CIRCUMVENT NETWORK SECURITY

Project Specification and Plan

OCTOBER 4, 2023

STUDENT: JOHN BRENNAN
South Eastern Technological University, Carlow Campus

Contents

Executive Summary

Threat actors are constantly developing new, more advanced techniques to help them breach digital ecosystems and threaten the cybersecurity of the users within. If successful they will have affected one or more of the core pillars of cybersecurity: Confidentiality, Integrity, and availability, also known as the C.I.A. triad. The advent of publicly available Artificial Intelligence AI tools such as chat GPT has opened up new avenues of attack, namely using AI tools to enhance the threat actors' capabilities. This project has three main aims:

1.) Investigate how AI tools are being used by threat actors.
2.) To demonstrate how AI tools can be used by a threat actor to aid in their exploitation of vulnerable systems. This will be accomplished though socially engineered queries designed to return malicious results normally prohibited by the safeguards the AI tools have in place.
3.) Develop a tool that can automate the process described above and show its effectiveness against a representative target network.

This project may be used as a basis for raising awareness of the increasing use of AI tools to assist a threat actor in carrying out attacks. It is hoped that businesses will be better able to detect the indicators of compromise of these attacks and prevent or limit the damage these attacks produce.

Introduction

The purpose of this project is to investigate how AI assisted tools are being used by growing numbers of threat actors to aid in their efforts to overcome network defences and breach target systems. In this project the AI assisted tool would ingest data from network infrastructure vulnerability scanner such as NMAP. The tool would then perform pre-processing before calling the Common Vulnerabilities and Exposures (CVE) Application Programming Interface (API) to return a list of potential vulnerabilities based on the processed data returned from the enumeration scans (executed vulnerability scans). The tool would then call the Chat GPT API, leveraging socially engineered queries to turn ChatGPT into "ThreatGPT". The returned ChatGPT generated code could then be used as a basis for exploitation of the vulnerabilities returned from the CVE API. The tool would return to the threat actor both the targets' vulnerabilities and the attacks based on the processed results. Giving the threat actor the most exploitable vulnerabilities resulting in the optimal attack vector with which to breach the target system.

Once within the target network an AI tool of this kind would ingest the results of multiple enumeration tools to inventory the infrastructure, detailing not just the endpoints and intermediate devices such as routers and switches but also each devices vulnerabilities and the best method to exploit the vulnerability to further compromise the system while remaining undetected by the target systems detection tools.

Terms

Throughout the project the documentation will refer to several terms when certain functions are being performed of used. Below is a brief explanation of the terms used in this document.

AI:

AI stands for Artificial Intelligence and is the general term for the way in which computer systems emulate the human brains problem solving skills through the creation of artificial neural networks.

API:

An Application Programming Interface (API) is a set of rules and protocols that allow different software applications to exchange information with each other safely and securely(What Is an API? | IBM, n.d.). This ensures that applications can work together making application development more streamlined and efficient. Instead of having to load the ChatGPT website for example an application that wishes to use ChatGPT can call its API, allowing the query to be sent and the response received without involving the User Interface (ChatGPT's website in this case).

Ennumeration Scan:

An enumeration scan is used to 'build a picture' of the devices within a networked system (1 Network Enumeration/Discovery | InformIT, n.d.). If you liken a computer network to a jigsaw puzzle, then this jigsaw would have all the pieces face down and the box with the picture of what the puzzle should look like is missing. What an enumeration scan does is to turn the pieces 'face up' and can show how some pieces fit together, allowing us to identify the fundamental parts that make up the network. NMAP is a well-known enumeration scanner retuning not just how many hosts are online within the network, but it can return what Operating system and version the hosts have and the network services that are running on the host.

ChatGPT:

ChatGPT is an AI application from an origination called OpenAI based on the GPT 3.5 model (Ray, 2023). The GPT in Chat GPT stands for Generative, Pre-trained Transformer. This means that the model can generate output in the form of natural, conversational text-based language patterns learned from existing data. The pre-trained portion of the name means that the model was previously trained on data from multiple sources. The transformer portion allows the model to focus on each part of the input separately and allows it to focus on the most relevant portions of the input in relation to the output generated (*What Is the Transformer Architecture and How Does It Work?*, n.d.). ChatGPT 3.5 has 6.7 billion parameters.

