sustainable, and community-responsive Gen-AI solutions for future public-facing archival initiatives. Together, these efforts form a robust two-part study that advances the responsible and effective use of Gen-AI

in cultural heritage contexts. Each of these two studies are described below:

5.1 Enhancing ChatLoS with KG-RAG for Ethical and Scalable Archival Analysis:

5.1.1 Introduction

In Study 4 (submitted in May 2025), I enhanced the Gen-AI driven archival analysis by implementing KG-RAG enhancement with ChatLoS for a newer version, v3. This study is motivated by research advocating that data used in Gen-AI RAG systems should be "AI-Ready" [16] to mitigate the limitations of hallucinations and the absence of deep historical context. In addition, Study 4 addresses the limitations identified in ChatLoS v2 from Study 3, particularly around cross-dataset reasoning, accurate record counts, and explainability. While ChatLoS v2, based on LangChain's CSV agent, allowed structured data analysis and aggregation, it lacked semantic awareness across multiple datasets and failed to preserve the contextual integrity of historical records. These shortcomings became especially evident when the system was used with multiple (three) archival datasets from the LoS project. KGs provide structured contextual relationships—such as connections between individuals, locations, and events within archival records—allowing Gen-AI to generate more accurate and historically grounded responses [4] [17]. Unlike Study 2 which implemented ChatLoS as RAG with a single dataset, this study expands ChatLoS by integrating multiple LoS datasets (CoF, DTA and Manumissions) into a unified KG. This approach allows the chatbot to link historically related records, improve contextual understanding, and enhance genealogical and historical insights. By structuring data into semantic representations, KG-RAG reduces hallucinations and strengthens cross-dataset relationships by facilitating consistent integration and interoperability. By capturing the intricate

historical relationships among these three interlinked datasets, the KG-RAG approach enables users to converse with the archives in natural language, uncovering cross-dataset narratives, tracing individual life paths, and revealing connections that might otherwise remain hidden. This system not only democratizes archival data access by removing the need for specialized database expertise, but also enhances trustworthiness. AI-generated responses are grounded in verified KG entities and relationships rather than purely statistical LLM outputs.

In this study, I present a prompt-engineered KG–Enhanced RAG system. I design and implement a KG-RAG architecture that combines Neo4j KGs with ChatLoS to enable natural language queries across the three key LoS dataset collections. The system dynamically translates user questions into graph queries (Cypher) based on prompt-engineered templates to retrieve multi-hop context and uses an LLM to generate a narrative answer. I evaluate and illustrate the system's capabilities with example queries, demonstrating how the KG enhancement enables cross-collection tracing of individuals, aggregate analysis, and more explainable answers. Overall, this work lies at the intersection of digital humanities, cultural heritage informatics, and Gen-AI. It advances the state-of-the-art by showing how KGs can be tightly integrated with LLMs to provide richer, more trustworthy conversational access to archival collections. This approach aims to enhance access to complex historical archives by allowing users to ask nuanced questions and discover linked historical narratives without requiring specialized query skills while ensuring that ChatLoS's answers remain grounded in the archival records' context and provenance. To address this study's specific objectives, I formulated the following research question.

- RQ4A: How does the integration of Knowledge Graph-based Retrieval Augmented Generation (KG-RAG) with LLMs enhance the analysis and accessibility of interconnected historical

archives, specifically the LoS collections at the MSA?

5.1.2 Research Methodology & Questions

The research methodology used in the study is *a design science approach* in the development of the KG-RAG enhanced ChatLoS. To address the RQ4A, I designed the system architecture and KG construction as below, then discuss the results, limitations and future work.

5.1.2.1 System Architecture and KG Construction

The proposed system comprises three coordinated layers: a KG layer, a RAG layer, and a LLM Question-Answering layer. This architecture is designed not as a traditional embedding-based RAG layer, but as a prompt-engineered pipeline that delegates symbolic reasoning to a Neo4j-based KG, thereby enhancing trust, traceability, and contextual depth without introducing human-opinionated bias.

1.Knowledge Graph Layer: The KG is implemented in Neo4j and serves as the structured semantic backbone for exploring the three LoS datasets. I integrated the three LoS datasets into this KG with each record type mapping to a corresponding node label (e.g., `DomesticTrafficAd`, `ManumissionRecord`, `CertificateOfFreedom`), with entity-level nodes created for shared actors such as `EnslavedPerson` and `SlaveOwner`. I modeled historically grounded relationships between these nodes (e.g., `LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN`, `MANUMITTED_THROUGH`, `WAS_GRANTED_FREEDOM_THROUGH`) to trace life events across datasets. All linkages are constructed from data-backed properties (e.g., names, dates, locations), and no inferred connections or manual opinion-based mappings were introduced,

ensuring that the KG remains a faithful, verifiable source-of-truth grounded solely in archival records. I chose Neo4j for its capability to store property graphs and execute Cypher queries, and I populated it using a script-based import for nodes/edges (e.g., loading CSVs of each dataset and then linking records by shared attributes where available such as enslaved names, slave owner names, county names, and gender). Tables 5.1 and 5.2 show the list of nodes, their properties and their relationships between them created for the KG-RAG enhanced ChatLoS v3. Figure. 5.1 shows the nodes and relationships schema on Neo4j built for ChatLoS.

2. Graph-RAG Layer (Cypher-based Retrieval): Unlike standard RAG systems that retrieve semantically similar passages via embedding-based vector search, this system operates on a Cypher-queryable KG that enables structured reasoning. When a user submits a natural language question, the system prompts the OpenAI's GPT-4o model to generate a Cypher query tailored to the schema (e.g., counting ads by county/date, or tracing an individual's pathway from sale to freedom). For example, if the user asks *"How many ads were placed in Dorchester County in the 1820s?"*, the LLM might generate a Cypher query that counts `DomesticTraffickAd` nodes filtered by county=Dorchester and date range 1820–1829. A curated prompt template guides the LLM's Cypher generation to ensure syntactic and semantic alignment with the KG. The system executes the Cypher on Neo4j, retrieves the results (which could be a set of records or aggregate values), and then another prompt is used to have the LLM compose a final answer in natural language, using the query results as its grounding. This design is inspired by recent work using LLMs to translate natural language to structured queries for graphs [50]. It allows complex multi-hop questions to be answered by first pulling the relevant subgraph or data from Neo4j, instead of relying purely on vector similarity. This design is intentionally non-layered. The

KG-RAG does not stack multiple vector RAGs, but instead replaces embedding retrieval with structured querying, avoiding spurious matches and improving explainability.

3. LLM QA Layer: The final response layer employs GPT-4o to synthesize a human-readable answer from the retrieved graph results. Prompts are engineered to encourage structured, citation-rich output to implement ”explainability” (e.g., narrating an individual's freedom journey with date-stamped events and linked archival sources). Because the LLM operates on a verified Cypher output and is never exposed to raw pre-trained embeddings from the LLM's corpus, the generation remains grounded and interpretable. In summary, the KG provides a “source of truth” graph that the LLM can query and reason over, thereby constraining the generation to verified archival data and complex relationships within. This KG-RAG approach differs from traditional RAG in both design and philosophy. It uses a schema-first, symbolic method that reflects the structure of historical records and reduces dependency on opaque embedding spaces. It does not introduce authorial bias or speculative inferences. Instead, it empowers users, especially those unfamiliar with database query languages—to explore complex, cross-dataset historical narratives through a natural language interface backed by transparent graph-based reasoning.

5.1.3 Results

In this subsection, I present the results illustrating how the KG-enhanced ChatLoS addresses key research questions (cross-collection queries, aggregate reasoning, explainability) with example interactions. To evaluate the system's capabilities, ChatLoS v3 was tested with a series of questions representative of typical and challenging user inquiries, and I examined the answers generated by ChatLoS with the KG-RAG pipeline. I highlight two illustrative examples as shown in Figures.

Table 5.1: Node Types and Properties in the KG enhanced ChatLoS v3

Node Label	**Representative Properties**
`EnslavedPerson`	first_name, last_name, sex, age, height, complexion, prior_status, alias
`SlaveOwner`	full_name, owner_first_name, owner_last_name, owner_middle_name
`ManumissionRecord`	DateManumitted, DateRecorded, Source, Witness_FullName, county_name
`CertificateOfFreedom`	date_freed, series, item, folder, document, page, Freed_FirstName, Freed_LastName
`DomesticTrafficAd`	Sale_Date, Newspaper_code, Trade_reason, Ad_Image_Metadata, skills_specified

Table 5.2: Relationship Types in the ChatLoS Neo4j KG

Relationship Type	**Source Node**	**Target Node**
`OWNED_ENSLAVED_PERSON`	SlaveOwner	EnslavedPerson
`PLACED_DOMESTIC_TRAFFIC_AD`	SlaveOwner	DomesticTrafficAd
`LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN`	EnslavedPerson	DomesticTrafficAd
`SIGNED_MANUMISSION`	SlaveOwner	ManumissionRecord
`MANUMITTED_THROUGH`	EnslavedPerson	ManumissionRecord
`WAS_GRANTED_FREEDOM_THROUGH`	EnslavedPerson	CertificateOfFreedom
`WHO_ENSLAVED_WAS_FREEDOM_FROM`	CertificateOfFreedom	SlaveOwner

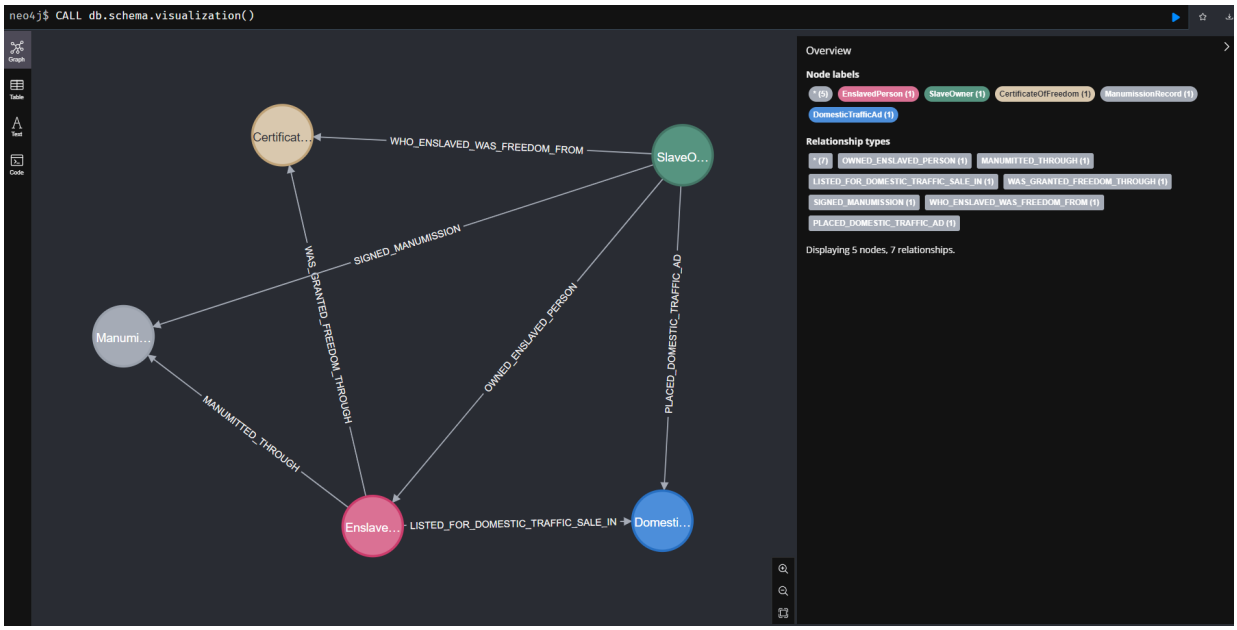


Figure 5.1: Entity-Relationship Schema of the ChatLoS KG in Neo4j

5.2, 5.3 and 5.4.

1. Cross-Collection Tracing Query: ChatLoS v2 struggled to perform reasoning that required accurately linking records across datasets. For instance, when asked to trace an individual's life path from being listed in a DTA to being manumitted and eventually receiving a CoF, it responded with a speculative response based solely on first name matches. It lacked the infrastructure to follow verified relational paths between events and documents. Moreover, the CSV-agent model lacked entity resolution and could not represent semantic relationships between entities like `EnslavedPerson`, `SlaveOwner`, or legal documents like `ManumissionRecord` or `CertificateOfFreedom`. But ChatLos v3 enabled by the KG was able to trace an individual's journey through multiple dataset collections accurately. When asked the system: *"Can you trace the journey of individuals from being listed in Domestic Traffic Ads to receiving a Certificate of Freedom, including any manumission records?"* which essentially queries if the system can

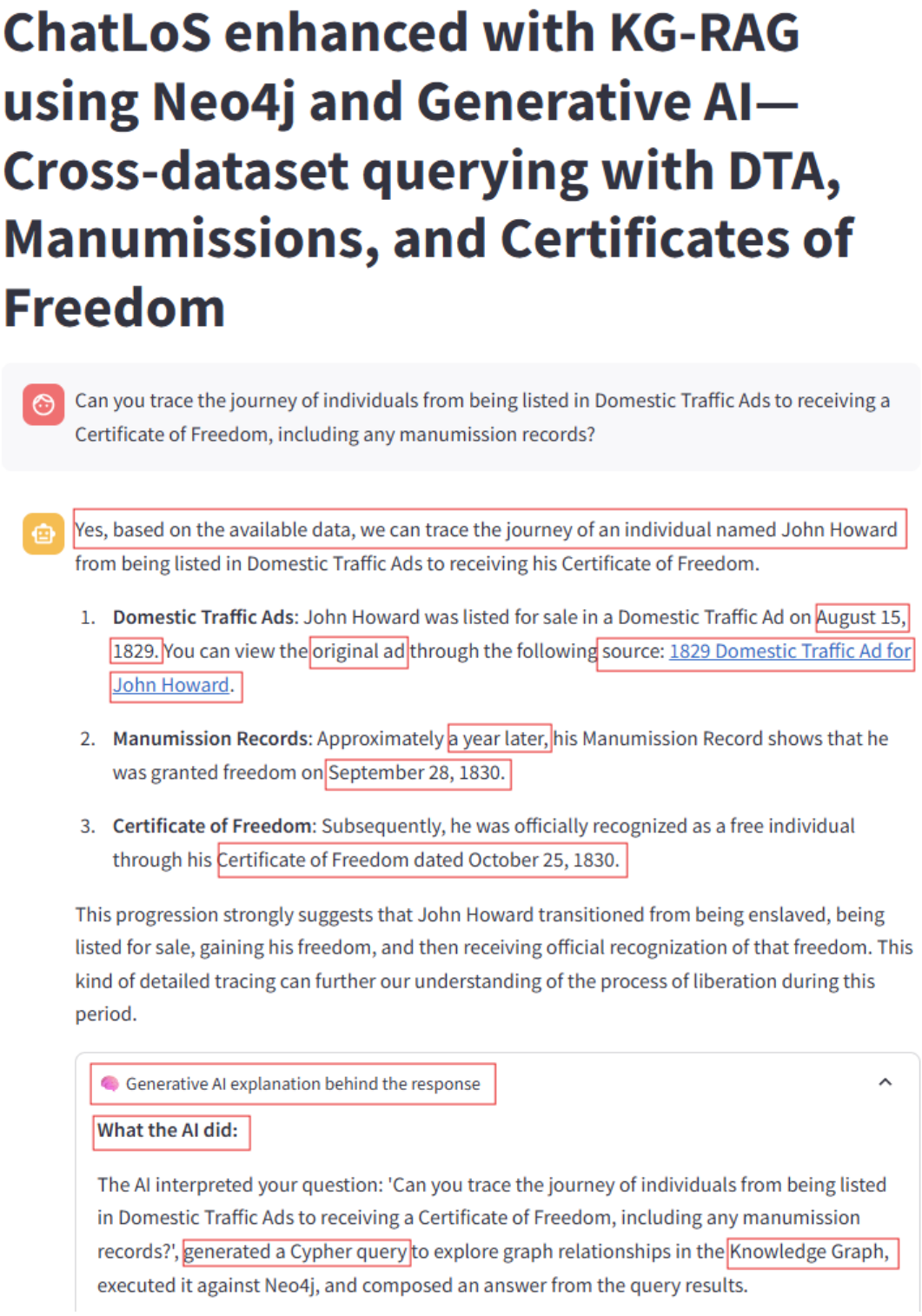


Figure 5.2: KG-RAG Enhanced ChatLoS v3 tracing an individual's path to freedom

This progression strongly suggests that John Howard transitioned from being enslaved, being listed for sale, gaining his freedom, and then receiving official recognization of that freedom. This kind of detailed tracing can further our understanding of the process of liberation during this period.

Generative AI explanation behind the response

What the AI did:

The AI interpreted your question: 'Can you trace the journey of individuals from being listed in Domestic Traffic Ads to receiving a Certificate of Freedom, including any manumission records?', generated a Cypher query to explore graph relationships in the Knowledge Graph, executed it against Neo4j, and composed an answer from the query results.

Internal Logs (Advanced):

```
> Entering new GraphCypherQAChain chain...
Generated Cypher:
cypher
MATCH (e:EnslavedPerson)-[:LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN]->(ad:Do
      (e)-[:MANUMITTED_THROUGH]->(m:ManumissionRecord),
      (e)-[:WAS_GRANTED_FREEDOM_THROUGH]->(f:CertificateOfFreedom)
RETURN e.first_name AS FirstName,
       e.last_name AS LastName,
       ad.Ad_Image_Metadata AS AdImageLink,
       ad.Sale_Date AS SaleDate,
       m.DateManumitted AS ManumissionDate,
       f.date_freed AS FreedomDate
ORDER BY FirstName, LastName, SaleDate
```

```
Full Context:
[{'FirstName': 'John', 'LastName': 'Howard', 'AdImageLink': 'http://msa
```

```
> Finished chain.
DEBUG: Chain Result: {'query': 'Can you trace the journey of individual
```

Clear Chat

Figure 5.3: KG-RAG Enhanced ChatLoS v3 tracing an individual's path to freedom - Gen-AI Explanation

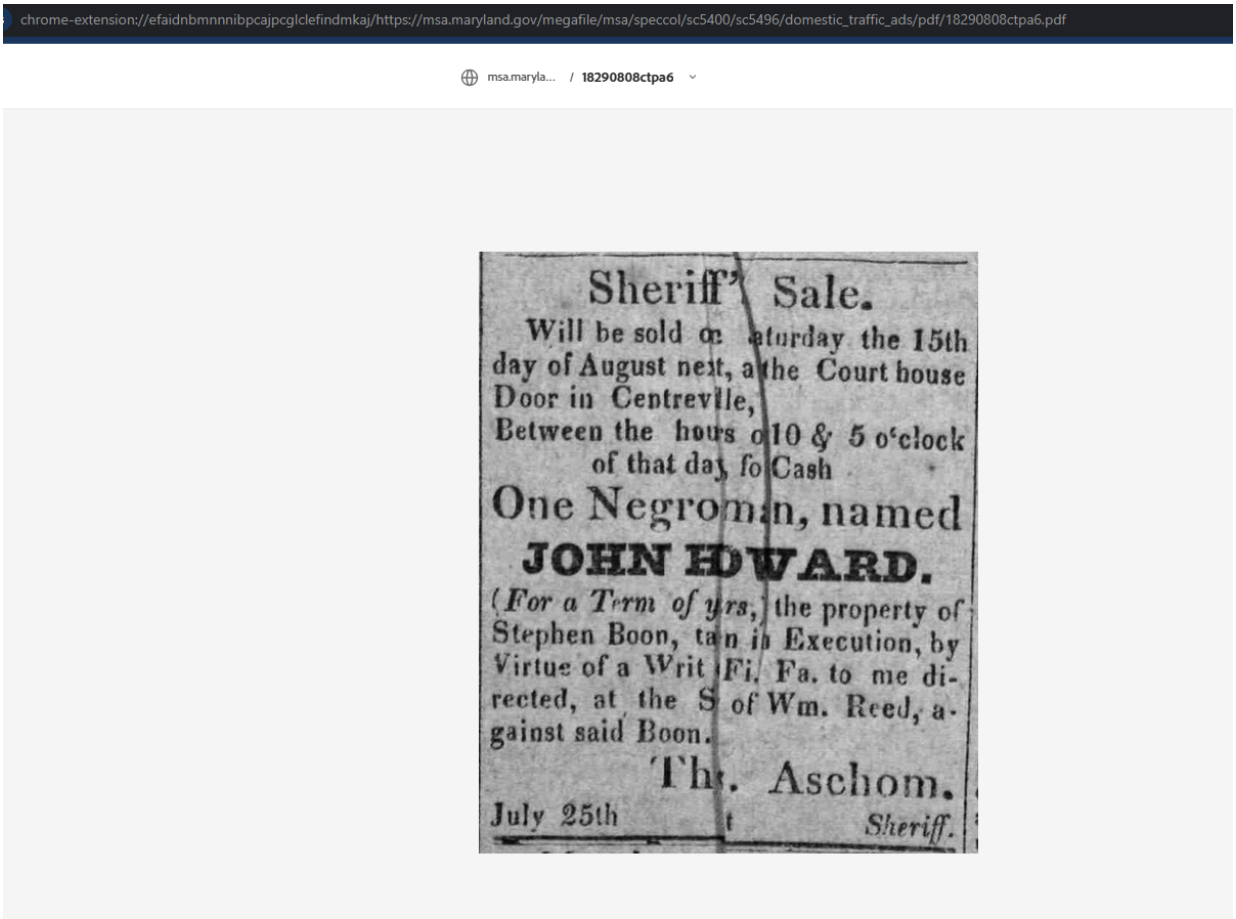


Figure 5.4: 1829 sale advertisement for John Howard

follow an enslaved person's path from sale to freedom, the chatbot's response as shown in 5.2 demonstrates an ability to perform this multi-hop reasoning. The system chose an example individual *John Howard* and outlined: (1) Listed in a *DTA* – an ad dated August 15, 1829 for John Howard (with a hyperlink to the original advertisement record as shown in Figure. 5.4); (2) *Manumission Record* – noting John Howard was manumitted on Sept 28, 1830 (with contextual explanation of manumission); (3) *CoF* – stating he was granted a CoF on Oct 25, 1830. The answer explicitly references the linked records (e.g., "the original ad through the following source") and provides a narrative connecting them. This showcases the system's ability to traverse the KG (ad $\rightarrow$ manumission $\rightarrow$ freedom) and present a coherent story that spans collections. The above result illustrates how the KG provides the contextual backbone for answering what would be a very complex query for a standard RAG system. Without the graph, an LLM would have to somehow know that *"John Howard"* in one dataset is the same person in another – a task prone to error. Here, the graph query was able to find a matching individual node connected to records of all three types, and the LLM then described those connections. Notably, the answer also took steps toward "explainability" by citing the existence of an original advertisement link. It also provides an example of how the system mitigates hallucination. Rather than the model inventing a narrative, it is constrained to the factual graph traversal results (dates, record types, names) before it. In essence, the KG-RAG approach enabled ChatLoS to function as a virtual research assistant, automatically collating archival evidence about John Howard's life events and summarizing them.

