

Can GPT-4 Think Computationally About Digital Archival Practices?

William Underwood
Joan Gage

Computational Archival Science Workshop #8
IEEE Big Data Conference 2023
Sorrento, Italy, December 17, 2023



UNIVERSITY OF
MARYLAND



COLLEGE OF
INFORMATION
STUDIES

Advanced
Information
Collaboratory

Digital Archival Practices

- Classify digital records according to record schedules
- Recognize personal identifiable information (PII) in digital records
- Protect the integrity of digital records
- Review records for restrictions on release
- Define the requirements for long-term preservation of digital records
- Improve access to relevant records
- Reason about the trustworthiness of archival systems
- Restrict access to digital records
- Extract metadata from digital records
- Describe digital images
- Describe digital record series
- Backup and recovery of digital records
- Appraise long-term value of digital records
- Convert file formats of digital records
- Recognize file format of digital records
- Render (view, play) digital records
- Verify digital signatures
- And many more

Computational Thinking

“Computational thinking is a way humans solve problems.”

“Computational methods and models give us the courage to solve problems and design systems that no one of us would be capable of tackling alone.”

“Computational thinking is a fundamental skill for everyone, not just for computer scientists. To reading, writing, and arithmetic, we should add computational thinking to every child’s analytical ability.”

Jeanette Wing. Computational Thinking. *Communications of the ACM*, Mar 2006

Data Practices

Collecting Data

Creating Data

Manipulating Data

Analyzing Data

D. Visualizing Data

Modeling & Simulation Practices

Using Computational Models to Understand a Concept

Using Computational Models to Find and Test Solutions

Assessing Computational Models

Designing Computational Models

F. Constructing Computational Models

Computational Problem Solving Practices

Preparing Problems for Computational Solutions

Programming

Choosing Effective Computational Tools

Assessing Different Approaches/Solutions to a Problem

H. Developing Modular Computational Solutions

Creating Computational Abstractions

I. Troubleshooting and Debugging

Systems Thinking Practices

Investigating a Complex System as a Whole

Understanding the Relationships within a System

Thinking in Levels

Communicating Information about a System

Defining Systems and Managing Complexity

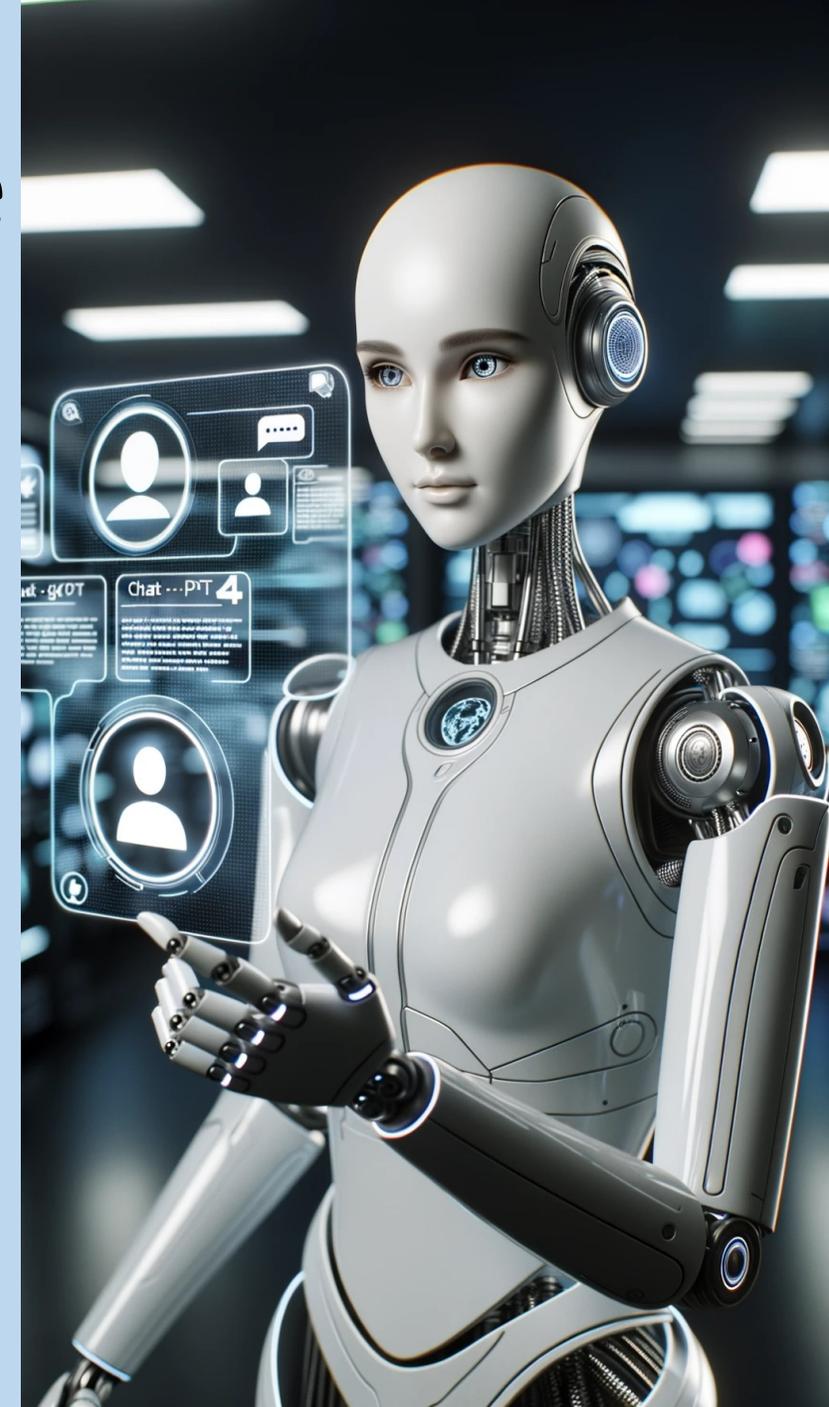
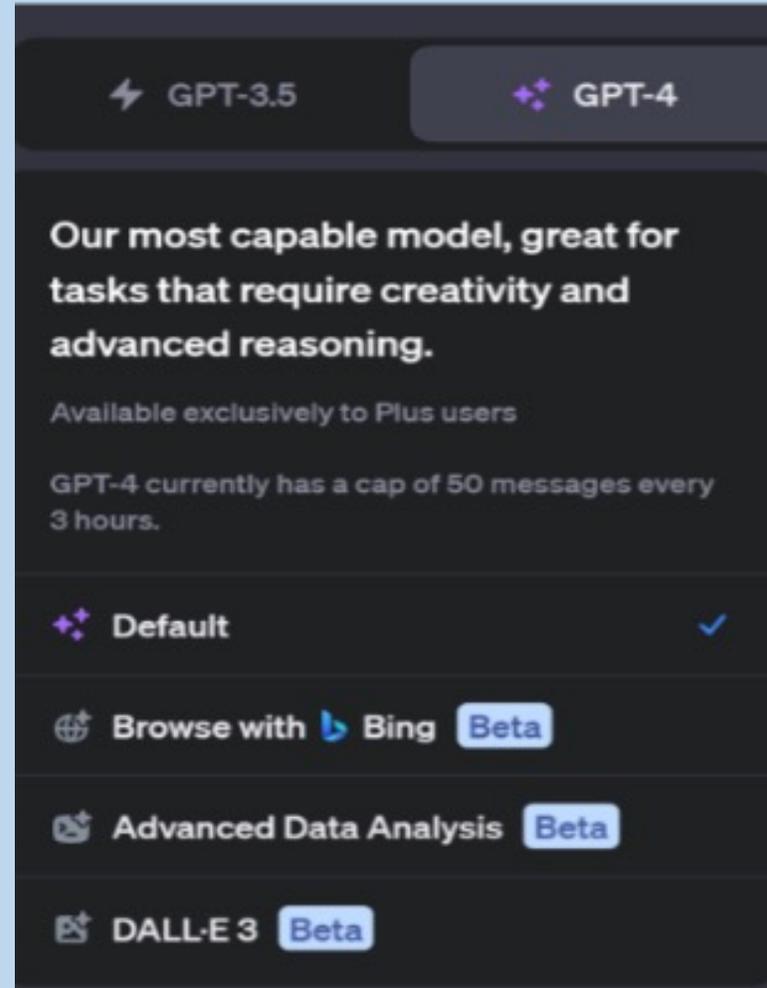
Computational Thinking (CT) Practices Learned By MLIS Grad Students in the Detection of Personally Identifiable Information (PII) Case Study