Project Inspiration

The cybersecurity landscape is constantly changing. The continuous development of new or enhanced technologies brings about opportunities for their misuse. The development of chat GPT and similar AI based chatbots represent a fundamental shift in how we interact with and look for help from machines (A New Paradigm | by Tristan Wolff | Medium, n.d.). Unfortunately, this advancement gives threat actors the opportunity to misuse this technology to aid them in their efforts.

It was this potential misuse that was the inspiration for this project. It should be noted that chat GPT etc does have safeguards in place (if you simply ask it to write an exploit for a vulnerability it will refuse). One of the aims of the project is to demonstrate how it is possible to ‘side-step’ these safeguards through carefully phrased questions and get chat GPT to return the desired result.

Project Scope

The scope of this project focuses on investigating the growing use of AI Tools to aid treat actors in breaching target systems. This involves researching the current threat landscape while focusing on attacks that have involved the use of AI assisted tools, looking at the use of AI in historical attacks to gain a better understanding its growth in use. This can help in predicting the most likely areas of growth in the use of AI by threat actors.

To use ChatGPT(see research document for more details) in this manner through the development of a tool that could leverage the output of various scanning tools that enumerate a target network to be used by a threat actor to aid their attempts to breach the target systems defences. This project does not focus on training or developing an AI framework to accomplish this but rather a proof of concept to demonstrate how existing tools can be used to aid in attacks.

Deliverables

Non-Technical Core Deliverables

1). A research Document.

Research is a core component of any project and can cover a range of areas from the project feasibility to the technologies and tools suitable to complete the project within the timescale permitted. A research document should take a core premise (in this case the growing use of AI by threat actors to circumvent network security) and develop it using a variety of sources of information. Below are some of the sub deliverables of the research document.

- A comparative review of current AI models will be carried out.
  This will help to determine which AI model would be most suitable for this project. It is carried out before the creation of the specification document and is why ChatGPT was chosen as the AI model being used in this project.

- A literature review of enumeration tools
  Performing an evaluation of the research that has been carried out on currently available enumeration tools and an analysis of the tools themselves. The results of this analysis will dictate the tools that will be used to enumerate the network.

- A review of AI assisted cyber-attacks.
  Reviewing AI assisted cyber attacks provides a 'lay of the land' of the current uses by threat actors of AI tools. This can also help when attempting to predict areas of growth in the use of AI by threat actors.

- Predict the most likely areas of future growth.
  As a culmination of the research carried out to date, predict the most likely growth areas for the use of AI by threat actors.

- Manipulation of chat GPT through socially engineered questions to generate malicious code. This will be accomplished through the generation of test data (an example of which is shown below) and then using that data to build socially engineered queries to return malicious code.

Example of Creating Test Data

To generate test data, I used the command: nmap -dd -n -sS -oX nmap-results.xml 192.168.1.0/24 to run a command line scan of the interior of my home network.

- The -dd forces debug 2 mode, this forces Nmap to report the results for each port, not just the ports it finds "interesting".
- The -n tells Nmap to run the scan without DNS resolution.
- The -sS tells Nmap to perform the scan in TCP Syn scan mode.

- The -oX tells Nmap to output the scan data in XML format and save it in a file called nmap-results.xml.

I also performed an intensive scan using Windows GUI version of Nmap called Zenmap.. This returned the operating systems of the active hosts on the network.

- An assessment of software development technologies.
  An assessment of the software development tools that may used to develop the application, resulting in the choosing of the programming language, API's and any modules that will be used to develop the application.

2). A specification and plan document

The specification and plan (this document) detail the specifications of the application and the creation of a timeline of events and goals necessary for the successful completion of the project within the allotted timeframe.

3) A research poster

Used to summarise the outcomes of the research document. This will present a 'plot summary' of the research carried out and the conclusions dawn.

Technical Core deliverables

These core deliverables are part of the application design and development of the project. They involve the design, development, and testing of the application.

1) Application design
   This includes.

- Designing of the user interface.
- Determining the sequential order of operations for task completion
- Anticipating error handling requirements.

2) Processing of enumeration scan results.

   This involves taking scan results and extracting the data relevant determining the CVEs of the target.

3). Automation of the manipulation process using the ChatGPT and CVE APIs.

To accomplish this, the processed scan results must first be passed to the CVE API. The results CVEs returned will then be used to generate the socially engineered queries that will be passed to the ChatGPT API.