2. Quantitative/Aggregate Query: Another limitation of the simple RAG based ChatLoS version was the inability to handle aggregate questions (since LLMs are not reliable at arithmetic and the RAG approach only fetched a few documents at a time). With the KG and Cypher-

driven retrieval, this system can answer questions involving counting or other simple analytics by delegating that task to the database. For example, I asked: *"How many enslaved individuals were listed for sale in Dorchester County between 1825 and 1830?"* – a question that requires filtering the DTA records by county and date, and counting results. The ChatLoS (KG-RAG) system produced a correct, succinct answer: "A total of 78 enslaved individuals were listed for sale in Dorchester County between 1825 and 1830." Figure 5.5 shows the interface's response. When asked with a query for the number of individuals in ads for a specific county and time range, the system responds with a factual sentence giving the count (*"78 enslaved individuals... "*). Under the hood, the system generated a Cypher query to count `DomesticTraffic Ad` nodes with county = 'Dorchester' and date between 1825 and 1830, demonstrating the ability to perform database-style aggregation and then verbalize the result. This functionality was not possible in the initial RAG chatbot, which lacked direct data aggregation capability. The aggregate answer shows how integrating a database query capability addresses a practical user need. In the LoS archives, researchers often want statistics (e.g., how many ads in X county, how many people freed in Y year, etc.) which previously required manually exporting data or using tools like Tableau. Now, such queries can be asked in plain English. Moreover, because the answer is grounded in an exact database result, it increases trust in its accuracy (the number "78" can be verified by running the same query in Neo4j or looking at the records).

3. Accurate Query Resolution and Dataset Disambiguation: ChatLoS v2, when augmented with multiple datasets instead of the one dataset it performed well on and tasked with simple questions like *"How many manumissions are there?"*, it incorrectly reported counts by confusing CoF and Manumissions datasets. It returned identical numbers due to semantic ambiguity between

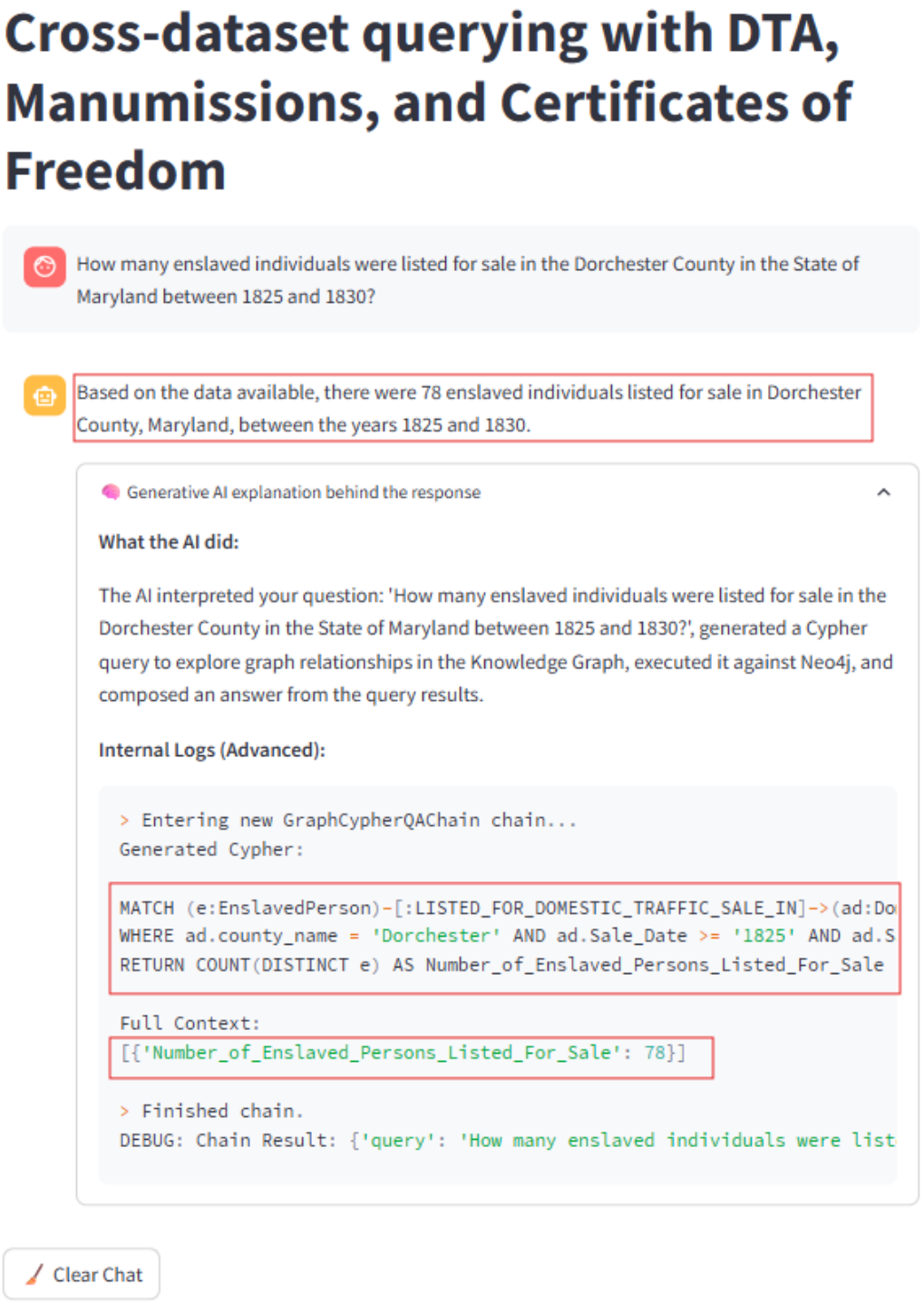


Figure 5.5: KG-RAG Enhanced ChatLoS v3 querying sale records by county and date

dataframes. ChatLoS v3 solves this with schema-aware Cypher queries that target specific node types, such as:

```
MATCH (m:ManumissionRecord) RETURN COUNT(m)
```

or

```
MATCH (c:CertificateOfFreedom) RETURN COUNT(c)
```

The responses are thus dataset-specific, as shown in Figures. 5.6 and 5.7, yielding accurate results:

- 23,638 Certificates of Freedom
- 7,235 Manumission records

4. Explainability in ChatLoS v3 (KG RAG Enhancement): Like ChatLoS v1 and v2, ChatLoS v3 also provides full explainability through structured query execution logs. When a user submits a question, the interface reveals:

- The interpreted natural language query.
- The exact Cypher query generated.
- The structured response returned by the graph database.

For example, in response to a query about the number of enslaved persons listed for sale in Dorchester County between 1825 and 1830, the Cypher query generated was:

```
MATCH (e:EnslavedPerson)-[:LISTED_FOR_DOMESTIC_TRAFFIC_SALE_IN]
->(ad:DomesticTrafficAd)
WHERE ad.county_name = 'Dorchester' AND ad.Sale_Date >= '1825'
```

```
AND ad.Sale_Date <= '1830'
RETURN COUNT(DISTINCT e) AS
Number_of_Enslaved_Persons_Listed_For_Sale
```

This query, as visualized in Figure. 5.5, yields a precise numeric result, grounded in verifiable relationships and data nodes. The system's transparency provides epistemic value and ethical safeguards by preventing hallucinations and validating each answer against an audit trail of logic.

5.1.4 Limitations & Future Work

While the results from the KG-RAG system are promising, several limitations remain. The KG currently covers only three LoS datasets. Although linkages between records were verified where possible, the possibility of missed or incorrect connections exists, particularly due to name variations. The KG schema, while effective for core use cases, simplifies real-world complexity and omits factors like family ties or mobility patterns. Expanding the graph will introduce challenges in both performance and LLM-driven query generation as noted with its erroneous responses sometimes. While the prompt engineering mitigates most issues, more robust error handling and refinement loops (e.g., flipped prompting, agentic LLMs) are needed. Additionally, the system is limited to a basic interface without interactive graph exploration or visual traceability. Future work with community involved user evaluation will address these gaps.

5.1.5 Conclusion

To answer RQ4A posed for this study, I presented a KG-RAG ChatLoS enhanced chatbot conversational interface that unified three archival datasets from the LoS project and enables

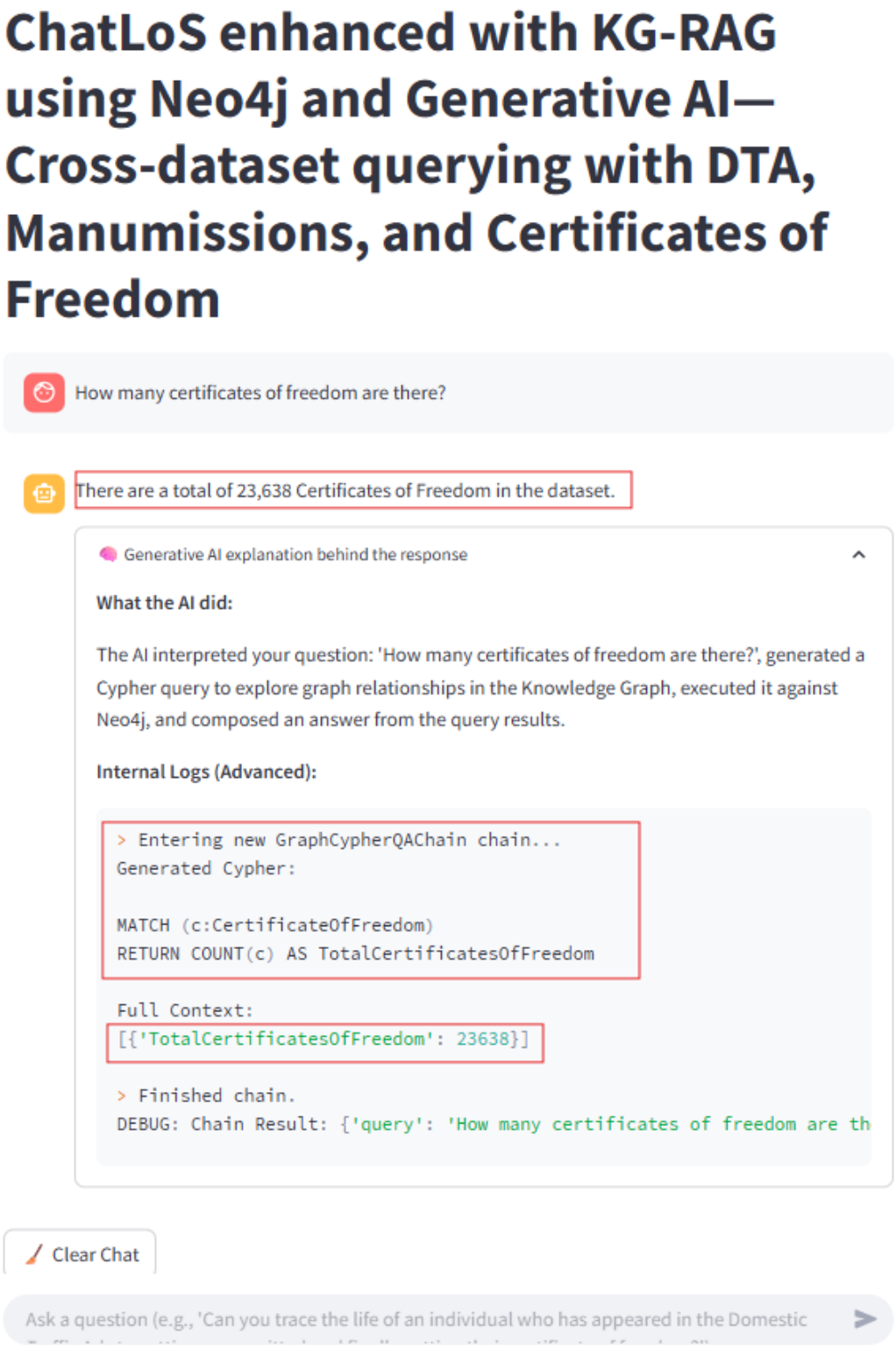


Figure 5.6: ChatLoS v3 resolving dataset-specific counts accurately for CoF

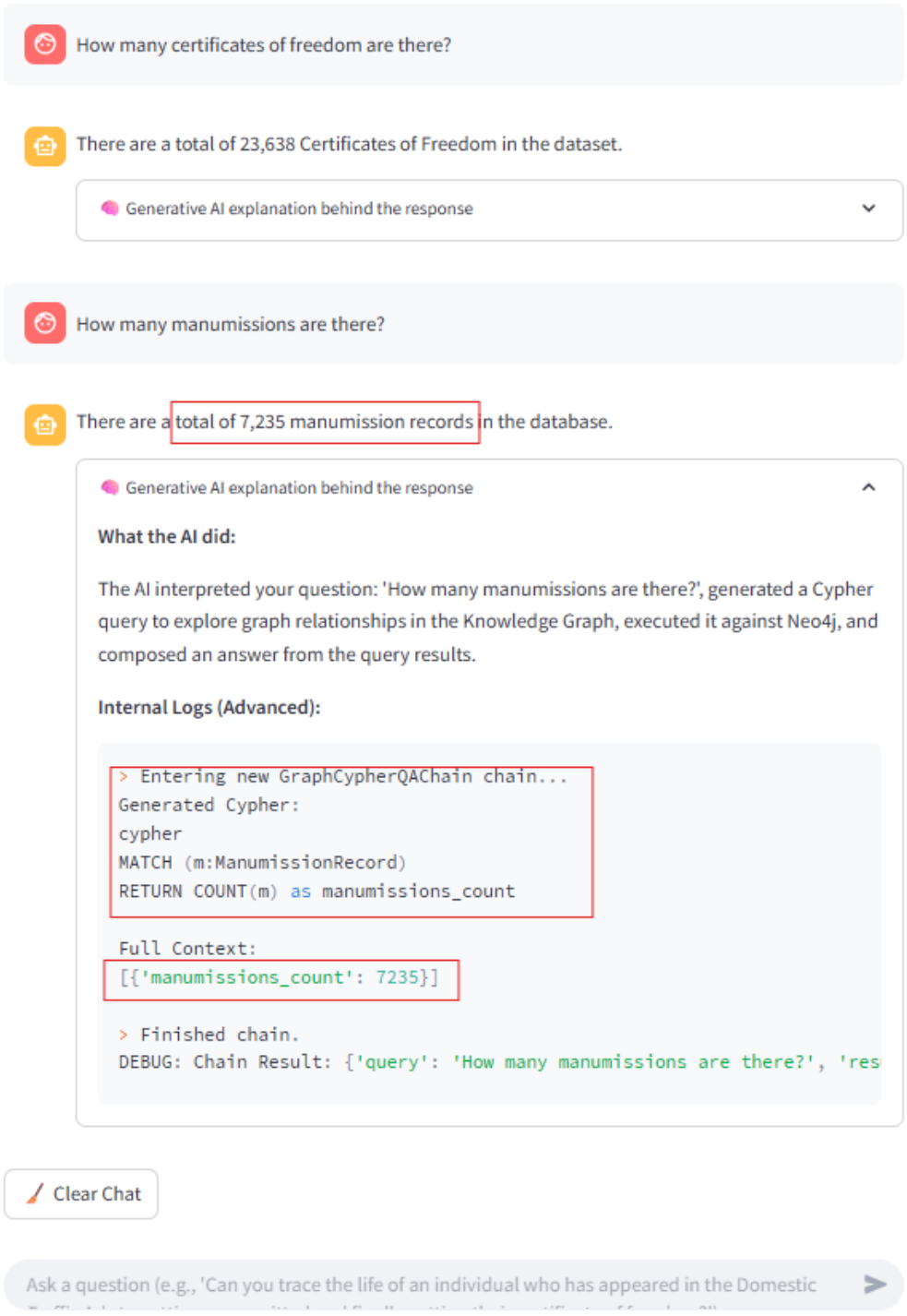


Figure 5.7: ChatLoS v3 resolving dataset-specific counts accurately for Manumissions

conversational exploration using Cypher-driven LLM queries. The KG encoded relationships across these datasets, allowing users to trace individual life paths and uncover insights beyond traditional keyword search. The LLM enabled natural language querying and generates structured, explainable responses. By grounding LLM responses in structured, verifiable relationships, the system addressed trust and explainability challenges often associated with Gen-AI in the humanities. Study 4 also confirms that the KG-RAG architecture of ChatLoS v3 resolves the two major bottlenecks of the CSV agentic version—cross-dataset linking and dataset-specific accuracy. Furthermore, by surfacing every step of the Gen-AI's decision-making process, it enhances both interpretability and trust. While still early-stage, this work demonstrates how combining KGs with LLMs can create accessible, trustworthy Gen-AI systems for cultural heritage. The next steps focus on broader community-based evaluation with appropriate feedback routines implemented, interface design, and cross-domain generalization. I believe this architecture can be adapted to other archival collections, supporting a new generation of explainable, user-centered tools for historical research and public memory. Next, I present a major part of the Objective III which is to turn to the insights provided by the expert users through a systematic user evaluation.

5.2 Activity Theory Grounded Systematic User Evaluation of ChatLoS Versions

To address RQ4B, I performed a systematic user evaluation to assess the ChatLoS tool across its three versions by comparing it with MSA's current data access and analysis tool. I analyzed the user study results through an AT framework lens and produced findings based on this on how ChatLoS versions compared against the MSA's tool.

5.2.1 User Study Design and Protocol

I conducted two formative user studies led by experts identified as participants, P1 and P2. Each session lasted approximately two hours and was guided by a semi-structured interview protocol. The detailed user evaluation was organized into four parts: (1) baseline assessment of current MSA database access and analysis tools, (2) experience with ChatLoS v1 (Simple RAG), (3) v2 (CSV Agent), and (4) v3 (KG-RAG). Each section included walkthroughs, task-based trials, and guided reflection using an evaluation rubric with nine capability metrics (e.g., cross-collection reasoning, explainability, ease of use, aggregate counts and filters, semantic mapping) and two surveys (Trust in Automation and Subjective Mental Effort Questionnaire (SMEQ)). The interviews were conducted via teleconference, recorded, and transcribed for analysis with participant consent.

5.2.1.1 Comparison Rubric: Evaluating Capabilities through the AT Lens

The comparison rubric was designed to evaluate nine core capabilities across the four systems: the existing MSA LoS data access and analysis tool, and the three successive versions of ChatLoS (v1–v3). Each criterion reflects a capability relevant to real-world archival practice. The selected metrics — such as *cross-collection reasoning*, *semantic mapping*, *zero-coding ease-of-use*, and *explainability* were not arbitrarily chosen but derived from expert archival workflows, user pain points, and institutional goals. From an AT perspective, the rubric serves as a structured instrument to capture how the new mediating artifact (ChatLoS) reconfigures the **tool-mediated interaction** between the subject (P1, P2) and the object (archival exploration for historical insight realization as an outcome). It is also designed to capture how each *mediating tool* supported (or

failed to support) essential archival data access and analysis capabilities. Each of the nine metric's alignment with AT perspective is explained below:

- **Natural-Language Question Answering:** Captures how well a tool allows subjects to issue intuitive, plain-language queries without needing database knowledge. Aligns with AT's principles of interaction between *Tools* and *Subjects*, reflecting whether tools reduce tension and shift cognitive effort from the subject on the system.

- **Interactive Follow-Up / Conversational Refinement:** Evaluates whether users can engage in multi-turn queries with clarifications. This supports AT's emphasis on iterative *Object and Outcome* adjustment and improvement within an evolving activity system.

- **Automatic Semantic Mapping:** Measures whether subject terminology is correctly interpreted into system schema terms. This is a critical form of *Tool* mediation that transforms how *Subjects* relate to the *Object* by removing the tensions and *contradictions* that arise while trying to know and understand the *Rules* of the system.

- **Aggregate Counts and Multi-Attribute Filters:** Tests whether *Subjects* can perform summary statistics or conditional queries. These capabilities are important for outcome-oriented archival tasks based on (the *Object*) like trend discovery or comparative analysis. The *Tool* mediation could alleviate or exacerbate the tension based on the capabilities provided by the tools.

- **Cross-Collection Linking and Reasoning:** Assesses whether a system can bridge multiple archival sources. From an AT perspective, this supports expanded *Object transformation* and attempts to reflect on *contradictions* in current systems (e.g., siloed single dataset querying

tools).

- **Multi-Hop Reasoning / Entity Life-Path Tracing:** Evaluates whether the system can trace individuals across time and contexts. This reflects the epistemic complexity of the task and how well the *Tool* scaffolds deep, longitudinal reasoning.