| Digital Archival Task | CT Practice |
|--|--|
| A scanner was used to collect digital image data of a paper index card. | Collecting Data |
| <i>Abbyy FineReader</i> was used to create digital text from images of index cards. | Creating Data |
| The <i>General Architecture for Text Engineering (GATE)</i> was used to analyze the digital text and extract metadata by creating rules to perform Named Entity Recognition (NER). | Analyzing Data |
| <i>Refine</i> was used to clean and normalize the data. | Manipulating Data |
| The <i>Python</i> programming library called <i>matplotlib</i> was used to create graphs and charts to visualize and understand the results of the analysis. | Visualizing Data |
| A flowchart was created that represents the input metadata, the input FAR and Form26 databases, the computations on the data and the decisions necessary to conclude whether an index card has PII requiring restriction on release. | Designing Computational Models |
| Pseudocode was constructed from the flowchart. | Constructing Computational Models |
| The pseudocode was encoded into the <i>Python</i> programming language. | Programming |
| <i>Python</i> functions were developed for looking up dates in the FAR and Form 26 databases and for comparing birthdates from the databases with dates from the index cards. | Developing Modular Computational Solutions |
| <i>A Jupyter Notebook</i> was used to support debugging of the <i>Python</i> program for recognizing PII. | Troubleshooting and Debugging |

GPT Technology

- Generative Pre-trained Transformer 4 (GPT-4) is a decoder-only transformer deep neural network developed by OpenAI.
- It was pre-trained on an enormous and enormous corpus of text.
- Given a prompt from a user and possibly additional context for the prompt, the network predicts the response to the prompt.
- OpenAI reports that GPT-4 has recorded a score of 1410 (94% level) on the questions of the Scholastic Aptitude Test.
- GPT-4 sometimes generate responses that are incorrect.

ChatGPT+ Interface



Creating Data

“In many cases, scientists and mathematicians use computational tools to generate data. ... Computational tools allow for data creation at scales that would otherwise be impossible. Students who have mastered this practice will be able to define computational procedures and run simulations that create data they can use to advance their understanding of the topic under investigation.”

Conversation with ChatGPT+ about a Digital Archival Practice that Involves Creating Data

JO: Can you list 3 of the best OCR programs for Windows 10, to convert the images to text?

ChatGPT+: Certainly, Optical Character Recognition (OCR) software can vary in terms of accuracy, features, and ease of use. Here are three OCR programs that are highly regarded and compatible with Windows 10, each with its own set of advantages:

1. Adobe Acrobat Pro DC

Advantages:

High accuracy in text recognition. Integrated with Adobe's suite of products, making it easier to edit PDFs and other document types. Offers advanced features like form recognition, document comparison, and batch processing.

Disadvantages:

Expensive, usually requiring a subscription. May offer more features than needed for simple OCR tasks.

2. ABBYY FineReader

Advantages: Extremely accurate OCR capabilities. Supports multiple languages. Can recognize complex layouts, including tables and graphs.

Disadvantages: The cost can be high for the full version. The interface may be overwhelming for first-time users due to its range of features.

3. Tesseract

Advantages: Open-source and free to use. Good OCR accuracy, especially for high-quality scans. Extensible and can be integrated into custom solutions.

Disadvantages: Requires technical know-how for installation and usage. Lacks a user-friendly interface; generally run through the command line or integrated into other software.

Evaluation

ChatGPT+ [GPT-4 alone] describes three methods of Optical Character Recognition showing that it knows how to create textual data from images containing text.

JO selected the option for ChatGPT+ to use Advanced Data Analysis, inserted an image of one of the 3x5 cards and asked ChatGPT+ to extract text from the image. ChatGPT+ used Python Tesseract, one of the methods it previously recommended, to OCR the image.

ChatGPT+ not only knows how to OCR images of text but is able to do so.

Summary of ChatGPT's Knowledge of Digital Archival Tasks and CD Data Practice Ability

| Digital Archival Practice | GPT-4 Know How | GPT-4 with Plugins Performance | CT Data Practice |
|---|----------------|--------------------------------|-------------------|
| Digitization (Scanning) | ✓ | | Collecting Data |
| OCR | ✓ | Python Tesseract | Creating Data |
| Clean and Normalize Text | ✓ | Python Library | Manipulating Data |
| Metadata Extraction | ✓ | Chrome Extension | Analyzing Data |
| Create Bar Chart for Distribution of Record dates | ✓ | Advanced Data Analysis | Visualizing Data |

Investigating a Complex System as a Whole

“Investigating a complex system as a whole relies on the ability to define and measure inputs and outputs of the system. This is especially critical in the sciences because so many phenomena are the result of very large-scale, complex interactions. Being able to black box the details of the underlying systematic interactions and focus on the system as a whole makes it possible to understand the characteristics of the system in aggregate. Computational tools such as models and simulations are especially useful in such investigations. ... Students who have mastered this practice will be able to pose questions about, design and carry out investigations on, and ultimately interpret and make sense of, the data gathered about a system as a single entity .”