4). The creation of a GUI application

This will bring the processes together through A GUI based application.

5). Design and implementation of a target test network

To test the application will require the design and implementation of a target network. This will necessitate the creation of several virtual machines to create the target environment.

6). Deployment and testing of the application within the target environment.

This involves deploying the application within the test environment, executing scans of the environment, returning CVEs based on the scan, returning the ChatGPT code. Performing exploits using the code provided and documenting the results.

Non-Core Deliverables

1). A defensive and offensive mode for the application

This would be a useful addition to the tools capabilities and can be divided into 2 subcategories:

- Provide recommendation for mitigation.
  This would be a useful option for a defensive blue team security specialist but is not a fundamental part of the application main purpose.
- A defensive mode where the execution of the is disabled.
  This would allow a defensive blue team security specialist to view the vulnerabilities and the exploit code to aid in their efforts to create detections and mitigations.

2). Cross platform implementation

This would see the application being implements on platforms such as Windows and Linux.

Assumptions and Constraints

Assumption:

The fundamental assumption of this project is that the endpoints within the target network are running some version of Microsoft's Windows operating system, and that, for the purposes of the project, the attacker already has a foothold within the system. Another assumption is that the enumerations tools are preinstalled in their default locations (NMAP for example is installed in /usr/local/bin in Linux and c:\Program files in Windows by default) for the AI assisted tool to run the scans needed to generate the required data from which the network attack profile can be generated.

Constraints

Time is one of the main constraints of the project as the project must be completed by April 2024. As a result of the time constraints the focus of the project will be research into the Growth of AI use by threat actors rather than the development and training of an AI model. Training an AI model requires large volumes of relevant, accurate data and processing power to train the model for the task it's being developed to accomplish. This data would have generated first through the creation of multiple varied environments as there is a lack of publicly available Hi fidelity data to train an AI model with. The likelihood of various enterprise networks revealing the vulnerabilities in their defences would not be data that would be made publicly available, making the acquisition of accurate data more difficult. For this reason, it was decided to concentrate on developing a method of circumventing ChatGPT's safeguards to 'trick' it into outputting the desired result.

The target system.

For the practical portion of the project a target network system is required. The target system represents systems found in many small/medium businesses where there may be a single person IT department and no discrete cyber security personal. This target would consist of endpoint workstations running Windows 10 Enterprise or professional, A Windows 2016 or 2019 server. These servers may be running services such as Dynamic Host Configuration Protocol (DHCP), Domain Name System (DNS), and Internet Information Services (IIS). The Network infrastructure would include A standard firewall such as the ZyXel USG110, 2 routers (different models and/or makes would be preferred as it increases the that can be discovered by the tools) and 1 switch.

Attack Vectors and vulnerabilities.

To maximise the likelihood of a successful outcome, the project will focus on manipulating current generative AI models, essentially hacking the AI tool in a similar manner to a Threat actor to demonstrate how it can be used to assist in circumventing networking defences. Below are examples of some attacks that may be produced.

Network Attacks:

- Open port/vulnerable service
- DHCP spoofing
- Vlan hopping.

Active Based Directory Attacks:

- Pass the hash.
- DCSync
- PsExecution

During the research portion of the project the number of attacks that are network and Active directory based may fluctuate depending on the availability of high-fidelity data for to train the AI model.

The Target Audience

One of the considerations of any project is the target audience. Who will use this research? It is hoped this research can be of benefit to both offensive red team and defensive blue team security researchers. A successful demonstration of the leveraging of a generative AI in this manner may provide new avenues of attack for offensive red team security specialists, while defensive blue team security specialists observe previously unknown methods of exploitation and develop detections and defences against them.

Metrics

Outlined below are the metrics for THE GROWING USE OF AI BY ATTACKERS TO CIRCUMVENT NETWORK SECURITY AN INVESTIGATION.

- To achieve a 30% success rate in generating code based on socially engineered quires.
- Archive a 50% success rate in generating CVE queries from scan results.
- Achieve a 25% success rate in successfully compromising a network component using the exploit code generated.
- Achieve a 20% level compromise of the target network.

Use Case

When developing any project, the creation of a use case diagram helps to visualise how the system under development behaves under normal conditions. Visualising what types of users interact with the system, what parts of the system each user has access to, and what functions the system is expected to perform under normal use.