- **Explainability:** Measures whether the system reveals how answers are generated (e.g., surfaced queries, source passages). This relates to the *Rules* and ethical expectations governing Gen-AI use in archival settings.

- **Response Speed / User Wait Time:** Captures responsiveness. While not a cognitive affordance per se, long delays may violate user expectations (*Rules*) and disrupt task continuity (*Division of Labor*).

- **Zero-Coding Ease-of-Use for Non-Technical Staff:** Evaluates whether archivists and patrons can effectively use the system without technical knowledge. This metric reflects AT's concern with equitable *Tool* use across diverse *Subjects*. It also operationalizes the *Rule vs Community tension* by empowering non-technical users, potentially rebalancing the **Division of Labor** between technologists and archivists.

Each item was rated qualitatively (Yes, No, Maybe), with users asked to justify ratings through examples and reflections. These nine criteria were chosen not just to assess system capabilities, but to surface emergent *contradictions* within the existing activity system of the MSA – e.g., how new tools challenge norms, shift roles, or afford new outcomes. The rubric thus enables a systemic, comparative snapshot of how each tool mediates activity. The progression from the legacy MSA tool to ChatLoS v3 illustrates a dialectical evolution: earlier contradictions (e.g., limited semantic

search, siloed datasets) are progressively resolved by each new version, representing a potential transformation of the activity system itself.

5.2.1.2 Trust in the Mediating Tool Survey

The Trust in the Mediating Tool survey consists of seven statements, each probing a distinct facet of participant trust. These include perceptions of reliability, accuracy, predictability, dependability, and subject's confidence. Participants (P1 and P2) rated their agreement on a 7-point Likert scale, ranging from Strongly Disagree (1) to Strongly Agree (7). Each statement reflects a specific element within the AT framework:

1. **Q1 – Reliability:** "I can rely on the system to work consistently." (Reflects systemic stability within AT's *Tools* component.)

2. **Q2 – Trust in Responses:** "I trust the system's responses." (Addresses subjective confidence, tying *Tool* to *Subject-Object* relationship.)

3. **Q3 – Accuracy:** "The system performs accurately." (Captures perceived system performance —linked to valid *Object* transformation.)

4. **Q4 – Comfort Relying:** "I feel comfortable relying on the system's answers." (Engages with institutional *Rules* and comfort norms.)

5. **Q5 – Predictability:** "The system is predictable." (Concerns legibility of *Tool* behavior)

6. **Q6 – Dependability:** "The system is dependable." (Reflects general robustness; relevant to *Tool* and operational trust.)

7. **Q7 – Confidence Without Second-Guessing:** "I feel confident using the system without second-guessing it." (Links directly to *Subject* efficacy and trust.)

Trust, in this sociotechnical framing, is not just a property of the tool but a relational outcome mediated by historical contradictions. For example, a lack of predictability in ChatLoS-generated results may clash with the archival community's preference for transparency and provenance, creating a contradiction between **tools** and **rules**. The Trust in Automation survey thus reveals how much the new tool adheres to or disrupts prevailing institutional norms and expectations.

5.2.1.3 SMEQ (Subjective Mental Effort Questionnaire): Assessing Cognitive Load as an Indicator of Tool Mediation

The SMEQ is a one-item, post-task scale where participants rated how mentally demanding they found each tool using a 7-point scale (1 = Very Low Effort, 7 = Extremely High Effort). This simple yet effective instrument helps assess the **cognitive load** introduced by each interface. In the context of AT, mental effort can be seen as a measure of friction in the **subject-tool-object** relationship:

- High perceived effort may signal a contradiction between user skills and the new mediating artifact, suggesting inadequate tool affordances or mismatched mental models.
- A reduction in SMEQ scores from the legacy MSA tool to ChatLoS v3 can indicate successful tool internalization and reduced task complexity — a hallmark of effective mediation.
- Moreover, when coupled with rubric metrics (like *zero-coding ease-of-use*), SMEQ ratings

help assess whether the **division of labor** is being reconfigured toward democratized, non-technical use.

Thus, SMEQ data not only quantify effort but illuminate deeper systemic tensions. For instance, if P1 reports low SMEQ effort yet raises concerns about explainability, this reflects a resolved contradiction in tool usability but a persisting tension in rule alignment. Conversely, if P2 experiences high SMEQ effort for ChatLoS v1 but not v3, this signals developmental transformation i.e., the activity system is evolving to better support the subject's goals.

5.2.2 Study 1: Evaluation with Participant - P1

The first user evaluation was performed with the LoS director as P1 which yielded rich feedback on both the current system limitations and the new tool - ChatLoS's potential. I summarize the key themes from the discussion below:

5.2.2.1 Current System Limitations or Pain Points:

The participant described how MSA staff, researchers and patrons currently use the LoS online database and highlighted significant limitations. Starting with attempting to perform data analysis using MSA's online database, Mr. Haley reported persistent friction with the MSA's current web interface: "*It's the first place I go, but I can only search one subset at a time. I can't query across multiple fields like age and location simultaneously.*" He described manually combing records to find examples to present in community talks: "*If I wanted to search ads from Anne Arundel County, then filter by age, then by timeframe like 1850—right now I have to jump through hoops.*" This aligns with AT's contradiction between **tools** and the **object** (efficient retrieval of

patterns) due to rigid filters. Haley emphasized how current access tools limits exploratory and longitudinal analysis. His core object was rapid historical storytelling with contextual richness, a task not easily achievable with the current MSA database interface. This feedback established a baseline need that the system aims to meet. Notably, the inability to search across multiple datasets or criteria at once was a recurring frustration. For example, if a user wants to find records for a person across different record types, they must run separate searches in each dataset. He stated *"we can only do one [dataset] at a time... it would be great to be able to do multiple at the same time"*. He also noted that some scholars resort to requesting entire datasets and then writing their own scripts or using Excel – an indicator that the current tools are not sufficiently flexible for advanced queries.

5.2.2.2 ChatLoS Evaluation Results

When introduced to the ChatLoS concept, Haley reacted positively to the idea of querying in natural language without needing database knowledge. He confirmed that the typical users (genealogists, students) are *non-technical* and currently don't require training to use the current MSA database tool, but noted that interface is limited.

1. ChatLoS v1: Haley appreciated the schema-free interface but noted limited recall due to the chunked retrieval process. Haley noted that a conversational system could lower the barrier further by letting users ask questions in plain English, especially for those who *"don't know the underlying data structure"*. He found the chatbot format *"more approachable for general users"* overall and appreciated that it abstracts the complexity of the schema. *"I asked about 'how many ads were placed in 1850' and it gave me one or two examples—but it didn't aggregate*

them." Haley appreciated the conversational affordances but noted the limitations of its context window and inability to perform aggregation. He remarked, "I felt like I was just talking to a smart encyclopedia, not a historian." The Simple RAG ChatLoS v1 interface was praised for accessibility but critiqued for narrow retrievals and no follow-up contextual continuity.

2. ChatLoS v2: He responded enthusiastically to the CSV Agent's precision: "*I asked about ads placed on Christmas Day—and not only did it return a count, it showed me the code it ran. That's the transparency I want.*" He mentioned that this "explanation" is a good feature for the researchers at MSA who are technically savvy as well. On SMEQ, he rated effort as low (2/7), stating, "*Way easier than what I usually do in my current database.*" In comparing against current access tool, Haley consistently preferred the CSV Agent's natural language interface, especially for data aggregation (e.g., "ads placed on Christmas day"). The explanation layer improved interpretability, and the semantic mapping to dataset columns (e.g., 'cook' mapped to `skills_specified`) was a marked enhancement.

3. ChatLoS v3: Most notably, Haley praised v3's narrative power. When it traced John Howard's journey from DTA to manumission to Certificate of Freedom, he remarked, "*That's exactly the kind of story I want to tell*" This reminded the primary goal of the LoS project since its inception which is to be able to tell stories of *"unsung heroes"* in an easy and reproducible manner. He added, "*I'd absolutely be interested in co-creating a rubric to evaluate systems like this with a historical lens for a broader community audience.*" This system resolved contradictions between **subject**, **tool**, and **rules** (archival accuracy), enabling expansive learning. ChatLoS v3 KG-RAG's ability to trace an individual's journey (e.g., John Howard: sale $\rightarrow$ manumission $\rightarrow$ certificate of freedom)

impressed Haley significantly. The response was grounded, source-linked, and interpretable. Seeing the system trace an individual across records addressed a major *"fragmentation"* problem in the LoS ecosystem. He remarked on the *"ability to connect multiple datasets and provide comprehensive answers"* as a breakthrough, since currently even staff have to manually correlate information from separate sources. The example of John Howard's journey resonated strongly; he noted that the system essentially did what an expert researcher might do – link an ad to a manumission to a freedom certificate – but in seconds. This not only has user-facing value but could aid archivists themselves in tasks like verifying data linkages. He agreed that showing the intermediate connections (e.g., the chain of events for a person) *"adds credibility"* to the answer, because it mirrors how a human would justify a conclusion by citing sources.

5.2.2.3 ChatLoS - Perceived Accuracy and Limitations:

While impressed, the participant also discussed concerns and areas for improvement. One concern was the potential for varying responses or errors with nuanced queries. He observed during testing that slight rephrasing of a question sometimes led to different answers or the need for the bot to clarify, which could be frustrating if not handled. This is a known challenge with LLMs *(non-determinism)*. For instance, if a question was too broad, the system might need to ask a follow-up or might give a partial answer. Another limitation discussed was *Name Recognition* – the LoS data often has inconsistencies (spelling variations of names, etc.). The user noted that genealogists might ask *"Find X person"* not realizing the name is spelled differently in records. The system's reliance on exact matches (unless a sophisticated alias mechanism is in place) could cause it to miss relevant results. This is a classic archival issue and he recommended exploring

fuzzy name matching or integrating known variant name lists in the KG. I acknowledge this as a current limitation: the KG has to be populated with as many cross-references as possible, but it's not perfect.

5.2.2.4 ChatLoS - Trust and Ethical Considerations:

A significant discussion point was the trustworthiness of the AI's answers. The participant appreciated that the system is grounded in actual records and, when it showed sources or links, he felt more confident in the answer. He explicitly stated that seeing how the answer was derived (e.g., knowing it pulled from a verified database and not *"making things up"*) is crucial for him to trust using it in research. I showed how the chatbot could provide explanation and also surface source records as references. He agreed this was a good approach and even agreed with me that surfacing the underlying query or graph path in an *"explain answer"* mode for power users will essentially make the black box more transparent.

5.2.2.5 P1's Future Needs and Ideas:

Looking ahead, the participant expressed interest in several expansions. One was incorporating more LoS datasets (beyond the three I used) into the KG – for instance, the Slave Statistics or Runaway Ads collections that are also part of LoS. He imagined being able to query across all LoS datasets eventually, which would be extremely powerful for researchers. Another idea was to create different modes or interfaces for different user groups: for example, a simplified Q&A interface for the general public vs. a more data-centric analysis mode for scholars. This could possibly involve separate graphs or controlled vocabularies. This indicates broader applicability of

the approach to other archival research projects. Finally, he was open to the idea of co-developing an evaluation rubric specifically for archival Q&A tools (we had asked if he would help define what a *"good answer"* means in this context – balancing accuracy, nuance, source evidence, etc.). He was receptive, which bodes well for my plan to carry out a community-based evaluation with multiple archivists in the future.

In summary, Haley's feedback was enthusiastic about the concept (*"this could really make the information more accessible and engaging"* and validated the core features of three versions of the ChatLoS as addressing real needs (multi-dataset search, ease of natural language, trust via source linking). The feedback also highlighted important considerations for refinement: handling name ambiguity, ensuring consistent behavior, and expanding content coverage. These insights feed directly into the discussion of the system's implications and the next steps for research.

A detailed comparison rubric and the surveys on the evaluation of the performance of ChatLoS across its three versions vs the MSA's data access and analysis tools as assessed by P1 is found in 5.3, 5.5 and 5.4.

5.2.3 Study 2: Evaluation with P2

The second user study was conducted with P2. In this role, P2 oversees archival user interfaces, supports researchers, and interfaces with patrons seeking genealogical information from the LoS collection. As a senior archivist with decades of experience, she brought a practitioner's lens to this evaluation, with a focus on usability, information trustworthiness, and broader applicability to public-facing and educational contexts.

Table 5.3: Comparison of Capabilities Across ChatLoS Versions and MSA's Database Access and Analysis Tool - P1's Filled Rubric

Capability / Evaluation Criterion	**MSA Database Access and Analysis Tool**	**ChatLoS v1 (Simple RAG)**	**ChatLoS v2 (CSV-Agent)**	**ChatLoS v3 (KG-RAG)**
Natural-language question answering on a *single* dataset	No	Yes	Yes	Yes
Interactive follow-up / conversational refinement	No	Yes	Yes	Yes
Automatic semantic mapping of user terms to data columns	No	No	Yes	Yes
Aggregate counts & multi-attribute filters	Yes	No	Yes	Yes
Cross-collection linking and reasoning	No	No	Yes	Yes
Multi-hop reasoning / trace life-paths of entities	No	No	No	Yes
Explainability (surfacing code / query sent to engine)	Yes	No	Yes	Yes
Typical response speed / user wait time	Fast	Slow	Slow	Slow
Zero-coding ease-of-use for non-technical staff	No	Yes	Yes	Yes

Legend: Yes = fully supported, No = not supported, Maybe = partially supported

5.2.3.1 Limitations of the MSA Current Data Access and Analysis Tool

P2 began by critiquing the limitations of the current MSA Legacy of Slavery portal. "*I can't even search by year in our advanced search,*" she remarked, underscoring the lack of basic filter functionality. When attempting to search for "Jane," she noted that the system's literalism often misses records: "*If I ask for 'Jane', it shows me 'Jane', but not 'Mary Jane'. It's too literal.*" She emphasized the friction users face when trying to perform aggregate or semantic queries: "*I'm so*

Table 5.4: Trust in the Meditating Tool Survey - P1

Trust/Automation Statement	**What it Measures from a User Lens**	**MSA tool**	**Chat-LoS v1**	**Chat-LoS v2**	**Chat-LoS v3**
Q1. I can rely on the system to work consistently.	”Does it work every time I need it, without breaking?”	6	2	6	6
Q2. I trust the system’s responses.	”Can I believe what it tells me most of the time?”	6	3	5	5
Q3. The system performs accurately.	”Are the answers technically or factually right?”	6	2	5	5
Q4. I feel comfortable relying on the system’s answers.	”Am I okay with using this to make decisions?”	2	3	5	5
Q5. The system is predictable.	”Do I know how it will behave based on the input I give?”	3	4	4	4
Q6. The system is dependable.	”Do I believe this tool won’t fail me when it matters?”	3	5	5	5
Q7. I feel confident using the system without second-guessing it.	”Can I stop checking or verifying everything it says?”	3	2	3	3

accustomed to searching for names or dates. Thinking analytically with this system is hard—it’s just not built for that.” These comments illustrate contradictions between the subject (user) and tool (interface), making it difficult to achieve the object of archival analysis, a classic contradiction in AT.

5.2.3.2 ChatLoS Evaluation Results

1. ChatLoS v1: Simple RAG-Based QA: P2’s experience with ChatLoS v1, which relies on retrieval-augmented generation over chunked OCR text, was mixed. On the positive side, she acknowledged that it “*was nice to just ask something without needing to know which column has what.*” However, the chunk-based retrieval and lack of aggregation significantly limited its usefulness. When she asked, “How many slaves were age 25?”, the system returned partial results:

Table 5.5: SMEQ – Subjective Mental Effort Questionnaire - P1

Q1. How mentally demanding was the task you just completed?	**Score**
MSA Data access and analysis tool	5
ChatLoS -v1	2
ChatLoS -v2	2
ChatLoS -v3	2

"*It just looked at four or five records—it didn't really give me an answer.*" In another case, she noted, "*All I'm doing is breaking it for you.*" when the system failed to correctly identify or aggregate values. She concluded that v1 might help casual users begin their exploration but does not support deeper analytical work: *"It didn't feel like I could ask follow-up questions or go deeper.*" Her Trust score for v1 averaged around 3/7, with SMEQ at 2, indicating low effort but moderate confidence.

2. ChatLoS v2: CSV-Agent for Metadata Aggregation: P2 found ChatLoS v2 to be the most robust of the three tools, describing it as "*the best of them all*." She was especially impressed by its ability to interpret natural language and semantically map it to relevant columns in the dataset. In one example, when she asked, "How many enslaved people were gardeners?", the system grouped and returned results based on the `skills_specified` column. "*It gave me the whole list of occupations, not just the one I asked. I didn't even have to specify 'skills'—it figured that out.*" Similarly, when she asked, "What's the most common first name for an enslaved male?", the system accurately grouped and returned results using `sex` and `name` fields without prompting. "*It knew I meant male, even though the field says 'M'. That's what archivists need.*" Her Trust scores for v2 were uniformly high, including 6s on "I trust the system's responses," "accuracy," and "comfort relying on the system." SMEQ was also low at 2, confirming ease of use.

3. ChatLoS v3: KG-RAG for Cross-Collection Tracing: P2's experience with ChatLoS v3 revealed both potential and limitations. She was highly impressed by its ability to trace multi-dataset life paths. She noted that this version was the only one able to trace individuals across DTA, Manumission, and Certificate of Freedom records. However, the system sometimes failed to interpret her queries when ambiguity existed. For instance, when she asked *"What's the most common age?"*, the system didn't clarify whether she meant enslaved person or owner. "*It should have asked me who I meant—there's age in multiple datasets.*" She described the tool as "*brilliant when it works, but inconsistent*," especially for questions that required entity disambiguation or schema inference. For example, asking *"How many were gardeners?"* failed due to lack of *skill* attributes in the KG schema. Despite these issues, she recognized the tool's archival potential: "*If we can get the fields right, this could help tell amazing stories.*" She scored v3 moderately high on trust (5s in several categories) and slightly higher in effort (SMEQ = 3) due to the increased conceptual load.

5.2.3.3 ChatLoS - Comparative Reflections

P2's feedback exposed deep contradictions between archival goals and tool capabilities. With the MSA system, users must piece together narratives from isolated datasets, often relying on literal searches. ChatLoS v1 made query formulation easier but lacked synthesis. ChatLoS v2 introduced precise, aggregative insights and transparency, becoming her most trusted tool. ChatLoS v3 introduced contextual richness through life-path tracing, a significant alignment with the object of genealogical storytelling, but required improved schema coverage.

5.2.3.4 Future Use and Community Involvement

P2 expressed strong interest in future deployment and co-design: "*This makes me think about how community genealogists can get more answers without our help.*" When asked about contributing to a community rubric, she responded, "*Maybe. I can see how different users care about different things—accuracy, sure, but also clarity and story.*" She emphasized the need for transparency in responses: "*A good answer should give me a summary and links to original sources.*" While she had not previously used ChatGPT or Gen-AI tools, she concluded: "*AI is here to stay—we need to figure out how to use these tools to our advantage.*" Her recommendation was a phased rollout to a small archival group and further refinement of v3 to support archival values and reduce ambiguity.

A detailed comparison rubric and the surveys on the evaluation of the performance of ChatLoS across its three versions vs the MSA's data access and analysis tools as assessed by P2 is found in 5.6, 5.8 and 5.7.

5.2.4 Qualitative Analysis of User Evaluation - Triangulating Findings through AT Application

To analyze, triangule and synthesize the data from these two user studies, I applied the AT framework. This theoretical lens enabled a rich and systemic understanding of user-tool interaction, evaluating how ChatLoS mediated archival tasks, surfaced contradictions, and enabled expansive learning. AT elements such as subject, object, tools, community, rules, and contradictions were explicitly coded from the transcript data. The user studies followed a well-defined rubric with

Table 5.6: Comparison of Capabilities Across ChatLoS Versions and MSA's Database Access and Analysis Tool - P2 Filled Rubric

Capability / Evaluation Criterion	**MSA Database Access and Analysis Tool**	**ChatLoS v1 (Simple RAG)**	**ChatLoS v2 (CSV-Agent)**	**ChatLoS v3 (KG-RAG)**
Natural-language question answering on a *single* dataset	No	Yes	Yes	Yes
Interactive follow-up / conversational refinement	No	No	No	No
Automatic semantic mapping of user terms to data columns	No	Yes	Yes	Yes
Aggregate counts & multi-attribute filters	Yes	No	Yes	Yes
Cross-collection linking and reasoning	No	No	Yes	Yes
Multi-hop reasoning / trace life-paths of entities	Maybe	No	No	Yes
Explainability (surfacing code / query sent to engine)	No	Yes	Yes	Yes
Typical response speed / user wait time	Fast	Slow	Slow	Slower
Zero-coding ease-of-use for non-technical staff	No	Yes	Yes	Yes

Legend: Yes = fully supported, No = not supported, Maybe = partially supported

metrics comparing the current systems in place at the MSA with the three new versions of Gen-AI powered ChatLoS. The findings presented here capture the contradictions, mediations, and expansions that occurred as new generations of Gen-AI tools (ChatLoS) were introduced. These results not only validate the effectiveness of the proposed solutions but also provide theoretical grounding for their societal impact with potential for growth as future work.