Conversation with ChatGPT+ about a Digital Archival Practice that Involves Systems Thinking

JO: The Record Life Cycle Model distinguishes managing active digital records from archiving inactive records that are of long-term value. Using the IDEF0 modeling language, what would be the inputs, outputs, controls, and mechanisms to the Context Diagram of the function Archive Digital Records?

ChatGPT+: The Record Life Cycle Model outlines the stages that records go through from creation to disposal, including stages like creation, use, storage, and eventual disposition. Archiving digital records is a particular function within this cycle that deals with long-term storage, maintenance, and provision for digital records that have become inactive but are still valuable.

When it comes to IDEF0 modeling language, you would typically develop a Context Diagram to represent the highest level in the functional decomposition of a system. In this case, the system's primary function would be "Archive Digital Records."

Here's how the different elements (Inputs, Outputs, Controls, and Mechanisms) might break down for the "Archive Digital Records" function:

Inputs:

- **Inactive Digital Records:** These are records that are no longer actively used but identified as valuable for long-term retention.
- **Metadata:** Information describing the characteristics, context, and content of records.
- **Archiving Policies:** Guidelines that define what, when, and how to archive.

Outputs:

- **Archived Digital Records:** Inactive digital records that have been successfully stored in the archival system.
- **Access Logs:** Records that show who accessed the archived records, when, and what operations were performed.
- **Compliance Reports:** Documentation verifying that the archived records meet legal and organizational standards.

Continued Conversation with ChatGPT+ about a Digital Archival Practice that Involves Systems Thinking

Controls:

- **Legal Requirements:** Laws and regulations that mandate the retention and safeguarding of certain records.
- **Organizational Guidelines:** Internal policies that guide what types of records to archive and for how long.
- **Retention Schedules:** Specific timelines outlining how long various types of records must be kept.
- **Data Integrity Checks:** Ensuring that the data remains unaltered and accessible over time.

Mechanisms:

- **Archiving Software:** The system used to actually store and manage the digital records.
- **Storage Infrastructure:** Physical or cloud-based storage solutions where the digital archives are kept.
- **Human Resources:** Trained staff responsible for managing the archiving process.
- **Security Measures:** Firewalls, encryption, and other technologies to protect the integrity and confidentiality of archived records.

In a Context Diagram, you would show the main function ("Archive Digital Records") in a central box, and then have arrows connecting to that box to represent each of the Inputs, Outputs, Controls, and Mechanisms. This will provide a high-level view of how the function interacts with the various elements.

By carefully thinking through these aspects and how they interrelate, you can lay the foundation for a more detailed IDEF0 model that decomposes the "Archive Digital Records" function into its sub-functions, providing a comprehensive view of what the archiving process entails

Evaluation

GPT-4's response to JO's question in this section shows that it not only has knowledge of the archival process, but how the IDEF0 method can be used to define a context diagram for the Archive Digital Records function.

In any case, it is concluded that ChatGTP+ has the knowledge of the function Archive Digital Records and the CT knowledge to investigate the function Archive Digital Records and thus to think computationally about a complex system as a whole.

Summary of the ChatGPT's Knowledge of a Digital Archival Practice and CT Systems Thinking Ability

| Digital Archival Practice | GPT-4 Know How | CT Systems Thinking Practice |
|---|----------------|---|
| IDEF0 Context Diagram of the function Archive Digital Records | ✓ | Investigating a Complex System as a Whole |
| Subfunctions of Archive Digital Records | ✓ | Understanding the Relationships Within a System |
| Subfunctions of Record Storage and Preservation | ✓ | Thinking in Levels |
| Graphical Representation of the IDEF Model of Archive Digital Records | ✓ | Communicating Information about a System |
| Boundaries of the Archive Digital Records System | ✓ | Defining Systems And Managing Complexity |

What's Next? Research Questions

- Can GPT-4 with plugins explain and perform modeling and simulation in support of digital archival practices?
- Can GPT-4 with plugins explain and solve digital archival problems using computational problem solving practices?
- Can a learning environment in which a conversational agent similar to GPT-4 be constructed in which Archival Studies students collaborate with the agent to learn computational methods and tools for performing digital archival tasks?

Read the Paper for Details

Try out the Technology

- OpenAI url
- Try Microsoft CoPilot in Bing
- Try Google Bard