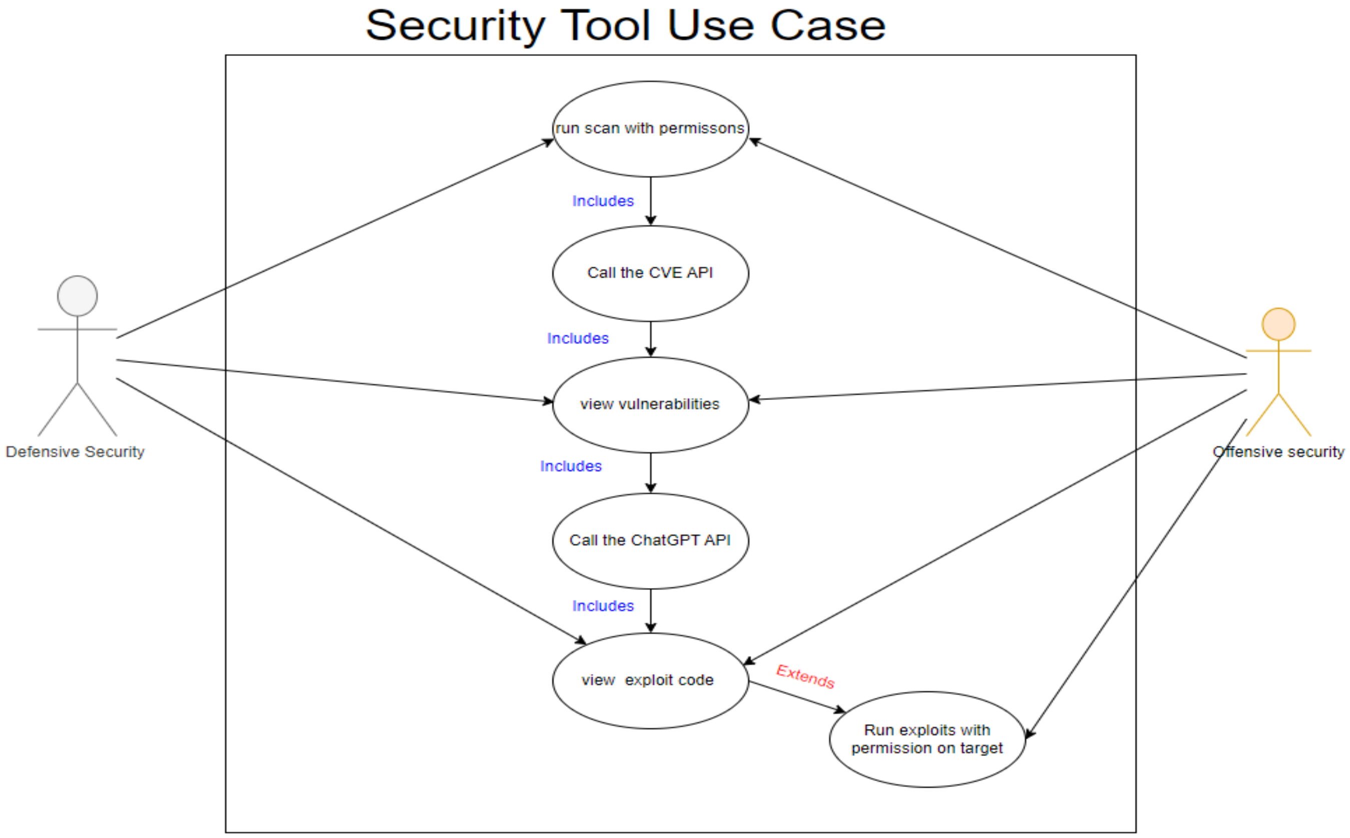


Fig 1. above shows a use case diagram for the application.