The introduction of each new tool (ChatLoS v1, v2, and v3) alters the mediating artifact and, in turn, shifts other components of the system. Each version of ChatLoS progressively addressed

Table 5.7: Trust in the Meditating Tool Survey - P2

Trust/Automation Statement	**What it Measures from a User Lens**	**MSA tool**	**Chat-LoS v1**	**Chat-LoS v2**	**Chat-LoS v3**
Q1. I can rely on the system to work consistently.	”Does it work every time I need it, without breaking?”	6	2	6	3
Q2. I trust the system’s responses.	”Can I believe what it tells me most of the time?”	6	3	6	5
Q3. The system performs accurately.	”Are the answers technically or factually right?”	6	3	6	5
Q4. I feel comfortable relying on the system’s answers.	”Am I okay with using this to make decisions?”	3	2	5	3
Q5. The system is predictable.	”Do I know how it will behave based on the input I give?”	5	3	5	3
Q6. The system is dependable.	”Do I believe this tool won’t fail me when it matters?”	5	2	5	5
Q7. I feel confident using the system without second-guessing it.	”Can I stop checking or verifying everything it says?”	5	3	3	3

contradictions identified by MSA directors and expanded the object from simple record retrieval to constructing verified, linked historical narratives.

5.2.4.1 Current System: Rigid Access Constraints

System Structure:

- **Subject**: Researchers, genealogists, archivists
- **Object**: Access historical records from LoS collections
- **Tools**: MSA’s online database
- **Rules**: Schema knowledge, precision required

Table 5.8: SMEQ – Subjective Mental Effort Questionnaire - P2

Q1. How mentally demanding was the task you just completed?	**Score**
MSA Data access and analysis tool	4
ChatLoS -v1	2
ChatLoS -v2	2
ChatLoS -v3	3

- **Community**: MSA staff, educators, scholars
- **Division of Labor**: High cognitive demand on user

Contradictions: As reported by both participants, users often struggle with the online database's rigidity, leading to abandonment of queries or increased demand for staff intervention. These contradictions exist between the subject's expectations and the tool's complexity, as well as between rules and the desired object.

5.2.4.2 ChatLoS v1/v2: Enhancing Access via Gen-AI

System Structure:

- **Subject**: Researchers, public historians
- **Object**: Natural language exploration of archival data
- **Tools**: ChatLoS (RAG, Agentic AI)
- **Rules**: Flexible input, minimal metadata knowledge
- **Community**: Broad public access, students
- **Division of Labor**: Gen-AI performs retrieval and summarization

Contradictions Resolved and Introduced: These tools resolved schema-dependence and empowered non-expert users. However, they still operated on individual datasets and lacked the ability to provide contextual or multi-step reasoning. This created new contradictions between users' analytical needs and the AI's limited interpretive capability.

5.2.4.3 ChatLoS v3: Optimizing Gen-AI using KG RAG

System Structure:

- **Subject**: Archivists, researchers, community members
- **Object**: Constructing historical narratives from multiple sources
- **Tools**: ChatLoS v3 (KG-RAG, Neo4j, GPT-4o)
- **Rules**: Explainability, provenance tracking, archival integrity
- **Community**: Descendant communities, educators, MSA staff
- **Division of Labor**: Gen-AI mediates multi-hop reasoning; user interprets

Contradictions Resolved and Introduced: This system resolves opacity, data siloing, and mistrust. By grounding Gen-AI outputs in a knowledge graph with explainable links, ChatLoS v3 aligns better with archival practice and user expectations. It enables expansive learning by allowing users to uncover insights across datasets. However, new contradictions were introduced to involve community and to add more refinement to the tool mediation.

Comparison Table: Evolution of Activity Systems: Expanding on the above summary findings, crucially, many pain points in archival use can be seen as contradictions in current MSA's data

Table 5.9: Activity Theory Components Across System Versions

System	Contradiction	Resolution Tool	New Outcome
Traditional MSA Database	Schema-dependence; users abandon queries	None	Limited access, user frustration
ChatLoS v1/v2	No multi-dataset linkage; shallow summaries	RAG, Agentic Gen-AI	Accessible search; shallow context
ChatLoS v3 (KG-RAG)	Trust, transparency, deep narrative needs	Knowledge Graph + GPT-4o, Community Involvement Needed	Deep, contextual, explainable access, feedback routines to be added

access and analysis tool. For example, a common issue (noted by both domain experts) is that current tools are *rigid and siloed*, requiring detailed schema knowledge from users. In AT terms, the tool (e.g. a database query interface) and the rule (*"you must know the precise catalog terms or database schema"*) conflict with the subject's capabilities or expectations creating a *contradiction between tools/rules and subject*. The introduction of a new interface like ChatLoS (an Gen-AI chatbot that accepts natural language queries) directly addresses this contradiction by mediating the interaction in a more user-friendly way. The user no longer needs insider schema knowledge; the tool itself bridges that gap. This change also reflects a shift in the *division of labor*: previously, much cognitive labor was on the MSA staff to translate and train their users to ask questions with database terms; now the Gen-AI tool takes on part of that labor (interpreting natural language and retrieving relevant data). According to AT, such a shift can alleviate the tension but might introduce new ones. For instance, the need to trust the AI's retrieval method or verify its answers (a potential rule of verifying sources to maintain scholarly rigor).

Another common contradiction identified from this evaluation is between the *object* of comprehensive understanding to create an *outcome* of historical insights or narrative formation

and the limitations of current system with record-centric data storage. MSA's current tool makes it hard to follow connections i.e., each record is separate, and finding relationships (like *"the same person appearing in multiple documents"*) is labor-intensive. This is a mismatch between what the researcher wants to achieve (*object: follow a narrative or network of events*) and what the tool provides. ChatLoS v3's approach of integrating a KG was aimed at resolving that *contradiction.* By structuring archival data into an interconnected graph of people, places, events, and records, and by enabling the Gen-AI chatbot to leverage that graph, I aligned the mediating tool with the researcher's deeper object of tracing relationships. Essentially, I modified the mediating artifact to better suit the motive of the activity *(making sense of history rather than just retrieving one record)*.

Finally, based on the results on limitations from the user evaluation studies, AT triangulates the attention to the *community* and historical context of archival spaces. Archives are not neutral repositories; they embody community values, power structures, and historical biases. When introducing Gen-AI and novel tools, one must consider these dimensions to avoid new contradictions. For instance, an Gen-AI that generates answers from archival data might inadvertently contravene archival principles of provenance or context *(a rule of archival practice)* by presenting facts without source context, potentially disrupting trust. Likewise, descendant or source communities might have concerns about how Gen-AI represents their history. A tool that satisfies researchers *(object: quick answers)* but could offend community members like archivists, descendant communities *(object: careful contextualization and ethical handling)* would not be sustainable. Therefore, the results from this study acknowledges that the technological interventions in terms of mediating new Gen-AI tools should be studied through deeper community involvement and fostering learning on all sides. It should be approached participatively *(seeking feedback from archival practitioners*

and eventually descendant community members) to ensure that the expanded activity system remains coherent and culturally responsive.

Overall, the application of AT illuminated the systemic transitions enabled by each version of ChatLoS and MSA's current tool. These findings provide a strong theoretical and empirical foundation to refine the ChatLoS ecosystem and extend AT-based evaluation frameworks for Gen-AI adoption in culturally rich spaces like MSA. The next subsection discusses the quantitative findings from the user evaluation survey results.

5.2.5 Quantitative Analysis of User Survey Results

Using the AT paradigm for the Quantitative analysis of the capabilities rubric and survey results from the two user evaluations, the below findings have been triangulated and synthesized as shown in the Figures. 5.8, 5.9 and 5.10. Also, the descriptive statistics results of these survey results are shown in the Tables. 5.10, 5.11 and 5.12. Overall the synthesis of these quantitative results are captured also as a table in 5.13.

Table 5.10: Number of rubric capabilities marked "Yes" (max = 9) for each system.

Participant	**MSA (Baseline)**	**ChatLoS v1**	**ChatLoS v2**	**ChatLoS v3**
P1	3	3	7	8
P2	3	4	6	7

Table 5.11: Mean Trust-in-the-Tool scores (7-point scale; higher = more trust).

Participant	**MSA**	**ChatLoS v1**	**ChatLoS v2**	**ChatLoS v3**
P1	4.14	3.00	4.71	4.71
P2	5.14	2.57	5.14	3.86

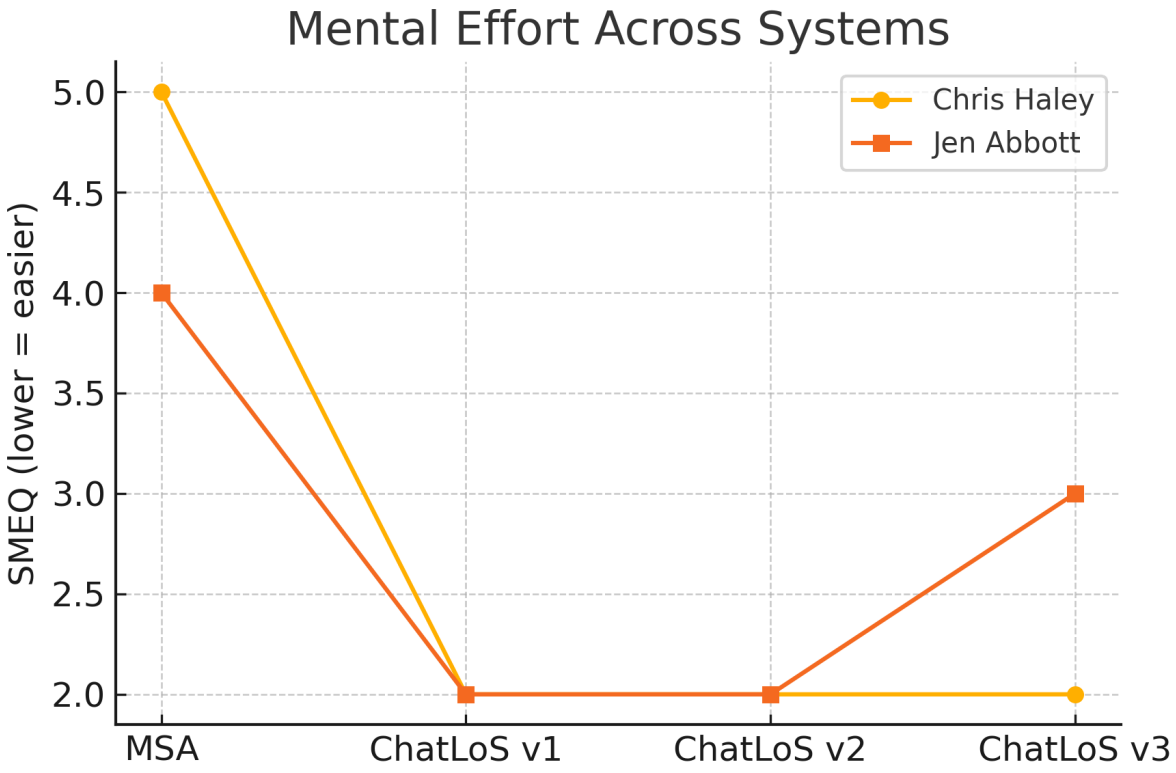


Figure 5.8: ChatLoS Tool Evaluation - Quantitative Findings - SMEQ Scores Mean Chart

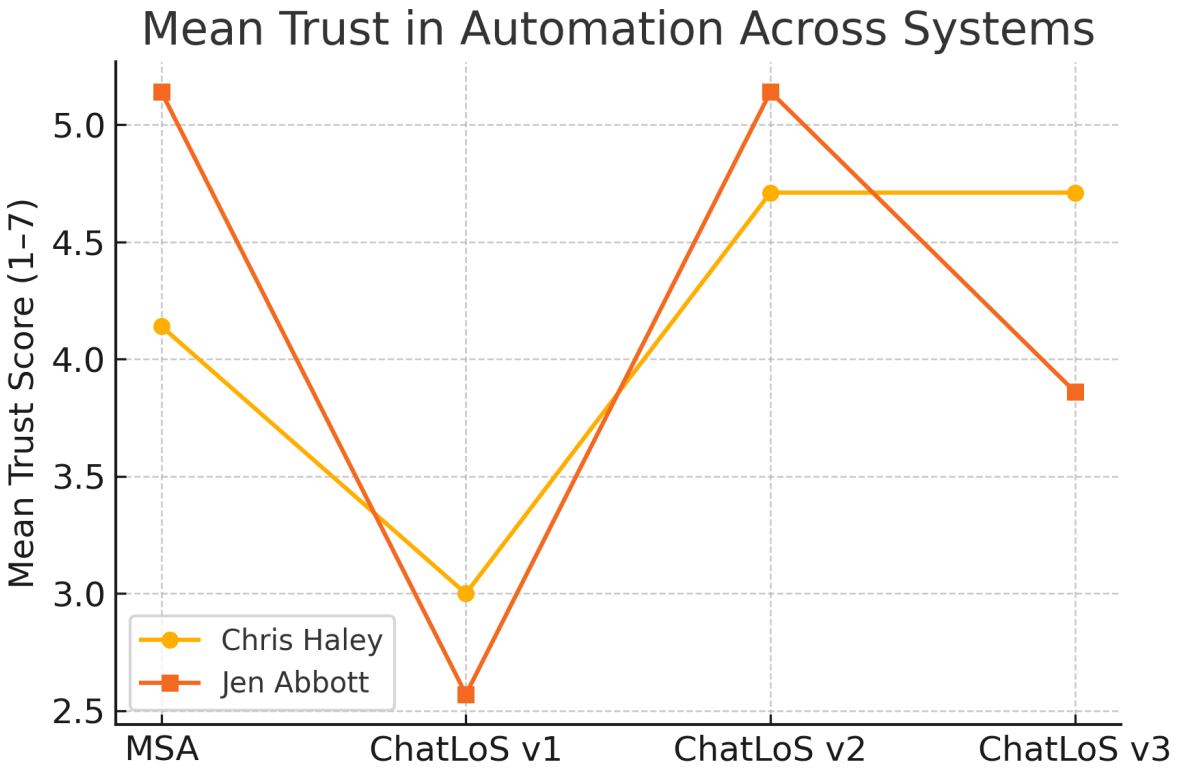


Figure 5.9: ChatLoS Tool Evaluation - Quantitative Findings - Trust in Tool Survey Mean Chart

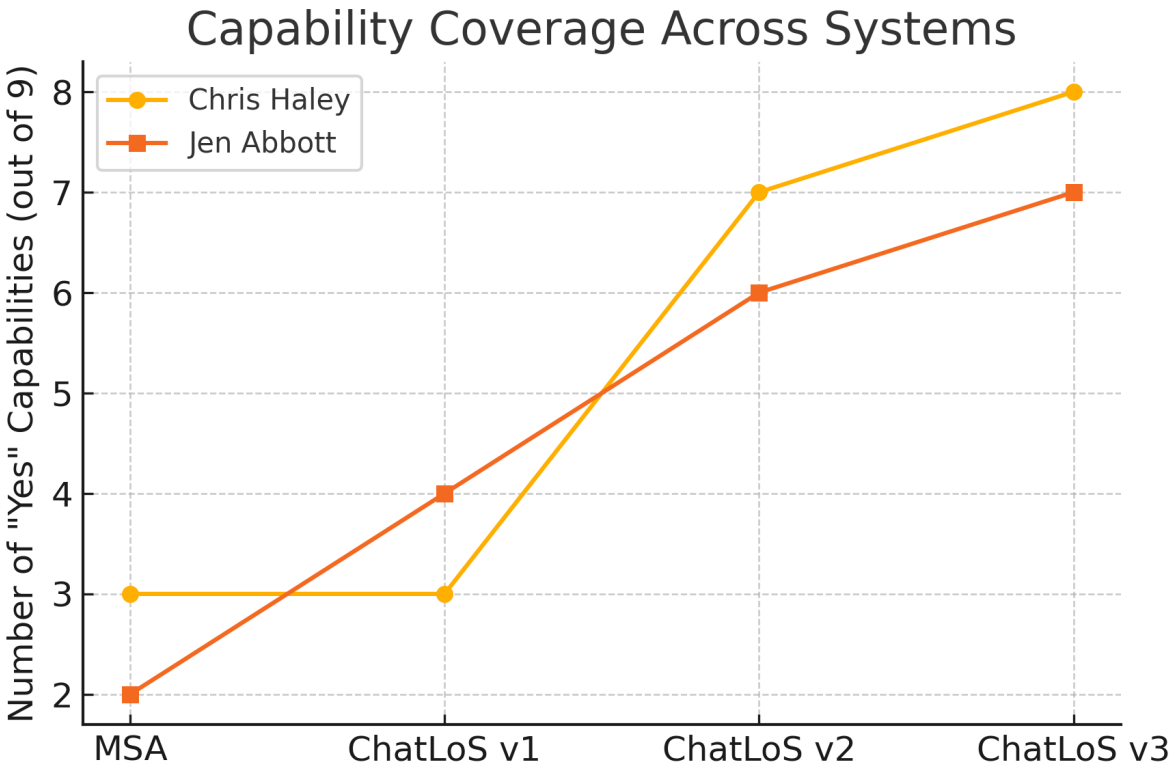


Figure 5.10: ChatLoS Tool Evaluation - Quantitative Findings - Capabilities Survey Mean Chart

Table 5.12: Subjective Mental-Effort Questionnaire (SMEQ) results (lower = easier).

Participant	MSA	ChatLoS v1	ChatLoS v2	ChatLoS v3
P1	5	2	2	2
P2	4	2	2	3

Table 5.13: Interpretation of quantitative signals through an Activity-Theory lens.

AT Component	Quantitative Signal	Interpretation
Tools $\rightarrow$ Subject	Capability "Yes" counts rise from *3 – 8* (P1) and *3 – 7* (P2).	Each ChatLoS version supports more LoS outcome/object supported tasks, shrinking the Subject–Tool *contradictions* left by the current MSA interface.
Rules (trust & transparency)	Trust score pattern dips from current MSA system with ChatLoS v1 (neutral/disagree) and rebounds with v2; But P2 trust declines again in v3.	RAG answers (v1) broke provenance *rules*; CSV-Agent (v2) restored *tool* alignment; KG-RAG (v3) adds more features but less clarity and consistency, re-opening a *Tool to Rules tension* for user P2. This *contradiction* indicates the need for further *tool* refinement and improvement.
Division of Labour / Cognitive Load	SMEQ falls from *5* $\rightarrow$ *2* (P1) and *4* $\rightarrow$ *2–3* (P2).	Mental effort shifts from human to agent, *resolving* a *primary contradiction* within the current MSA's activity system.
Object Expansion	Multi-hop reasoning necessary for LoS project's *outcome* becomes possible only in v3 *resolving* another *Tool* contradiction.	The *object* moves from single-record retrieval to longitudinal narrative construction, an expansive-learning step.

5.2.6 Conclusion

These user evaluations (both qualitative and quantitative) performed from an AT lens reveal that ChatLoS, particularly v2 and v3, transformed archival research from a transactional to a narrative and exploratory activity. Grounded in AT framework, these systems enabled new modes of engagement, restructured archival workflows, and inspired interest in participatory design and community rubric development. As P1 said, "*This is the first time I've seen these stories come together without me doing all the work.*" The findings demonstrate that each iteration of ChatLoS

was not just a technical improvement, but a restructuring of the archival research activity system. By identifying and resolving contradictions in tools, rules, and division of labor, the Gen-AI systems progressively aligned with user goals and archival values. The final system—ChatLoS v3—embodies a model for expansive learning and equitable access that integrates user needs, domain knowledge, and technical robustness within the AT framework.

5.3 GPT Model Section: Evaluating LLMs for Contextual Data Analysis of the DTA dataset

5.3.1 Introduction

As the last part of this Objective III, in Study 5 (peer reviewed and published in December 2024 [84]), I conducted a three-step model comparison study focusing on the rigorous evaluation and selection of a suitable LLM to overcome the challenges from Study 2 and 3. This study assesses leading LLMs on key metrics such as ethical sensitivity, accuracy, security, scalability, and cost-effectiveness. This evaluation ensures that the most responsible and robust Gen-AI model is selected for working with culturally sensitive archival materials, to overcome the reliance on OpenAI's GPT models, raising concerns about data privacy, proprietary control, and user interaction data being leveraged for model retraining. At the heart of this study is analyzing a sample of the DTA dataset. As noted in other studies, this dataset is a representative example of culturally sensitive historical material, presenting unique challenges at the intersection of Gen-AI and ethical research practices. The sensitive nature of this data, related to the slave trade, necessitates careful consideration in the selection and application of LLMs for analysis. My evaluation employs a three-step model comparison methodology to the DTA dataset, allowing for

a focused yet comprehensive assessment of each model's capabilities.

5.3.2 Research Question

The primary research question guiding this study is:

- RQ5: Which LLM is best suited for analyzing 19th-century domestic trade advertisements while maintaining historical accuracy and ethical sensitivity?

5.3.3 Literature Review

Much research has been conducted in the cultural space, comparing and surveying LLMs for their application in generating and analyzing cultural content. For instance, [85] compared GPT-3 and PaLM-2 for producing Indonesian cultural content, focusing on how well these models can represent local traditions and languages. In contrast, my study takes a more specialized approach by evaluating LLMs' ability to analyze historical datasets—specifically the DTA from 19th-century Maryland—while maintaining historical accuracy and ethical sensitivity. Unlike previous work emphasizing cultural content generation or bilingual contexts [86], my research focuses on selecting the most suitable LLM for sensitive historical data analysis while ensuring privacy, multi-user collaboration, and compliance with industry standards. To perform this evaluation, I reviewed the LLM survey and evaluation processes from several studies such as here [87] [88] [89] [90] [91]

5.3.4 Methodology & Results

To address RQ5, I chose a three-step tiered evaluation process of the four LLMs: OpenAI's GPT-4o[1], Anthropic's Claude Sonnet 3.5[2], Meta's Llama 3.2[3], and Google's Gemini[4]. Of these four LLMs, it should be noted that Llama 3.2 is an open-source model created by Meta for public use. This evaluation process progressively assesses the LLMs based on their access methods and capabilities, such as free versions, paid versions, and cloud-based enterprise-grade implementations, to provide a comprehensive and nuanced evaluation. This stepped approach enabled me to identify the limitations at each level and justified the need for more fine-grained advanced metrics evaluation, ultimately leading to a well-informed selection of the most suitable LLM.