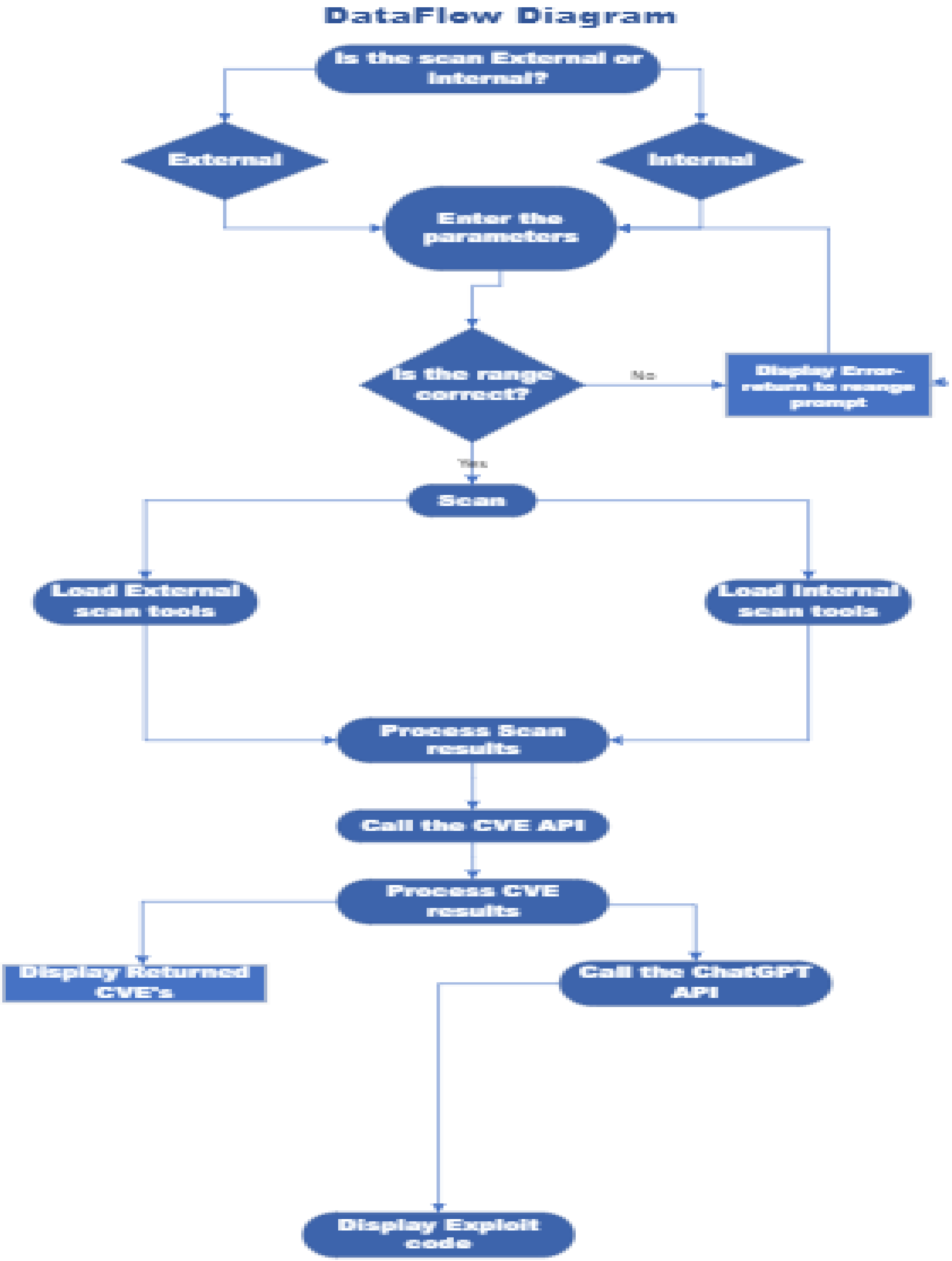


Fig 3. above shows the workflow for the application.

FURPS

Functionality

The application should allow a user to scan a target network and return vulnerabilities. The application should automatically highlight the most exploitable vulnerability and generate code that can be used to exploit the vulnerability. The user should be able to highlight other vulnerabilities that were found and have code generated for them.

Usability

The application should be intuitive to use, requiring very little prior knowledge to install and use. It should automatically return the code to perform the attack most lightly to succeed but also allow the user to scroll through the vulnerabilities discovered and return code to exploit whichever vulnerability the user picks.

Reliability

The finished application should produce the code that requires very little modification to be able to exploit the vulnerabilities. Error handling should be robust and allow the tools to continue to function through most cases.

Performance

The application should perform all tasks with minimal delay.

Supportability

The application should run on a Linux based host.