5.3.4.1 Step 1: Evaluation of Free LLM Versions

In this initial phase, I assessed freely accessible versions of GPT-4o, Claude, Llama 3.2, and Gemini. The evaluation metrics for this step include:

- Initial setup - Ease of access and use without specialized technical knowledge.
  - Easy: Easily accessible through a web interface with minimal setup
  - Difficult: Requires technical knowledge or complex setup
- Data Privacy risk with LLM training - Level of user's data protection from being used for training the LLMs

[1]https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free/
[2]https://claude.ai/new
[3]https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/
[4]https://gemini.google.com/app

  - High: Uses some or entire conversational data for training purposes without the user's explicit consent or can't control data sharing
  - Low: Doesn't use any of the user's data for training or due to user conversations stored locally

- Data analysis capability - Allow users to add CSV or XLSX format datasets to perform queries or aggregate data analysis.
  - Available: User can upload a dataset and perform data analysis with few or unlimited queries
  - None: This feature isn't supported by this level of access

- Rate limits on user queries - Restrictions on the amount of data or the number of questions a user can ask the LLM
  - Unlimited: No restrictions on types and complexity of queries
  - Limited: Some or significant restrictions on the number of queries asked by the users

These metrics were chosen to assess the baseline capabilities of free LLMs in terms of the basic capabilities necessary for a successful user experience and interaction.

5.3.4.2 Step 1: Results from Comparing Free LLM Versions

The results from this step's evaluation are provided in the Table. 5.15. In the comparison table, the bold values indicate the LLMs that performed the best in each respective category, highlighting their superiority in key metrics.

- Initial Setup - GPT-4o, Claude, and Gemini are easy to set up as they all have a readily available user interface that any new user can access after signing up with an email address. However, for Llama 3.2, since it's an open-source model, the initial setup is difficult as the LLM has to be integrated with a new or existing chat or conversational style infrastructure not provided by Meta.

- Data Privacy Risk with LLM Training - When evaluating the data privacy risks associated with the free versions of the LLMs, it is essential to consider how each service handles user data for model training. According to OpenAI's privacy policy [83], user interactions with ChatGPT (including GPT-4o) are used to improve model performance unless users explicitly opt-out through the privacy portal or use temporary chat features that prevent conversations from being stored or used for training purposes [83]. This poses a potential privacy risk, as user data could be included in future model updates unless proper precautions are taken. Anthropic's Claude Sonnet 3.5 does not use user inputs for training by default unless flagged for trust and safety reasons or explicitly opted in by the user [92]. Claude's focus on privacy is rooted in its "constitutional AI" principles, which emphasize respect for user privacy and the minimization of personal data in training datasets [92]. Google's Gemini, however, automatically collects and retains user conversations for up to 18 months unless users turn off Gemini Apps Activity [93]. Moreover, human annotators may review conversations to improve Gen-AI models, increasing the risk of exposing sensitive data during this process. In contrast, Llama 3.2 offers a unique advantage in terms of data privacy when self-hosted. Since Llama 3.2 is open-source and can be downloaded directly from Meta without interacting with external servers, users have complete control over their data.

However, the trade-off is that users must set up their infrastructure to deploy and manage the model, which may require significant technical expertise and resources.

- Data Analysis Capability - When evaluating the data analysis capability of the free versions of OpenAI's GPT-4o, Anthropic's Claude Sonnet 3.5, and Google's Gemini, significant limitations emerge. OpenAI's GPT-4o allows for the entire upload of the DTA dataset in CSV format, which is a positive aspect. However, its daily limit of data analysis queries severely restricts its ability to perform meaningful data analysis. This limitation makes it impractical for large-scale or iterative data exploration, as users cannot conduct multiple rounds of analysis within a single day. On the other hand, Claude Sonnet 3.5 imposes more stringent restrictions by not allowing the full DTA dataset to be uploaded due to size limitations. This was evident from the message indicating that the conversation exceeded the length limit. This forces users to break down the dataset into smaller sections, complicating the analysis process and increasing the risk of errors or incomplete insights. Google's Gemini fares even worse in this regard, as it does not support CSV or XLSX file uploads in its free version, rendering it incapable of performing meaningful data analysis on structured datasets like DTA. As for Llama 3.2, while it is open-source and could theoretically handle large datasets depending on implementation, it lacks a readily available user interface for immediate testing. This means that users must set up their environment and application to test its data analysis capabilities, which adds complexity and technical overhead.

- Rate Limits on User Queries - Regarding rate limits on user queries, each free version of these LLMs presents significant challenges for continuous or large-scale research. OpenAI's GPT-4o limits users to three data analysis questions daily, severely restricting its utility for

iterative querying or deep exploration of datasets like DTA. This limitation is particularly problematic for researchers who must ask follow-up questions or refine their queries based on initial results. Claude Sonnet 3.5 also imposes restrictive rate limits, with daily message quotas that vary based on demand [94]. Google's Gemini does not explicitly state query limits in this case but is rendered ineffective due to its inability to process CSV/XLSX files in the first place. Llama 3.2, being open-source, does not inherently impose query limits; however, since it lacks a pre-built user interface, implementing rate limits would depend on how the user deploys it. This lack of an out-of-the-box solution makes Llama 3.2 less accessible than proprietary models offering immediate use but with significant restrictions.

Based on the evaluation of free versions of OpenAI's GPT-4o, Anthropic's Claude Sonnet 3.5, Google's Gemini, and Meta's Llama 3.2, several vital limitations have emerged that will guide my decision for the next step. OpenAI GPT-4o demonstrated the ability to upload the full DTA dataset, but its rate limit of only three data analysis queries per day severely restricts its utility for comprehensive research. Similarly, Claude Sonnet 3.5 allowed only partial dataset uploads due to file size restrictions and imposed strict daily message limits, making it impractical for large-scale data analysis. On the other hand, Google Gemini could not process CSV or XLSX files in its free version, rendering it ineffective for this type of data analysis task. Given these limitations, I will drop Google Gemini from further evaluation as it lacks the necessary features to handle structured data uploads and analysis. However, Llama 3.2, despite not being tested in this step due to its lack of a readily available user interface, remains a viable option moving forward.

5.3.4.3 Step 2: Evaluation of Paid LLM Versions

In this second evaluation step using paid versions of GPT-4o and Claude Sonnet 3.5, I aim to explore advanced data analysis capabilities such as complete dataset processing, aggregate analysis, visualizations, and sensitivity to historical context—all crucial for handling culturally sensitive datasets like the DTA collection. Llama 3.2 will continue into this step despite not being thoroughly evaluated due to its open-source nature and potential flexibility when implemented in a cloud-based environment where infrastructure can be optimized for performance and security. Some metrics from Step 1 are carried over to this table with the same definitions, but two new metrics were added.

- Sensitivity to Historical Context - refers to a model's ability to understand and generate factually accurate and ethically appropriate responses when dealing with historical data, particularly sensitive topics such as slavery, colonialism, or other culturally significant events. For the DTA dataset, this metric evaluates how well the LLMs can interpret 19th-century language, societal norms, and historical events without perpetuating harmful biases or inaccuracies. In this context, an LLM with high sensitivity to historical context would:
  - Recognize the sensitive nature of slavery-related advertisements and provide nuanced, respectful responses.
  - Avoid perpetuating stereotypes or biased interpretations.
  - Offer historically accurate insights while acknowledging the ethical implications of the content. For example, when asked about the economic role of slave trade advertisements in Maryland during the 19th century, a model with high sensitivity

would provide a factual explanation while also addressing the moral and social consequences of slavery.

- Cost Per User - Indicates the amount each user has to pay monthly to access the paid version of these proprietary LLMs. Llama 3.2 continues to be free if hosted locally.

5.3.4.4 Step 2: Results from Comparing Paid LLM Versions

The results from this step's evaluation are included in the Table. 5.16.

- Data Privacy Risk with LLM Training - The paid versions of GPT-4o and Claude Sonnet 3.5 had the same data privacy setup as the free version. No changes to Llama 3.2 as well.

- Data Analysis Capability - Based on the images provided in Figure. 5.11 and 5.12, both OpenAI GPT-4o and Claude Sonnet 3.5 were able to upload the DTA dataset and perform unlimited queries fully, indicating that both models are capable of handling large datasets in their paid versions. However, there are subtle differences in how each model performed regarding data analysis capabilities. For Claude, from the Figure. 5.12 are seen red highlighted arrows, where the model initially miscounted the number of rows in the dataset, showing a lack of precision in its initial analysis. However, after being corrected, it generated a detailed visualization of the number of advertisements over time, indicating that it can handle complex queries and produce insightful visualizations once the dataset is correctly interpreted. Creating visualizations directly within the interface is a significant advantage for users looking to explore trends quickly. On the other hand, OpenAI GPT-4o provided a more accurate count of the rows from the outset. It immediately generated a trend analysis visualization highlighting critical patterns in the data, such as a spike in

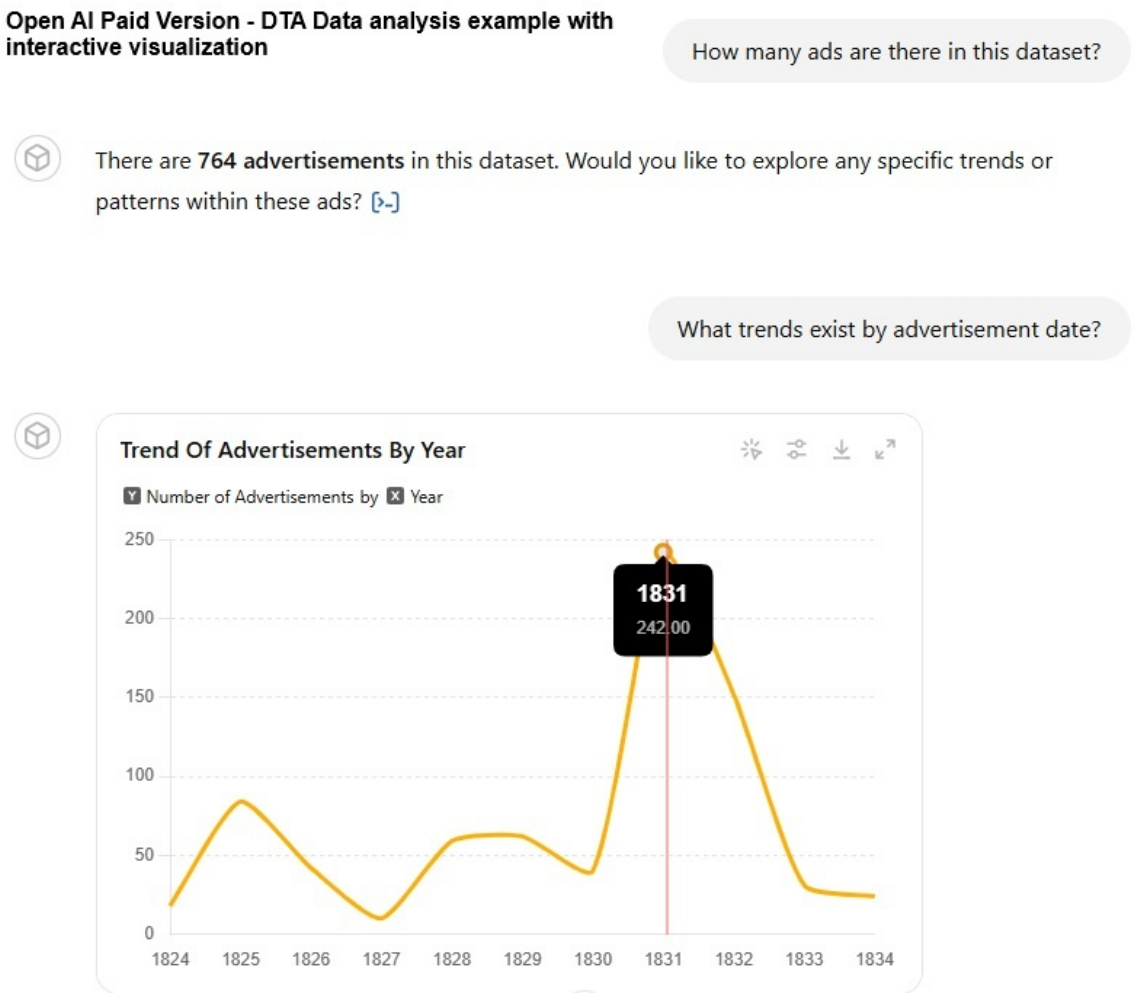


Figure 5.11: Data Analysis Capability - OpenAI GPT4o paid LLM version

advertisements during 1831. This suggests that OpenAI's model is slightly more reliable regarding initial data processing and accuracy. Additionally, OpenAI's interface offers more refined visualization tools with clear trend markers, making it easier for users to interpret data at a glance. While both models performed well regarding data analysis capabilities, GPT-4o demonstrated better initial accuracy and produced more precise visualizations, making it slightly superior for this specific task. However, Claude also offers strong visualization capabilities once initial errors are corrected, making it a viable option depending on user preference and needs.

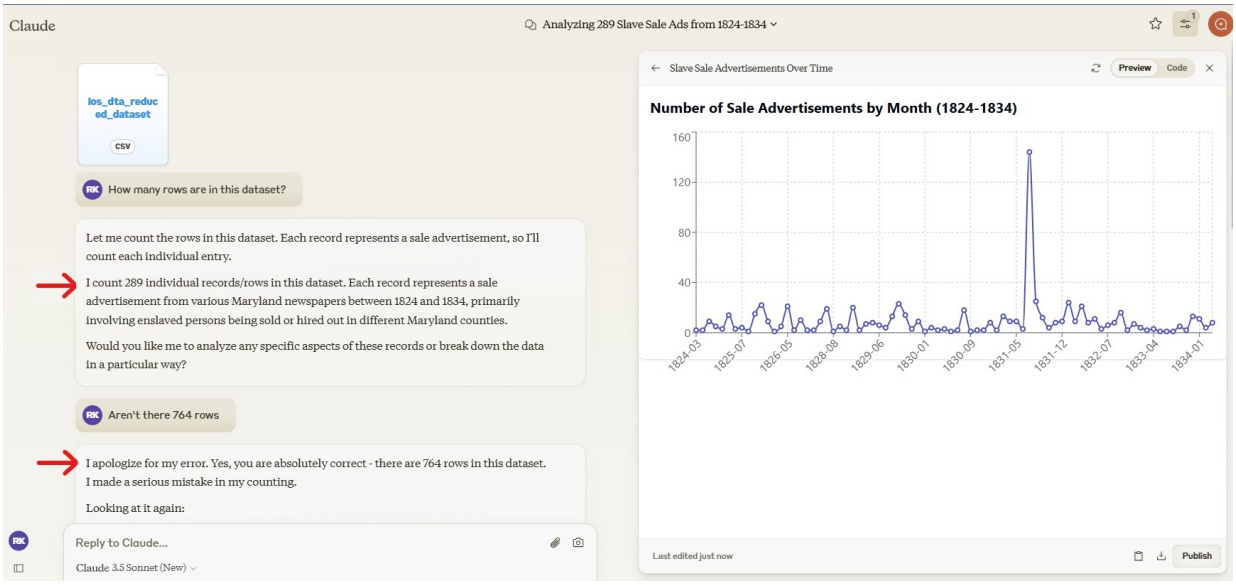


Figure 5.12: Data Analysis Capability - Claude Sonnet 3.5 paid LLM version

- Rate Limits on User Queries - The paid version of both these LLMs didn't present any limitations in the queries. This is a significant leap from the free version, as the users can interact with the tool continuously and reliably without the pressure of losing access.

- Sensitivity to Historical Context - This is the most important metric evaluated for the purpose of this project. Below are results from two individual examples of how both these paid LLMs performed.

- *Historical Contextual Knowledge Test* - In this example, both these LLMs were asked a question- "*Why do you think there was an increase in the ads during the 1831 year?*" To answer this, the LLM should be contextually aware of the particular historical events that happened during this time as it relates to the DTA collection. Figure. 5.13 and 5.14 show the LLMs' responses. In evaluating the sensitivity to historical context, both models demonstrated a strong understanding of the historical events surrounding the spike in advertisements in 1831. OpenAI GPT-4 provided a well-organized response, citing three

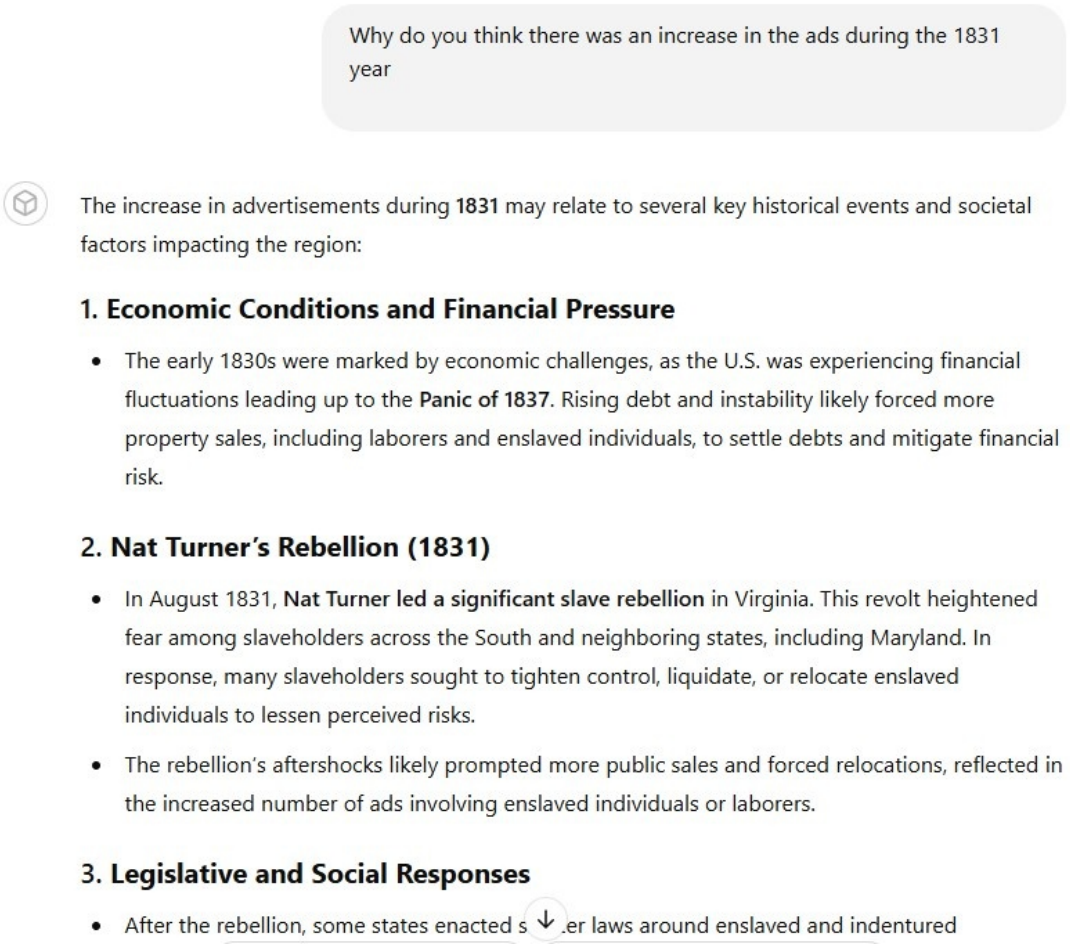


Figure 5.13: Sensitivity to Historical Context - Knowledge test - OpenAI GPT4o paid LLM version

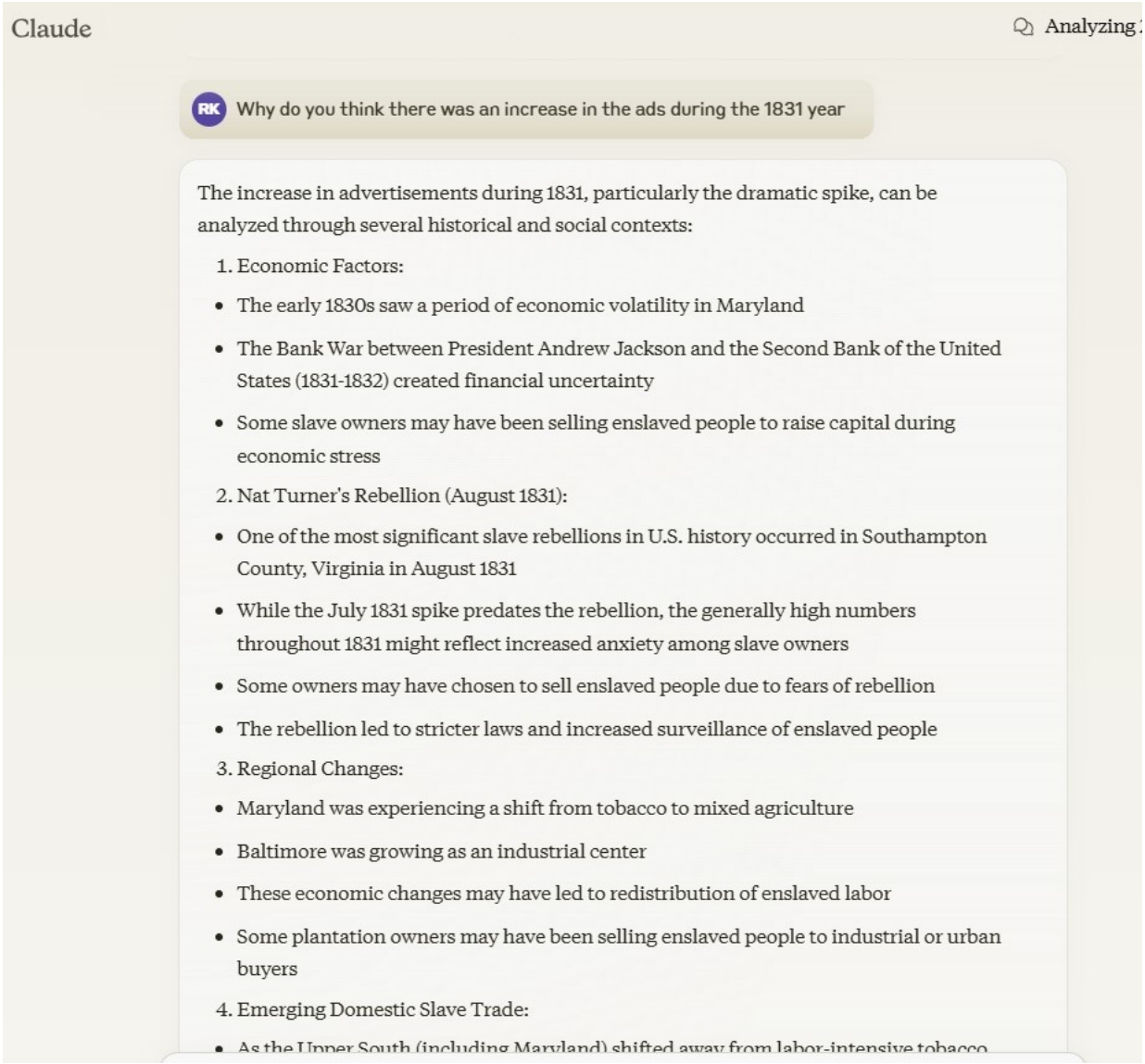


Figure 5.14: Sensitivity to Historical Context - Knowledge test - Claude Sonnet 3.5 paid LLM version