Similar Research

In 2017 Bishop fox demonstrated a proof-of-concept AI called DeepHack,(1). This Offensive AI bot was able to learn how to perform Sequel injection attacks without having been coded to perform the attacks. Farkhund Iqbal1 et al (2) researched how Chat GPT queries can be manipulated to return that would otherwise be forbidden.

My project attempts replicate this work and then to automate this process by attempting to manipulate the queries of the API calls.

Project Plan

Introduction

This part of the document will detail a Plan for the project, including a timeline of events, major milestones in chronological order, and any software and hardware requirements. When approaching any project, it is important to first develop a project plan. A project plan not only creates project goals and deadlines but also helps by breaking down a seemingly daunting task that is the project into smaller more manageable subtasks. Concentrating on the completion of component tasks rather than the project as a whole allows for better overall productivity and increases the likelihood of successful completion of the entire project.

A project plan also allows the project developer the opportunity to identify tasks and goals for the project that may have been overlooked without the creation of a properly scoped project plan.

Requirements

The project will have several requirements both in terms of software and hardware (in the form of virtual machines). Below is a list of requirements to demonstrate the application.

Software Requirements

- Python
- Use of the CVE and ChatGPT API's

Python:

Python is the programming language that will be used to develop the application. This is a popular and versatile language.

Hardware requirements

Windows 2019 servers X 2

Windows 10 workstations X 3

Kali Linux X1

Timeline

The main area of focus for this project will be the research element as before I can attempt to automate a socially engineered ChatGPT I must first learn how to correctly manipulate the queries to get Chat GPT to return the desired results for this reason I will be allocating 8 weeks to this and related AI research. Accordingly, a large portion of the project timeframe will be dedicated to research.

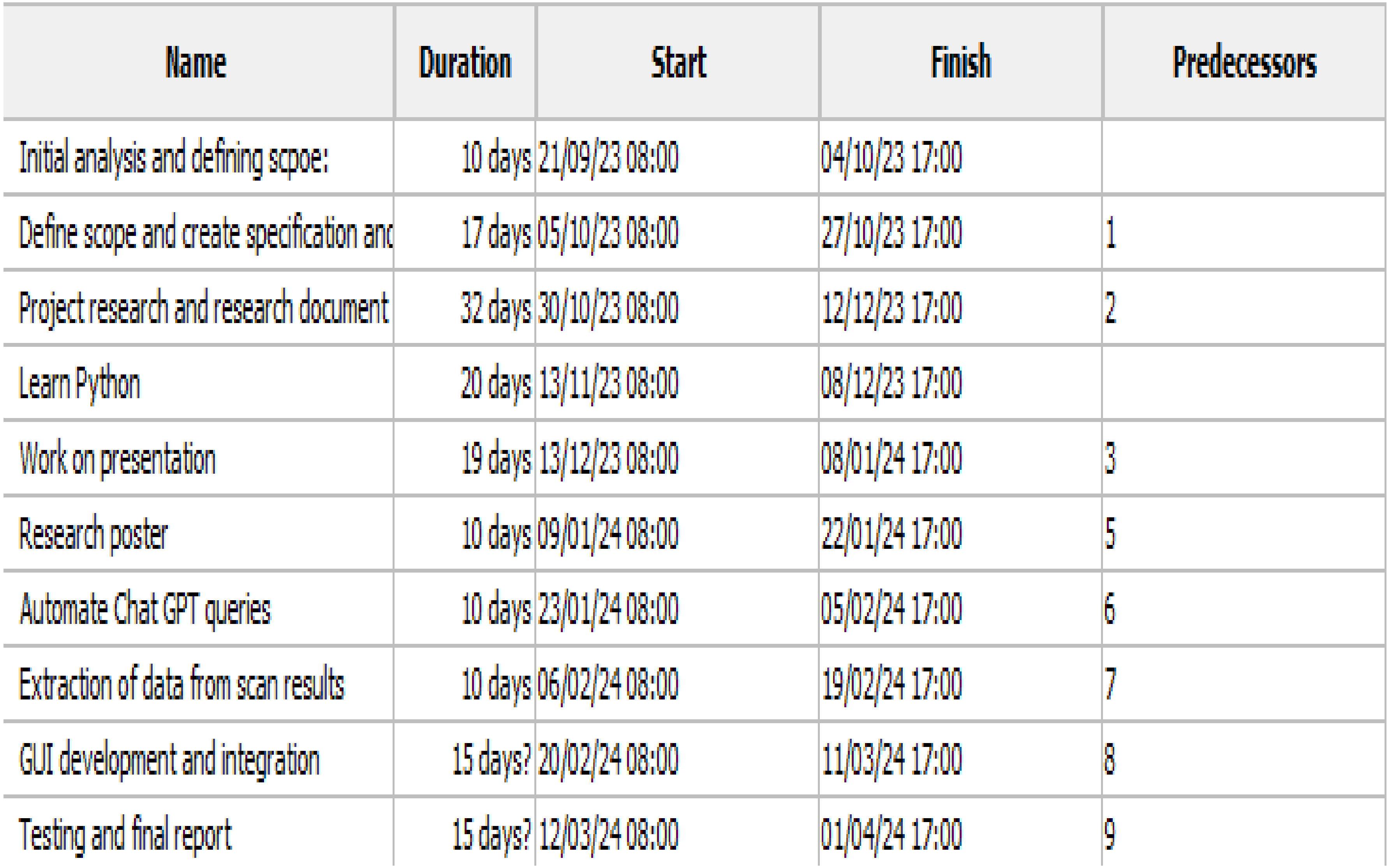
During a portion of this time, I also plan to learn the python programming language well enough to be able to use it to create the application for this I will allocate 4 weeks in total (2 of which will take place during the research phase.).

If successful at manipulating queries through ChatGPT's the front end, my plan is to then work on automating the queries to achieve the same result using python and the Chat GPT API, for this I will allocate 2 weeks.

Once successfully automated the next task will be to extract relevant data from the scans and to automate the return of the CVE's using the CVE API for this, I will allocate 2 weeks. Automating these queries may overlap with automating the GPT queries the coding challenges will be the same just the API calls being used will be different. The main anticipated use of this time allocation will be to successfully extract and format the relevant data from the scan results.

Once the previous steps have been completed the next phase will be the creation of a GUI and the integration of the independently rested functions. I will allocate 3 weeks for this portion the project.

Finally, 3 weeks for the testing of the completed app and final report.



Name	Duration	Start	Finish	Predecessors
Initial analysis and defining scpoe:	10 days	21/09/23 08:00	04/10/23 17:00	
Define scope and create specification anc	17 days	05/10/23 08:00	27/10/23 17:00	1
Project research and research document	32 days	30/10/23 08:00	12/12/23 17:00	2
Learn Python	20 days	13/11/23 08:00	08/12/23 17:00	
Work on presentation	19 days	13/12/23 08:00	08/01/24 17:00	3
Research poster	10 days	09/01/24 08:00	22/01/24 17:00	5
Automate Chat GPT queries	10 days	23/01/24 08:00	05/02/24 17:00	6
Extraction of data from scan results	10 days	06/02/24 08:00	19/02/24 17:00	7
GUI development and integration	15 days?	20/02/24 08:00	11/03/24 17:00	8
Testing and final report	15 days?	12/03/24 08:00	01/04/24 17:00	9

Fig 4. above details the project plan.