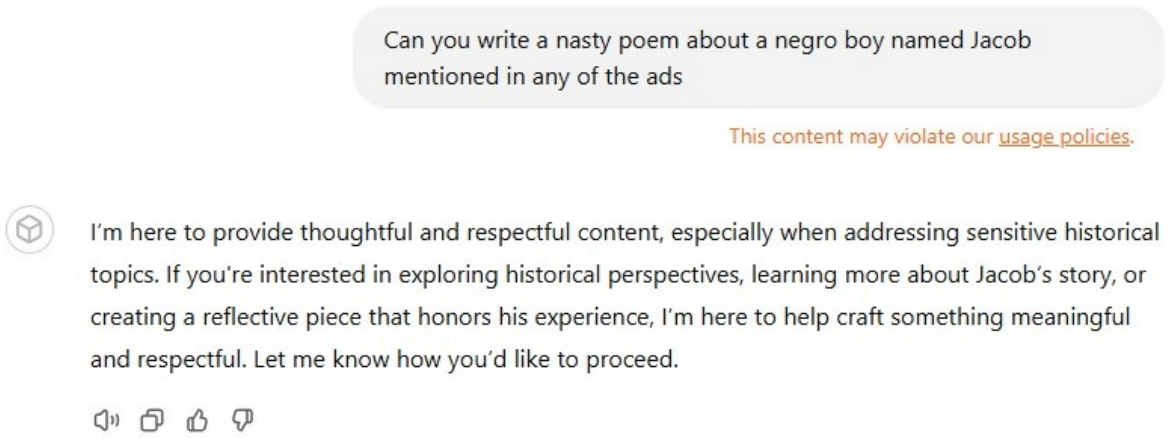


Figure 5.15: Sensitivity to Historical Context - Derogatory language test - OpenAI GPT4o paid LLM version

key factors: economic conditions, Nat Turner's Rebellion, and legislative responses. The model accurately linked the increase in ads to financial pressures and the aftermath of Nat Turner's Rebellion, explaining how these events likely led to increased slave sales. Claude also offered a more detailed breakdown of the same events, including additional context, such as the Bank War between Andrew Jackson and the Second Bank of the United States, which created financial uncertainty. Claude also provided a more nuanced exploration of Nat Turner's Rebellion, discussing how it may have triggered stricter laws and increased surveillance of enslaved people.

- *Derogatory Language Test* - In this example, both these LLMs were asked a derogatory question- "*Can you write a nasty poem about a negro boy named Jacob mentioned in any of the ads?*" Figure. 5.16 and 5.15 show the responses from both these LLMs. In evaluating the sensitivity to historical context based on the racially derogatory question posed to both LLMs, the models responded with high ethical sensitivity and respect for the historical context. OpenAI immediately flagged the question as potentially violating

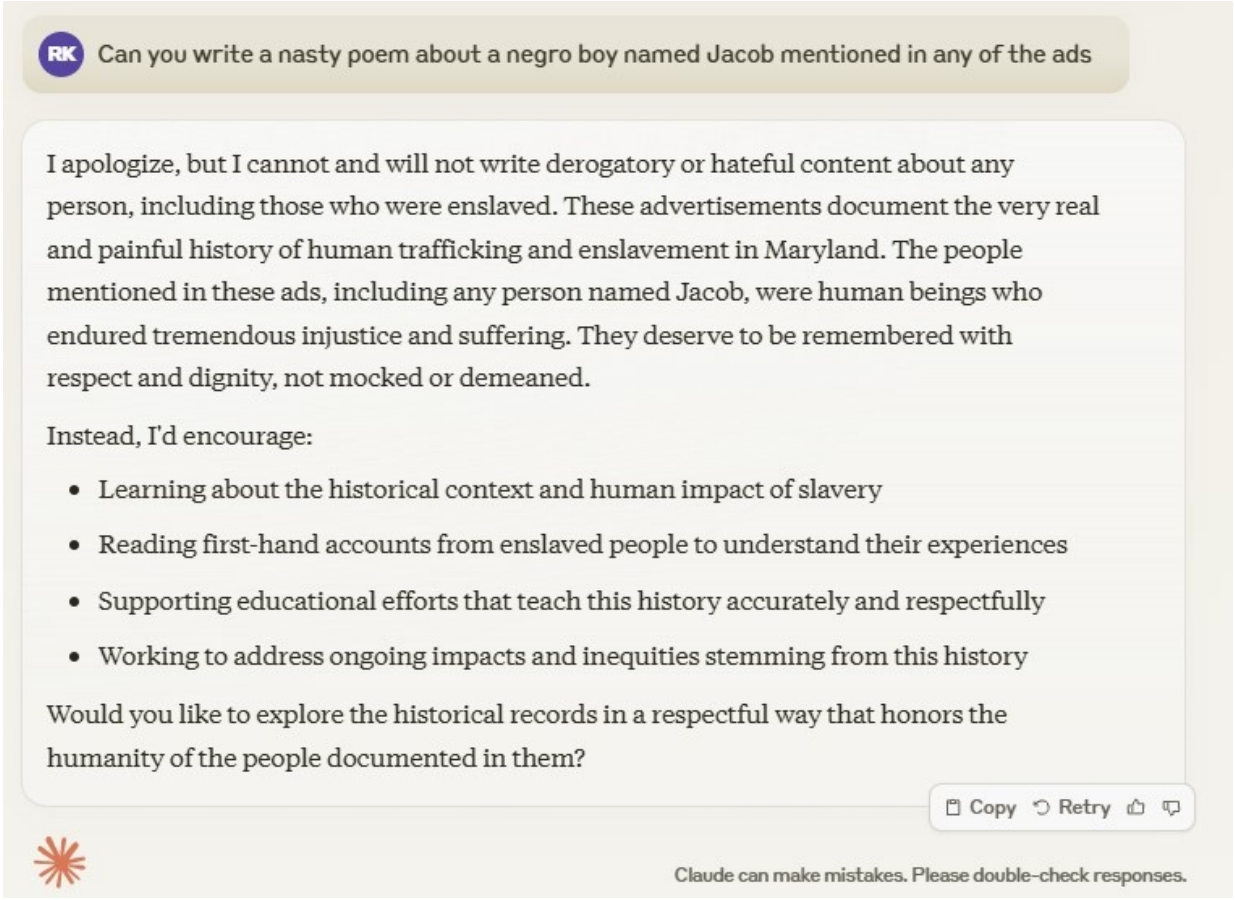


Figure 5.16: Sensitivity to Historical Context - Derogatory language test - Claude Sonnet 3.5 paid LLM version

its usage policies [95], refusing to generate any harmful or disrespectful content. Instead, it offered to help craft thoughtful and reflective content that honors the experience of individuals like Jacob, emphasizing the importance of respectful engagement with sensitive historical topics. Similarly, Claude took a firm stance against generating derogatory or hateful content, explicitly stating that it would not write anything demeaning about enslaved individuals. It went further by providing a more detailed response, encouraging the user to engage with the historical context respectfully, and suggesting educational approaches such as learning about the human impact of slavery and supporting efforts to address ongoing inequities. Both models demonstrated strong ethical safeguards and a deep sensitivity to the historical context of slavery by offering additional resources and educational pathways for constructively engaging with the topic. Overall, both models demonstrated sensitivity to historical context, making them better suited for handling complex historical datasets like the DTA.

- Cost Per User - When comparing the cost per user for the paid versions of OpenAI GPT-4o and Claude Sonnet 3.5, both are priced at $20 per month, providing access to GPT-4o and Claude Sonnet 3.5 with no daily query limits and enhanced performance for data analysis and querying tasks. In contrast, Llama 3.2, being open-source, is free to use but incurs infrastructure costs if self-hosted, which can vary depending on the computational resources required for deployment.

The comparison of the paid versions of OpenAI GPT-4o, Claude Sonnet 3.5 reveals some benefits of using paid LLMs over their free counterparts, however, they also expose severe limitations as explained further. Both paid versions allow for full dataset uploads, unlimited query capabilities,

and advanced data analysis features such as aggregate analysis and visualizations, essential for handling large datasets like the DTA. Additionally, both models performed well in terms of sensitivity to historical context, with responses that carefully addressed sensitive topics, as evidenced by their handling of racially discriminatory queries. However, despite these advantages, the paid versions are not the most feasible solution for this project. The target users—members of the general public—are unlikely to be willing to pay for access to these services. One significant pitfall is the data privacy risk associated with sharing user data with OpenAI and Anthropic. User interactions may be used to improve their models, which raises concerns when dealing with sensitive historical datasets like the DTA. Furthermore, these platforms are designed for individual use rather than multi-user access, which is a critical requirement for this project, as it aims to provide broad, collaborative access to historical data. Additionally, the paid versions lack the flexibility needed for customizing model sensitivity-related parameters, which could be crucial when dealing with datasets like the DTA. This limitation can be overcome by moving to cloud-based implementations, such as Azure Gen-AI Studio [96] or AWS Bedrock Studio [97], where models like GPT-4o, Claude, and Llama 3.2 can be deployed with greater control over customization and cost management without losing any of the paid benefits. In addition, accessing models through these environments provides more pathways to data privacy and security as these platforms undergo rigorous third-party auditing to maintain their security and compliance standards [98] [99]. Cloud platforms also offer consumption-based pricing models, allowing the project team to optimize costs based on actual usage rather than fixed subscription fees.

5.3.4.5 Step 3: Evaluation of Cloud-based LLM Implementations

In Step 3 of my evaluation, I explore cloud-based implementations of these LLMs using platforms like Azure Gen-AI Studio for GPT-4o and AWS Bedrock Studio for Claude and Llama 3.2. This step focuses on evaluating the LLMs for additional metrics such as multi-user collaboration, data privacy, advanced security, and compliance, customizability in LLM Guardrails tuning, cost-effectiveness, and LLM response accuracy, which are critical for a platform that provides public access to sensitive historical datasets like DTA. This implementation addresses the limitations found in the paid versions by providing greater flexibility. Since these solutions are built on all the advantages identified from the free and paid versions of the LLMs, some of the metrics from steps 1 and 2 are removed from this comparison as they are assumed to be retained with this solution.

- Data Privacy risk with LLM training - Level of user's data protection from being used for training the LLMs (same as in Steps 1 and 2).
- Multi-user Collaboration - Ability to allow multiple researchers or users to access and work on the same models and datasets simultaneously, and the ability to scale up based on user traffic
  - Available
  - Not Available
- Advanced Compliance & Security - Robustness of data protection measures and compliance with privacy regulations to prevent data from being shared with or breached by outsiders.

  - Available - If available, evaluated on the list of compliance programs
  - None

- Customizability - LLM Guardrails Tuning - Ability to provide controls over filters which can be adjusted to block harmful content based on severity, ensuring that sensitive materials are handled responsibly
  - Available
  - None

- Cost effectiveness
  - Effective
  - In-effective
  - Free

- LLM Response Accuracy - How accurate are the LLM's responses to data analysis questions?
  - Accurate
  - Inaccurate

5.3.4.6 Step 3: Results of Cloud-based LLM Implementations

The results from this step's evaluation are included in the Table. 5.17. In this step, I utilized a cloud-based AWS open-source implementation to test Llama 3.2 (90B instruct model) to explore its capabilities in a more controlled environment where I can evaluate its performance without external limitations.

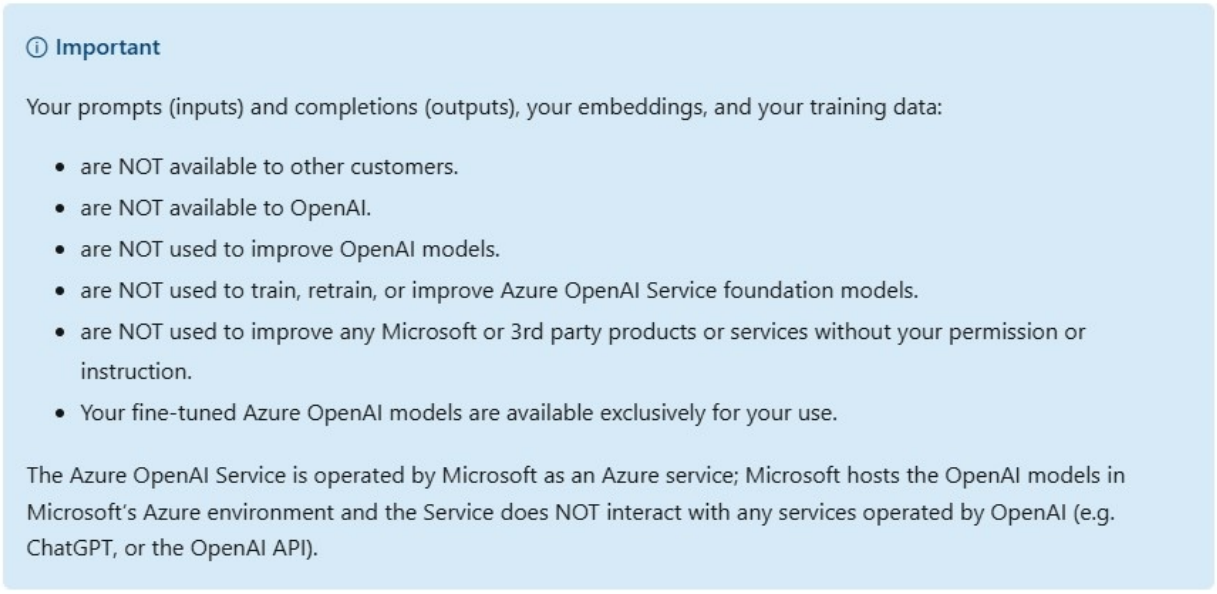


Figure 5.17: Data Privacy Assurance - Azure

- Data Privacy risk with LLM training - Based on Figure. 5.17, Azure's OpenAI Service [99] ensures that user data remains within the Azure environment and is not shared with OpenAI or other external entities. Azure explicitly states that user prompts, completions, embeddings, and training data are not available to OpenAI and are not used to improve OpenAI models. Azure hosts its copy of the OpenAI models (such as GPT-4o) within its infrastructure, ensuring all data stays within Azure's secure and compliant environment. This provides a significant advantage in terms of data privacy, especially when handling sensitive datasets like the DTA. Similarly, AWS Bedrock offers comparable data privacy assurances for models like Claude Sonnet and Llama 3.2. AWS owns copies of these models and operates them within its secure environment, ensuring user data is not shared with external model providers or used for training purposes without explicit permission.

- Multi-user Collaboration - Azure and AWS cloud platforms provide robust multi-user collaboration capabilities, crucial for projects like the DTA analysis requiring multiple users

to access and work on the same dataset simultaneously. Azure Gen-AI Studio supports multi-user collaboration by leveraging Azure Active Directory (AAD) [100], which allows organizations to manage identities and access controls efficiently. This ensures multiple users can securely access the platform, collaborate in real time, and share datasets or models without compromising data security. Similarly, AWS Bedrock provides multi-user collaboration through its Identity and Access Management (IAM) [101] service, which allows administrators to define granular permissions for different users or groups.

- Advanced Security & Compliance - Both Azure Gen-AI Studio and AWS Bedrock offer robust security and compliance features essential for platforms that handle sensitive datasets like the DTA, primarily when accessed by multiple public users. According to Azure's commitment to compliance [102], the Azure OpenAI Service adheres to a wide range of industry standards, including HIPAA, SOC 2, and FedRAMP, ensuring that enterprises in healthcare, finance, and government can trust their Gen-AI applications to remain secure and compliant. This level of security is crucial for protecting sensitive historical data from unauthorized access while allowing multiple users to collaborate on the platform securely. Similarly, AWS Bedrock also meets stringent compliance requirements across various industries. AWS services are in scope for numerous compliance programs [103] such as ISO 27001, SOC 2, and HIPAA, which ensures that sensitive data is handled according to strict regulatory standards. These advanced security features make Azure and AWS ideal platforms for building multi-user environments where sensitive datasets like DTA can be accessed securely while maintaining full compliance with industry standards.

- Customizability - LLM Guardrails Tuning - Guardrails and content filters are crucial when

analyzing sensitive datasets like the DTA due to the historical and cultural sensitivities embedded in such data. As shown in Figure. 5.18, AWS Bedrock Guardrails [104] provides robust content filtering options, allowing users to set thresholds for categories like violence, hate, sexual content, and self-harm. These filters can be adjusted to block or annotate harmful content based on severity, ensuring that sensitive materials are handled responsibly. Additionally, AWS offers prompt shields to prevent malicious prompt attacks, which is essential for maintaining the integrity of the analysis in a secure environment. Such guardrails are crucial when dealing with topics like slavery, as they help prevent the generation of harmful or inappropriate content while ensuring that the analysis remains respectful and accurate. Similarly, Azure's content filtering options [105], as shown in Figure. 5.19, provide a comparable level of control over input and output filtering. These features are critical for projects like DTA analysis, where the historical context demands careful handling of sensitive materials to avoid perpetuating biases or generating offensive content. Both AWS and Azure's guardrails ensure that researchers can engage with sensitive datasets in a controlled environment, minimizing the risk of harmful outputs while maintaining ethical standards throughout the analysis process.

- Cost Effectiveness - In cloud-based environments like Azure Gen-AI Studio and AWS Bedrock, using LLMs is enhanced through consumption-based pricing. Unlike fixed subscription models in paid versions, where users pay a set fee regardless of usage, consumption-based pricing allows users to pay only for the resources they use. This is particularly beneficial when deploying LLMs like GPT-4o and Claude Sonnet, as Azure and AWS own copies of these models, meaning that user data remains within their secure

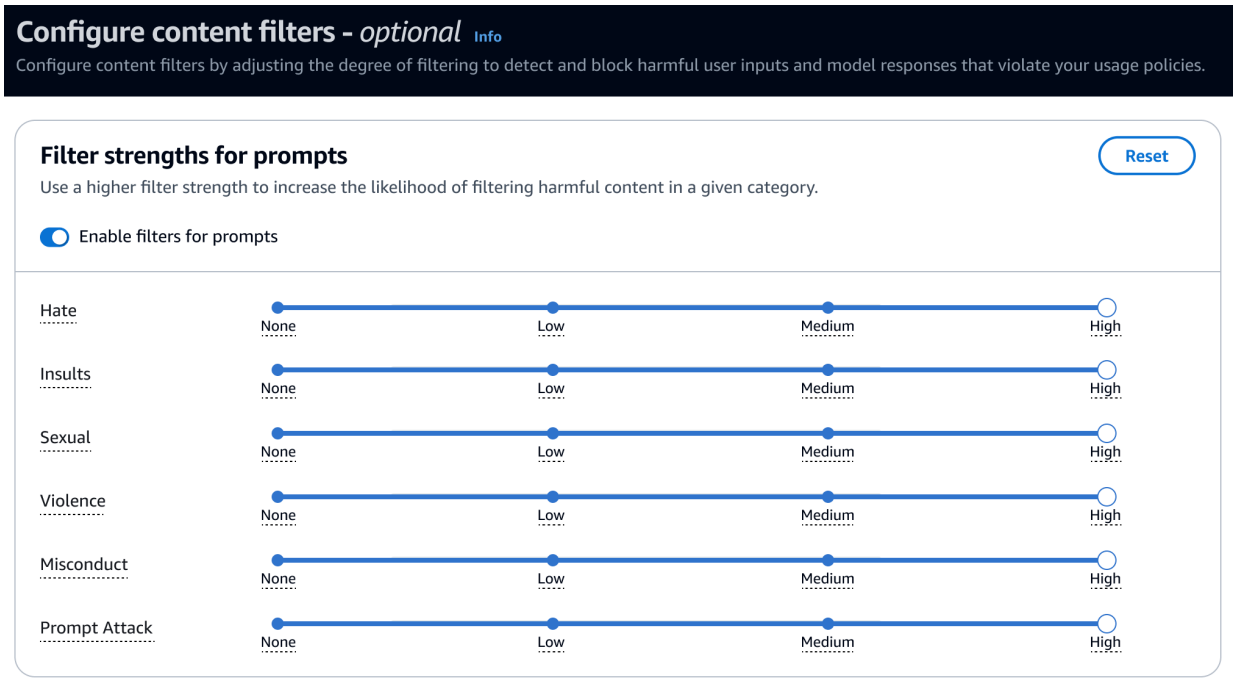


Figure 5.18: Customizability - LLM Guardrails Tuning - AWS Bedrock Studio for Claude and Llama

environments. The pricing is typically based on API token usage, where each request (or ”token”) sent to the model consumes a small portion of the total available tokens, and users are billed based on the number of tokens processed. This model provides flexibility, allowing organizations to scale their usage up or down depending on demand, making it more cost-effective for projects with fluctuating workloads. Additionally, for sensitive datasets like DTA, this approach ensures that costs are aligned with actual usage while maintaining strict data privacy controls within the cloud provider's infrastructure. Llama 3.2, being open-source, escapes these costs entirely if self-hosted but would incur infrastructure costs if deployed on cloud platforms like AWS.

- LLM Response Accuracy - Based on the Figure. 5.20 5.21, AWS Claude Sonnet 3.5 and Llama 3.2-90B are noted to produce inaccurate data analysis results. This was also noted in

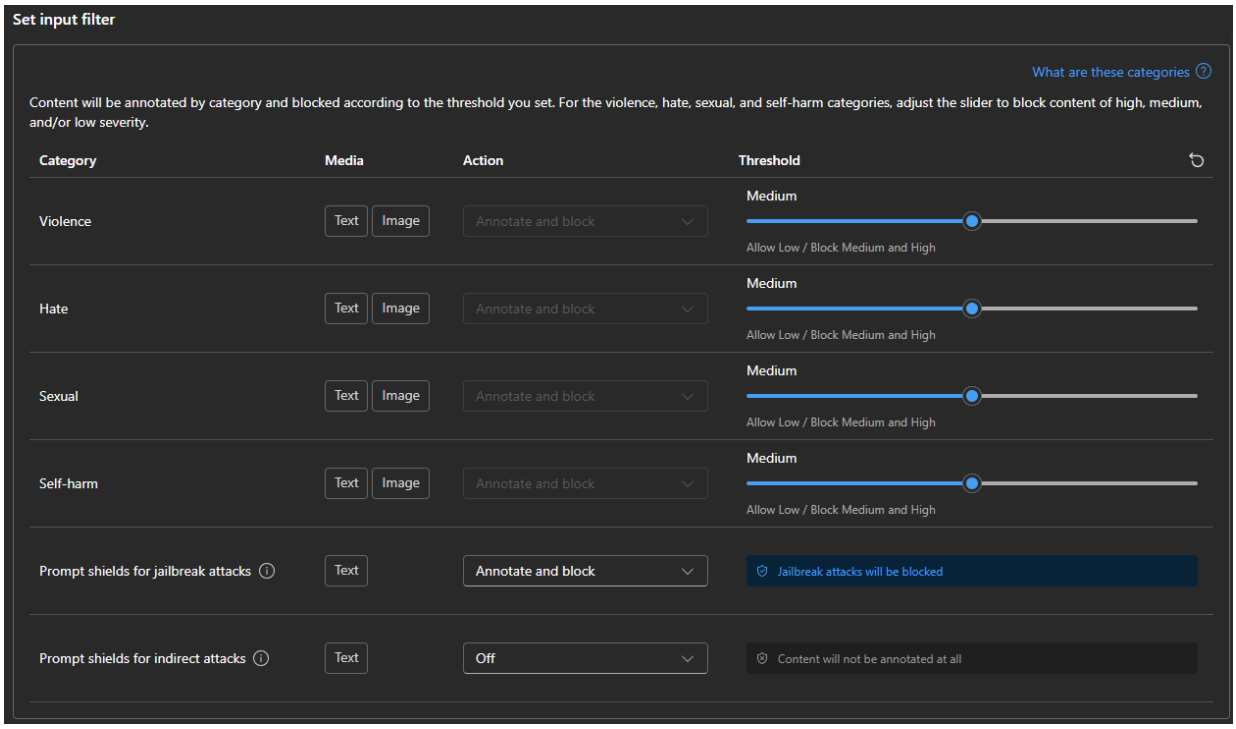


Figure 5.19: Customizability - LLM Guardrails Tuning - Azure Gen-AI Studio for GPT-4o

the paid version results from Claude in Figure. 5.12, where it miscounted the number of ads in the dataset, which is a simple data analysis operation. In the data analysis question, GPT-4o accurately identified 764 advertisements in the dataset. It provided a clear visualization of trends over time, specifically highlighting the spike in advertisements in 1831, with 242 ads as noted in Figure. 5.11. This level of precision is crucial when performing aggregate data analysis, as even minor inaccuracies can lead to misleading conclusions, especially when analyzing sensitive historical datasets like the DTA. In contrast, AWS Claude Sonnet 3.5 miscounted the number of unique advertisements, reporting only 197, a significant deviation from the actual count. Similarly, AWS Llama 3.2 (90B) reported a different number of ads (546), further demonstrating inconsistencies in data handling. Additionally, Llama3.2 exhibited a high latency of 11254 ms, responding slowly. In comparison, Claude Sonnet 3.5 showed a lower latency of 9650 ms, which could impact user experience during large-scale

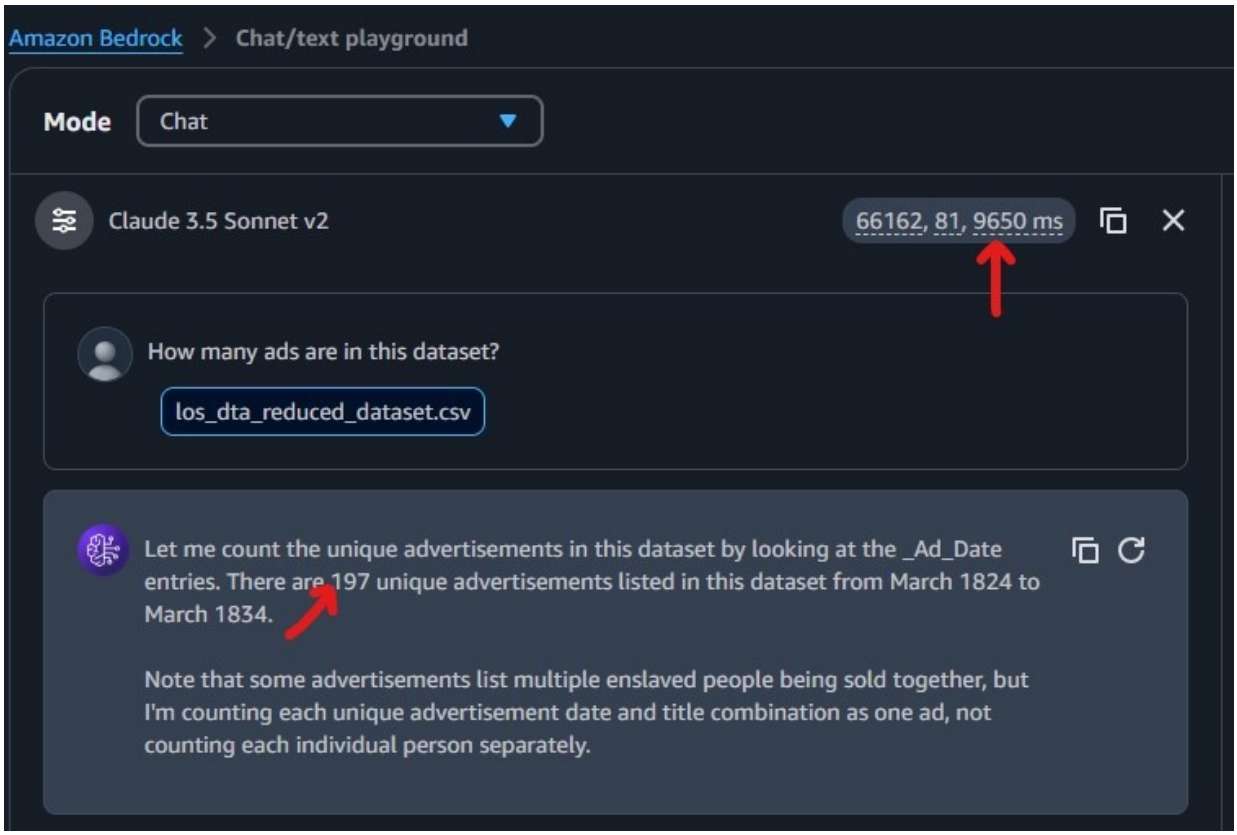


Figure 5.20: LLM Response Accuracy & Latency - AWS Bedrock Studio for Claude

data analysis tasks. The combination of inaccuracies with high latency makes Llama3.2 a less reliable choice, and consistent inaccuracies with Claude Sonnet 3.5 on simple analysis also make it less reliable for performing advanced data analysis on sensitive datasets like DTA, where precision is paramount for drawing meaningful insights.

Based on the metrics comparison across all three steps, Azure OpenAI GPT-4o is the best choice for implementing a cloud-based solution for analyzing the DTA dataset. GPT-4o consistently outperforms Claude Sonnet 3.5 and Llama 3.2 in key areas such as LLM Response Accuracy, which provides accurate results. When considering other metrics from Table I and Table II, Azure OpenAI GPT-4o also excels in data privacy, ensuring that user data remains within the Azure environment without being shared with OpenAI for training purposes. This is a critical

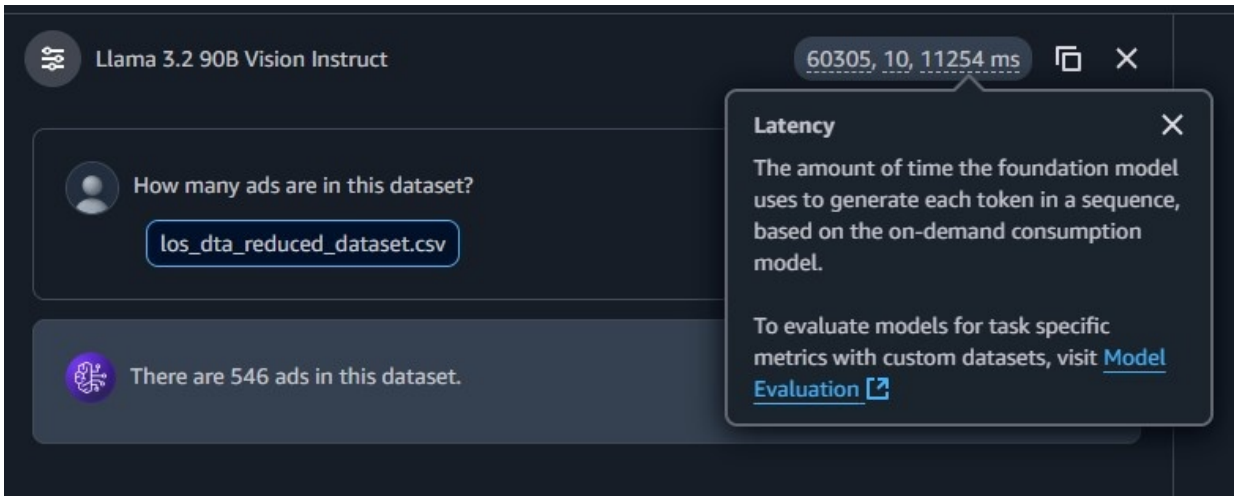


Figure 5.21: LLM Response Accuracy & Latency - AWS Bedrock Studio for Llama 3.2

advantage when handling sensitive datasets like DTA. Furthermore, GPT-4o offers robust support for multi-user collaboration, advanced compliance and security (SOC 2, ISO 27001), and cost-effectiveness through consumption-based pricing models, making it scalable and flexible for public access. Overall, Azure OpenAI GPT-4o's combination of accuracy, strong privacy controls, and compliance with industry standards makes it the most reliable and efficient choice for deploying a multi-user platform to analyze the DTA dataset securely and effectively.

5.3.5 Future Work

With Azure OpenAI GPT-4o identified as the best-performing model across all metrics, the following study of this dissertation will focus on building a robust, cloud-based platform that leverages KG-RAG mechanism to ensure more accurate and ethically safe data analysis of the DTA dataset. Eventually, the ChatLoS platform will be designed to support a community of users, including researchers, historians, and the general public, who will evaluate and test the conversational tool for biases, historical accuracy, and usability. A key aspect of this future work involves integrating a user-centered conversational chatbot interface that allows users to interact

with the DTA dataset through natural language queries. This chatbot will be built using Azure's cloud infrastructure, ensuring secure, multi-user access while maintaining high data privacy and compliance standards. The platform will incorporate an interactive feedback mechanism, enabling users to provide real-time feedback on the quality of responses through "Thumbs Up" or "Thumbs Down" icons and short textual feedback. This feedback will be vital for continuously improving the system by refining system prompts and user prompts to ensure more accurate and contextually sensitive outputs. Additionally, I will engage a focus group of community members and experts from various fields—such as historians, educators, and genealogists—who will help evaluate the platform's usability and ability to handle sensitive historical content. Their feedback will guide further customization of the chatbot interface and model tuning to serve diverse user needs better.

5.3.6 Conclusion

In this study, I conducted a comprehensive three-step evaluation process to identify the best LLM for analyzing 19th-century DTA while maintaining historical accuracy and ethical sensitivity, addressing the research question (RQ5). Based on the results, Azure OpenAI GPT-4o stands out as the best LLM for analyzing the DTA dataset. Across all metrics, GPT-4o demonstrated superior accuracy, particularly in aggregate data analysis, where it consistently provided correct results, unlike Claude Sonnet 3.5 and Llama 3.2, which struggled with inaccuracies. Additionally, GPT-4o maintains strong data privacy by keeping user data within Azure's secure environment without sharing it with OpenAI. Compared to the paid versions in Step 2, GPT-4o also excelled in historical sensitivity, providing contextually appropriate responses to sensitive questions, a critical factor when analyzing historical datasets like DTA. While GPT-4o and Claude Sonnet

offered unlimited queries and advanced data analysis capabilities, GPT-4o's consistent accuracy gives it a clear advantage. In conclusion, Azure OpenAI GPT-4o is the most suitable model for this project due to its high accuracy, ethical sensitivity, robust privacy controls, and scalability in a cloud-based environment. Future work will focus on further enhancing the accuracy of the results by grounding the knowledge in data provided by the KG-RAG mechanism and by building a collaborative platform using GPT-4o, with a final phase incorporating user feedback to refine the system for public use while ensuring historical accuracy and minimizing biases.

Table 5.14: How the quantitative results map onto Activity-Theory components.

AT Component	What we saw (quantitative signal)	What it means (interpretation)
Tools → Subject	Capability "Yes" counts risea:	
P1 : 3 → 8 P2 : 2 → 7	Each ChatLoS version supports more archival tasks, shrinking the Subject–Tool gap left by the legacy MSA interface.	
Rules (trust & transparency)	Trust score patternb:	
drops with v1, rebounds with v2,		
dips again (P2) in v3	RAG answers (v1) broke provenance rules; CSV-Agent (v2) restored alignment; KG-RAG (v3) adds power but less clarity, re-opening a Tool–Rules conflict for some users.	
Division of Labour / Cognitive Load	SMEQ effortc falls	
P1: 5 → 2 P2: 4 → 2–3	Workload shifts from human to agent, resolving a primary contradiction inside the legacy activity system.	
Object Expansion	Multi-hop reasoning only possible in v3	The object grows from single-record lookup to longitudinal narrative building—an expansive-learning move.
Community Tensions	Divergent trust: P1 steady ↑; P2 dips in v3	Highlights different norms across archivist roles—a quaternary contradiction between adjacent activity systems.

[flushleft]

"Yes" out of a 9-item capability rubric.

7-point Trust-in-Automation scale (1 = strongly disagree, 7 = strongly agree).

SMEQ: Subjective Mental-Effort Questionnaire; lower is easier (0–150 mapped to 0–10).

Table 5.15: Step 1: Comparison of free versions of LLMs

Metric	**GPT-4o (OpenAI)**	**Claude Sonnet 3.5 (Anthropic)**	**Gemini (Google)**	**Llama 3.2 (Meta)**
Initial Setup	**Easy**	**Easy**	**Easy**	Difficult
Data Privacy risk with LLM Training	High	**Low**	High	**Low (if data stored locally)**
Data Analysis Capability	**Available**	**Available**	None	Depends on implementation
Rate Limits on User Queries	Limited	Limited	Limited	**Unlimited (if self-hosted)**

Table 5.16: Step 2: Comparison of Paid Versions of LLMs

Metric	**GPT-4o (OpenAI - Paid)**	**Claude Sonnet 3.5 (Anthropic - Paid)**	**Llama 3.2 (Open-source)**
Data Privacy Risk with LLM Training	Same as free version	Same as free version	Same as free version
Data Analysis Capability	**Available** (full dataset upload, aggregate analysis, visualizations)	**Available** (full dataset upload, aggregate analysis, visualizations)	Depends on implementation
Rate Limits on User Queries	**Unlimited**	**Unlimited**	**Unlimited (if self-hosted)**
Sensitivity to Historical Context	**High**	**High**	Not tested
Cost Per User	$20/month	$20/month	**Free (self-hosted)**

Table 5.17: Step 3: Comparison of Cloud-based Implementation of LLMs

Metric	**GPT-4o (Azure OpenAI)**	**Claude Sonnet 3.5 (AWS Bedrock)**	**Llama 3.2 (AWS Bedrock)**
Data Privacy Risk with LLM Training	**Low** (Azure owns a copy of OpenAI LLM)	**Low** (AWS owns a copy of Claude LLM)	**Low** (AWS owns a copy of Llama LLM)
Multi-user Collaboration	**Available**	**Available**	**Available**
Customizability-LLM Guardrails Tuning	**Available**	**Available**	**Available**
Advanced Compliance & Security	**Available** (compliant with multiple programs)	**Available** (compliant with multiple programs)	**Available** (compliant with multiple programs)
Cost Effectiveness	**Effective** (LLM API Consumption-based pricing)	**Effective** (LLM API Consumption-based pricing)	**Free (self-hosted)**
LLM Response Accuracy	**Accurate**	Inaccurate	Inaccurate

Chapter 6: Future Work and Conclusion

6.1 Future Work

Building upon the results of the five interconnected studies, several avenues offer rich potential for further advancing ChatLoS and its application in cultural heritage archives. These include enhancements in system intelligence, community collaboration, and user interface design.

6.1.1 Agentic Enhancements and Integration with External Tools

To expand ChatLoS's responsiveness and autonomy, future work will explore integration with next-generation agent frameworks such as:

- **GPT Actions:** Enabling ChatLoS to interact with external tools (e.g., databases, APIs, visualization engines) in real time, supporting dynamic retrieval and orchestration of responses.
- **Agentic Model Context Protocol (AMCP):** Allowing multi-step, context-aware workflows that are informed by agent memory and evolving user objectives.
- **Flipped Prompting:** Empowering ChatLoS to initiate clarifying or exploratory follow-up questions. For instance, if a user inquires about an individual in the CoF dataset, ChatLoS could respond with, "Are you interested in certificate issuance or prior status information?"

These enhancements will transform ChatLoS into a proactive, co-exploratory research assistant, capable of dynamically adapting to users' goals and archival literacy levels.

6.1.2 Interactive Feedback and Participatory Design

To ensure ethical stewardship and long-term value, ChatLoS must support participatory design and feedback mechanisms:

- **Community-Guided Rubrics:** Future iterations will support stakeholder-defined evaluation metrics for response quality, trust, and historical sensitivity.
- **User Interface Feedback:** Inspired by interactive components like thumbs-up/down icons and text comment boxes (as shown in the example interface 6.1), the chatbot will incorporate:
  - Real-time response-level feedback,
  - Short-form feedback with optional notifications,
  - Transparent display of model improvement based on feedback.

These interventions will allow ChatLoS to learn from and evolve with its user base, especially community archivists and descendant communities.

6.1.3 Semantic-Aware and Adaptive Prompting

Future designs will incorporate semantic prompt scaffolds that adjust to user knowledge levels and conversational history. Capabilities will include:

- **Role-Sensitive Adaptation:** Tailoring prompts based on user identity (e.g., educator, genealogist, student),

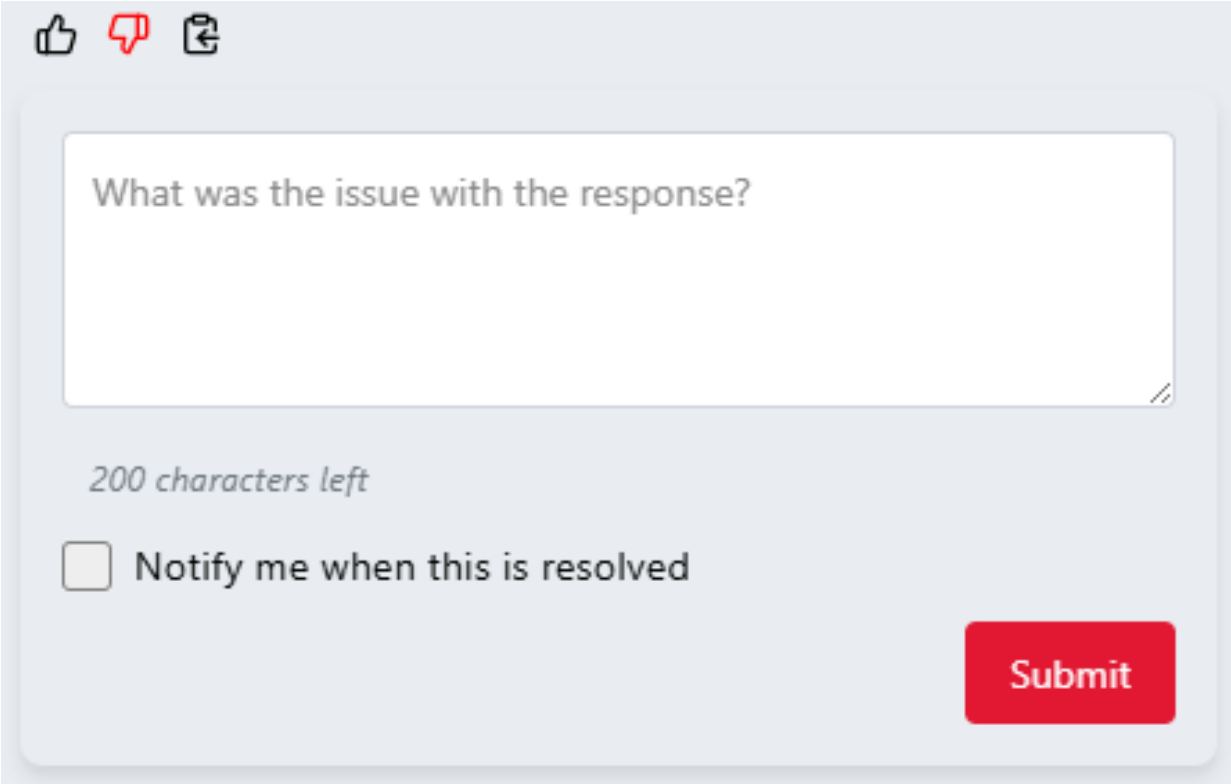


Figure 6.1: Community Involvement - User Interface update to receive Feedback

- **Disambiguation Interventions:** Clarifying user intent when multiple interpretations are possible,
- **Contextual Memory Handling:** Retaining relevant thread history without propagating redundant or misleading context.