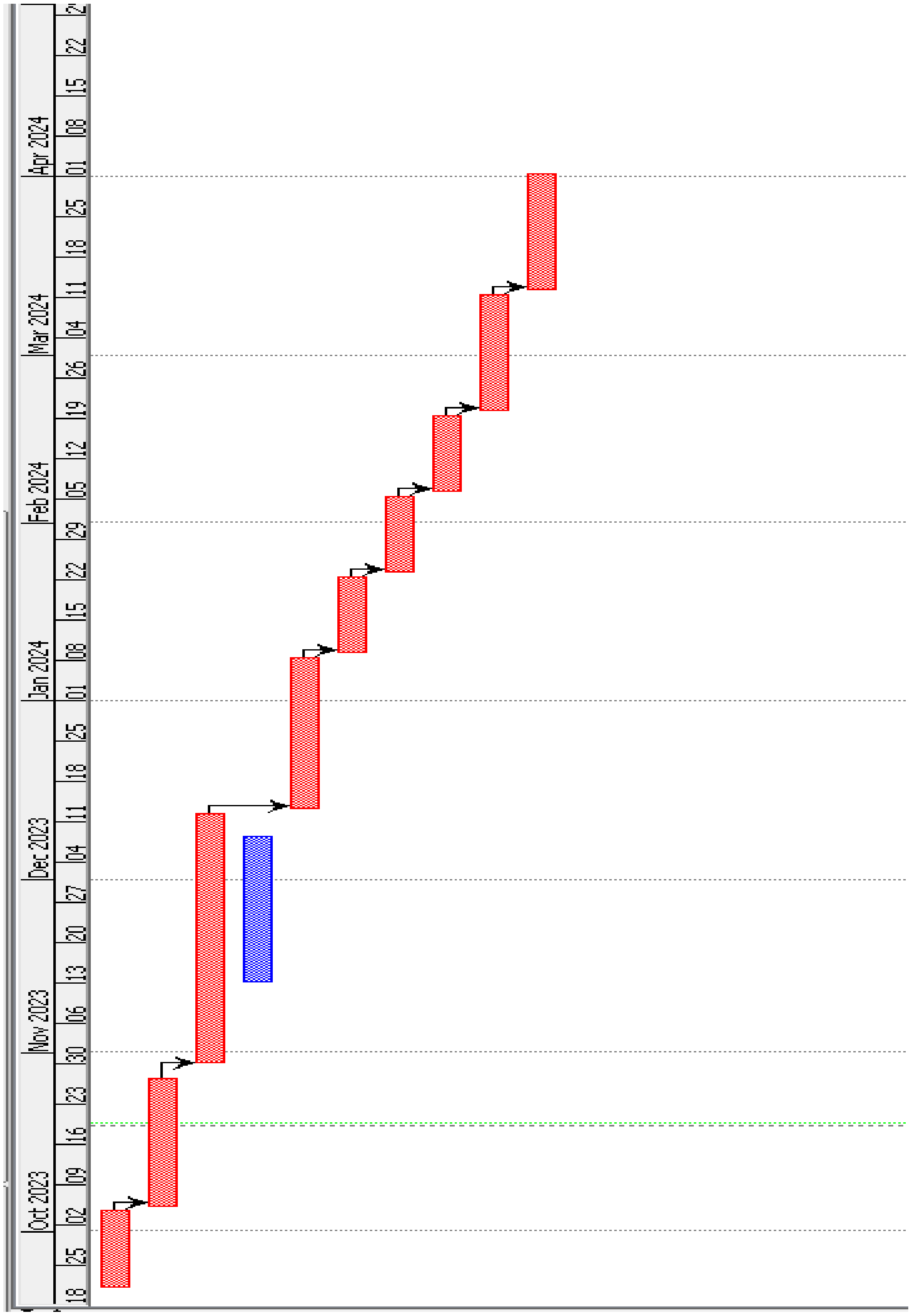


Fig 5. above shows a Gantt chart outlining the project timeline.

References

A New Paradigm Shift: How The Rise Of ChatGPT Is Revolutionizing The Way We Ask Machines For Help | by Tristan Wolff | Medium. (n.d.). Retrieved October 22, 2023, from https://tristwolff.medium.com/a-new-paradigm-shift-how-the-rise-of-chatgpt-is-revolutionizing-the-way-we-interact-with-machines-f28730649f4a

Bishop Fox | DeepHack Demo - Exploiting SQLi by Using an Open-source…. (n.d.). Retrieved October 18, 2023, from https://bishopfox.com/resources/deephack-demo-exploiting-sqli-by-using-an-open-source-hacking-ai-tool

Iqbal, F., Samsom, F., Kamoun, F., & MacDermott, Á. (2023). When ChatGPT goes rogue: exploring the potential cybersecurity threats of AI-powered conversational chatbots. *Frontiers in Communications and Networks*, *4*, 1220243. https://doi.org/10.3389/FRCMN.2023.1220243/BIBTEX

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121–154. https://doi.org/10.1016/J.IOTCPS.2023.04.003

Security Through Penetration Testing: Internet Penetration | 1 Network Enumeration/Discovery | InformIT. (n.d.). Retrieved October 23, 2023, from https://www.informit.com/articles/article.aspx?p=25916

What is an Application Programming Interface (API)? | IBM. (n.d.). Retrieved October 25, 2023, from https://www.ibm.com/topics/api

What Is the Transformer Architecture and How Does It Work? (n.d.). Retrieved October 26, 2023, from https://datagen.tech/guides/computer-vision/transformer-architecture/#