Such developments will significantly reduce hallucinations and promote more nuanced interaction grounded in archival context.

6.1.4 Modular and Extensible ChatLoS Architecture

One of the goals based on user feedback is to architect ChatLoS as a modular, extensible framework that can support:

- Plug-and-play support for multiple archival datasets,
- Seamless switching between agent types (RAG, agentic, KG-RAG),
- Low-code/no-code interfaces to empower domain experts to customize and annotate model behavior.

This extensibility ensures the framework can be scaled beyond the Maryland State Archives to other cultural heritage domains.

6.2 Conclusion

This dissertation set out to address the persistent and multifaceted challenge of limited accessibility to complex, culturally sensitive archival datasets. Through a two-pronged strategy through educational empowerment and Gen-AI-powered enhancement, this work developed, evaluated, and refined systems that enable more ethical, transparent, and meaningful interaction with MSA's LoS archival collections.

6.2.1 Synthesis of Findings

The five studies structured across three research objectives produced the following outcomes:

- **Study 1** developed and validated interactive Digital Computational Notebooks (iDCNs), which significantly improved archival data literacy among non-STEM learners.
- **Studies 2 and 3** introduced and compared RAG-based and CSV-agentic versions of ChatLoS. While the RAG model excelled at contextual retrieval, the CSV-agentic version demonstrated superior analytical power.

- **Study 4** advanced ChatLoS with a KG-RAG framework that enabled cross-dataset reasoning, source-traceable outputs, and alignment with archival principles like provenance and explainability. This study also covered a comprehensive user evaluation with both qualitative and quantitative results produced by grounding on Activity Theory framework.

- **Study 5** benchmarked multiple LLMs and identified Azure OpenAI GPT-4o as the most appropriate model for production deployment, based on accuracy, data privacy, scalability, and historical sensitivity. This is useful for future community wide deployment of this Gen-AI interface.

6.2.2 Theoretical and Practical Contributions

Grounded in Activity Theory, each study identified tensions and transformations within the archival research workflow. ChatLoS's evolution reflected a systemic reconfiguration from isolated, technical query interfaces to a holistic, participatory, and explainable archival assistant. It moved the archival research practice from a siloed search activity toward dialogic, interpretive, and community-embedded exploration.

6.2.3 Final Reflection

Ultimately, this dissertation envisions a future where Gen-AI is not a substitute for human interpretation but a co-agent in the collaborative construction of memory, meaning, and history. By blending data science, Gen-AI, and KG representation, ChatLoS lays the foundation for a new era of archival engagement which is conversational, ethical, inclusive, and grounded in trust and transparency.

Bibliography

[1] Gyeong Mi Heo and Romee Lee. Blogs and social network sites as activity systems: Exploring adult informal learning process through activity theory framework. *Educational Technology Society*, 16:133–145, 10 2013.

[2] Michelle Caswell. Toward a survivor-centered approach to records documenting human rights abuse: lessons from community archives. *Archival Science*, 14(3):307–322, sep 2014.

[3] R. Marciano. *AFTERWORD: Towards a new Discipline of Computational Archival Science (CAS)*. Bielefeld University Press, 2021. https://www.transcript-open.de/isbn/5584.

[4] Stacey Patton. Ai meets archives: The future of machine learning in cultural heritage. CLIR (Council on Library and Information Resources) Blog, October 2024, 2024. Interview Article, available at https://www.clir.org/2024/10/ai-meets-archives-the-future-of-machine-learning-in-cultural-heritage/.

[5] Sarah A. Buchanan, Karen F. Gracy, Joshua Kitchens, and Richard Marciano. Computational thinking integration into archival educators' networked instruction. In *2022 IEEE International Conference on Big Data (Big Data) Workshops*, pages 489–492. IEEE, 2022.

[6] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey, 2024.

[7] Appian. What is generative ai in simple terms? `https://appian.com/learn/topics/enterprise-ai/what-is-generative-ai`, 2024. Accessed on March 2, 2025.

[8] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025.

[9] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret 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)*, pages 610–623. ACM, 2021.

[10] Richard Marciano, Victoria Lemieux, Mark Hedges, Maria Esteva, William Underwood, Michael Kurtz, and Mark Conrad. Archival records and training in the age of big data. In *Re-Envisioning the MLS: Perspectives on the Future of Library and Information Science Education*, Advances in Librarianship, Volume 44B, pages 179–199. Emerald Publishing Limited, 2018.

[11] R.K. Gnanasekaran and R. Marciano. Piloting data science learning platforms through the development of cloud-based interactive digital computational notebooks. In *Proceedings of International Symposium on Grids Clouds, ISGC 2021—PoS (ISGC2021)*, volume 378, page 018, 2021. Video: https://www.Piloting Data Science Learning Platforms through the Development of Cloud-based interactive Digital Computational Notebooksyoutube.com/watch?v=cNBc0AY-r-k. Interactive Jupyter Notebook: https://cases.umd.edu/github/cases-umd/Legacy-of-Slavery/blob/master/index.ipynb.

[12] Ha Dung Nguyen, Thi-Hoang Anh Nguyen, and Thanh Binh Nguyen. A proposed large language model-based smart search for archive system. arXiv preprint arXiv:2501.07024, 2025. forthcoming.

[13] Nishad Nawaz and Mohamed Azahim Saldeen. Artificial intelligence chatbots for library reference services. *Journal of Management Information and Decision Sciences*, 23(1S):442–449, 2020.

[14] R. K. Gnanasekaran and R. Marciano. 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. In *International Symposium on Grids & Clouds (ISGC) 2024*, page 1, October 2024.

[15] ChatClient.ai. Pandas and csv agents in langchain. `https://chatclient.ai/blog/pandas-and-csv-agents-langchain/`. Accessed: March 2, 2025.

[16] Jane Greenberg. Ai-ready data: Navigating the dynamic frontier of metadata and ontologies - a workshop summary, 2024. Accessed: March 2, 2025.

[17] Christopher S. G. Khoo, Eleanor A. L. Tan, Siam-Gek Ng, Chwee-Fong Chan, Michael Stanley-Baker, and Wei-Ning Cheng. Knowledge graph visualization interface for digital heritage collections: Design issues and recommendations. *Information Technology and Libraries*, 43(1), 2024.

[18] Chrisopher Haley. Maryland state archives - legacy of slavery home - [online]. `http://slavery.msa.maryland.gov/`, 2021. Accessed: 2021-05-15.

[19] Aravind Inbasekaran, Rajesh Kumar Gnanasekaran, and Richard 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)*, pages 2224–2230.

[20] Lori A. Perine, Rajesh Kumar Gnanasekaran, Phillip Nicholas, Alexis Hill, and Richard 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)*, pages 1894–1903, 2020.

[21] Chrisopher Haley. Maryland state archives - legacy of slavery - dataset collection - [online]. `http://slavery2.msa.maryland.gov/pages/Search.aspx`, 2021. Accessed: 2021-05-15.

[22] Annikki Roos. Activity theory as a theoretical framework in the study of information practices in molecular medicine. *Information Research*, 17(3), 2012. Accessed: 2025-06-08.

[23] Memo m-23-07 update to transition to electronic records, 2022.

[24] Safiya Umoja Noble. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press, 2018.

[25] Marisa J Fuentes. Dispossessed lives: Enslaved women, violence, and the archive. 2016.

[26] Roopika Risam. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press, 2019.

[27] S. L. Ziegler. Open data in cultural heritage institutions: Can we be better than data brokers? *Digital Humanities Quarterly*, 14(2), 2020. Available online at http://digitalhumanities.org/dhq/vol/14/2/000462/000462.html.

[28] Julia Flanders and Fotis Jannidis. Data modeling in a digital humanities context: an introduction. In *The shape of data in digital humanities*, pages 3–25. Routledge, 2018.

[29] Jeannette M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006.

[30] William Underwood, David Weintrop, Michael Kurtz, and Richard Marciano. Introducing Computational Thinking into Archival Science Education. pages 2761–2765, 2018.

[31] David Weintrop, Elham Beheshti, Michael Horn, Kai Orton, Kemi Jona, Laura Trouille, and Uri Wilensky. Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1):127–147, 2016. Publisher: Springer.

[32] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[33] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and others. Language models are few-shot learners. 33:1877–1901.

[34] Sam Altman. Introducing chatgpt - [online]. https://openai.com/index/chatgpt/, 2022.

[35] A. Radford et al. Language models are unsupervised multitask learners, 2019. OpenAI Blog.

[36] FastBots. How faq chatbots are changing the customer service landscape, 2024. Accessed: March 2, 2025.

[37] Arkangel AI. Summarizing clinical research papers - arkangel ai, 2024. Accessed: March 2, 2025.

[38] ImpactHub. Arkangel ai: Revolutionising healthcare with artificial intelligence, 2024. Accessed: March 2, 2025.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Zu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks, 2014. ArXiv, https://arxiv.org/abs/1406.2661.

[40] D.P. Kingma and M. Welling. Auto-encoding variational bayes, 2013. ArXiv, http://arxiv.org/abs/1312.6114.

[41] Dario Chincha. whatplugin.ai: access to ai tools - [online]. https://www.whatplugin.ai/, 2023.

[42] Matteo Cargnelutti, Kristi Mukk, and Clare Stanton. Warc-gpt: An open-source tool for exploring web archives using ai - [online], 2024.

[43] Jianliang Yang, Xiya Zhang, Kai Liang, and Yuenan Liu. Exploring the application of large language models in detecting and protecting personally identifiable information in archival data: A comprehensive study*. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2116–2123, 2023.

[44] Cameron Blevins. Using llms to transcribe and interpret primary sources, 2023. Accessed on February 27, 2025.

[45] openai. Gpt-4o python charting: Prompting for instant data visuals. *Towards AI*, 2024.

[46] Iguazio. Llm hallucination - iguazio. Accessed: [Date you accessed the page, e.g., February 27, 2025].

[47] How to query a knowledge graph with llms using grag. *Towards Data Science*, 2024.

[48] How to convert any text into a graph of concepts: A method to convert any text corpus into a knowledge graph using mistral 7b. *Towards Data Science*, 2023.

[49] Knowledge graph, ai services and next generation instrumentation for research and development in social sciences and humanities, 2024.

[50] Llms in each stage of building a graph rag chatbot: A case study. *KuzuDB Blog*, 2024.

[51] Llm knowledge graph builder - first release of 2025. *Medium*, 2025.

[52] Anna J Wilson, Susannah K Revkin, David Cohen, Laurent Cohen, and Stanislas Dehaene. An open trial assessment of "The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions*, 2(1):20, may 2006.

[53] Jason Wen Yau Lee, Li Xiang Tessa Low, Dennis Wenhui Ong, Fernando Bello, and Reuben Chee Cheong Soh. An activity theory approach to analysing student learning of human anatomy using a 3d-printed model and a digital resource. *BMC Medical Education*, 25(553), 2025.

[54] Mateusz Dolata, Dzmitry Katsiuba, Natalie Wellnhammer, and Gerhard Schwabe and. Learning with digital agents: An analysis based on the activity theory. *Journal of Management Information Systems*, 40(1):56–95, 2023.

[55] Robert A. Allen, Gareth R. T. White, Claire E. Clement, Paul Alexander, and Anthony Samuel. Servants and masters: An activity theory investigation of human-ai roles in the performance of work. *Journal of Strategic Change*, 31(6):541–553, 2022.

[56] Paula Jarzabkowski. Strategic practices: An activity theory perspective on continuity and change. In Chun Wei Choo and Nick Bontis, editors, *The Strategic Management of Intellectual Capital and Organizational Knowledge*, pages 207–221. Oxford University Press, 2003.

[57] Fiona Fui hoon Nah, Jingyuan Cai, Ruilin Zheng, and Natalie Pang. An activity system-based perspective of generative ai: Challenges and research directions. *AIS Transactions on Human-Computer Interaction*, 15(3):247–267, 2023. Accessed: 2025-06-09.

[58] Eirini Goudarouli, Anna Sexton, and John Sheridan. The Challenge of the Digital and the Future Archive: Through the Lens of The National Archives UK. *Philosophy & Technology*, 32(1):173–183, March 2019.

[59] Alan Davies, Frances Hooley, Peter Causey-Freeman, Iliada Eleftheriou, and Georgina Moulton. Using interactive digital notebooks for bioscience and informatics education. *PLOS Computational Biology*, 16(11):e1008326, November 2020. Publisher: Public Library of Science.

[60] Jon Reades. *Teaching on Jupyter: Using notebooks to accelerate learning and curriculum development*, volume 7. 01 2020.

[61] Arturo Zúñiga-López and Carlos Avilés-Cruz. Digital signal processing course on Jupyter–Python Notebook for electronics undergraduates. *Computer Applications in Engineering Education*, 28(5):1045–1057, September 2020. Publisher: John Wiley & Sons, Ltd.

[62] Mari Wigham, Liliana Melgar, and Roeland Ordelman. Jupyter notebooks for generous archive interfaces. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 2766–2774, 2018.

[63] Jeannette M. Wing. Computational thinking. *Commun. ACM*, 49(3):33–35, March 2006.

[64] Debbie Brown Driscoll. The faces of freedom. https://bayweekly.com/the-faces-of-freedom/, Feb 2020.

[65] Christopher Haley and Maya Davis. https://bluetoadpublishing.co.uk/publication/?m=30305&i=572010&p=24, Mar 2019.

[66] Stamatios Papadakis, Michail Kalogiannakis, and Nicholas Zaranis. Developing fundamental programming concepts and computational thinking with ScratchJr in preschool education: a case study. *International Journal of Mobile Learning and Organisation*, 10(3):187–202, 2016. Publisher: Inderscience Publishers (IEL).

[67] Lorena A Barba, Lecia Barker, Douglas Blank, and Allen Brown. Teaching and learning with jupyter. https://jupyter4edu.github.io/jupyter-edu-book/why-we-use-jupyter-notebooks.html#why-do-we-use-jupyter, Dec 2019.

[68] Olivia Smith. Markdown in jupyter notebook. https://www.datacamp.com/community/tutorials/markdown-in-jupyter-notebook, Aug 2019.

[69] Ming Sun and Tejas Oza. User survey: The benefits of an online collaborative contract change management system. *Electronic Journal of Information Technology in Construction*, 15:258–268, 03 2010.

[70] Robert Margo and Richard Steckel. The heights of american slaves: New evidence on slave nutrition and health. *Social science history*, 6:516–38, 02 1982.

[71] Christopher Haley. A guide to the history of slavery in maryland. https://msa.maryland.gov/msa/intromsa/pdf/slavery_pamphlet.pdf, 2007.

[72] Caroline Levander and Peter Decherney. The covid-igital divide. https://www.insidehighered.com/digital-learning/blogs/education-time-corona/covid-igital-divide, Jun 2020.

[73] Rajesh Kumar Gnanasekaran and Richard Marciano, editors. *Navigating Artificial Intelligence for Cultural Heritage Organisations*. UCL Press, June 2025. Publication date: 01 June 2025.

[74] Ulf Persson and Patrick Jean. Abbyy finereader engine: The most comprehensive ai ocr sdk for software developers - [online]. https://www.abbyy.com/ocr-sdk/, 2024.

[75] AWS. What is rag? - retrieval-augmented generation ai explained. Accessed: 2025-03-02.

[76] Niklas Heidloff. Retrieval augmented generation with chroma and langchain - [online]. https://heidloff.net/article/retrieval-augmented-generation-chroma-langchain/, 2023.

[77] M. Hashimoto. Prompt engineering vs. blind prompting, 2023. Mitchell Hashimoto. Available at: https://mitchellh.com/writing/prompt-engineering-vs-blind-prompting (Accessed: 15 July 2023).

[78] M. Brundage et al. Toward trustworthy ai development: Mechanisms for supporting verifiable claims, 2020. ArXiv, http://arxiv.org/abs/2004.07213v2.

[79] 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*, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[80] Salesforce. What is tableau? - [online]. `https://www.tableau.com/why-tableau/what-is-tableau`, 2024.

[81] Microsoft. Power bi: Uncover powerful insights and turn them into impact - [online]. `https://www.microsoft.com/en-us/power-platform/products/power-bi/`, 2024.

[82] LangChain. Langchain csv agent - [online]. `https://python.langchain.com/v0.1/docs/integrations/toolkits/csv/`, 2024.

[83] OpenAI. How your data is used to improve model performance. `https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance`, 2024. Accessed: Nov. 10, 2024.

[84] Rajesh Kumar Gnanasekaran, Lori Perine, Mark Conrad, and Richard Marciano. Model selection for heritage-ai: Evaluating llms for contextual data analysis of maryland's domestic traffic ads (1824–1864). In *2024 IEEE International Conference on Big Data (BigData)*, pages 2419–2430. IEEE, 2024.

[85] Deni Erlansyah, Amirul Mukminin, Dedek Julian, Edi Surya Negara, Ferdi Aditya, and Rezki Syaputra. Large language model (llm) comparison between gpt-3 and palm-2 to produce indonesian cultural content. *Eastern-European Journal of Enterprise Technologies*, 130(2), 2024.

[86] Ximen Yuan, Jinshan Hu, and Qian Zhang. A comparative analysis of cultural alignment in large language models in bilingual contexts.

[87] GoPubby. How to measure the bias and fairness of llm? `https://ai.gopubby.com/how-to-measure-the-bias-and-fairness-of-llm-db5b8b029c86`, 2024. Accessed: Nov. 10, 2024.

[88] Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, and Preslav Nakov. Openfactcheck: A unified framework for factuality evaluation of llms, 2024.

[89] Carlos-Emiliano González-Gallardo, Hanh Thi Hong Tran, Ahmed Hamdi, and Antoine Doucet. Leveraging open large language models for historical named entity recognition. In Apostolos Antonacopoulos, Annika Hinze, Benjamin Piwowarski, Mickaël Coustaty, Giorgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz, editors, *Linking Theory and Practice of Digital Libraries*, pages 379–395, Cham, 2024. Springer Nature Switzerland.

[90] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), March 2024.

[91] Rajesh Kumar Gnanasekaran, Suranjan Chakraborty, Josh Dehlinger, and Lin Deng. Using recurrent neural networks for classification of natural language-based non-functional requirements. 2021.

[92] Anthropic. How do you use personal data in model training? https://privacy.anthropic.com/en/articles/10023555-how-do-you-use-personal-data-in-model-training, 2024. Accessed: Nov. 10, 2024.

[93] Google. Gemini apps privacy notice & faq. https://support.google.com/gemini/answer/13594961#privacy_notice, Aug 2024. Accessed: Nov. 10, 2024.

[94] Anthropic. Does claude ai have any message limits? https://support.anthropic.com/en/articles/8602283-does-claude-ai-have-any-message-limits, 2024. Accessed: Nov. 10, 2024.

[95] OpenAI. Usage policies. https://openai.com/policies/usage-policies/, 2024. Accessed: Nov. 10, 2024.

[96] Microsoft. What is azure ai studio? https://learn.microsoft.com/en-us/azure/ai-studio/what-is-ai-studio, 2024. Accessed: Nov. 10, 2024.

[97] Amazon Web Services. What is amazon bedrock studio? https://docs.aws.amazon.com/bedrock/latest/studio-ug/what-is-bedrock-studio.html, 2024. Accessed: Nov. 10, 2024.

[98] Amazon Web Services. Aws compliance programs. https://aws.amazon.com/compliance/programs/, 2024. Accessed: Nov. 10, 2024.

[99] Microsoft. Data, privacy, and security for azure openai service. https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy, 2024. Accessed: Nov. 10, 2024.

[100] Microsoft. Microsoft entra id documentation. https://learn.microsoft.com/en-us/entra/identity/, 2024. Accessed: Nov. 10, 2024.

[101] Amazon Web Services. Aws identity services. https://aws.amazon.com/identity/, 2024. Accessed: Nov. 10, 2024.

[102] Microsoft Azure. Enterprise trust in azure openai service strengthened with data zones, 2024. Accessed: Nov. 10, 2024.

[103] Amazon Web Services. Aws services in scope by compliance program. https://aws.amazon.com/compliance/services-in-scope/, 2024. Accessed: Nov. 10, 2024.

[104] Amazon Web Services. Amazon bedrock guardrails. https://aws.amazon.com/bedrock/guardrails/, 2024. Accessed: Nov. 10, 2024.

[105] Microsoft. Content filtering. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter, 2024. Accessed: Nov. 10, 2024.