Individual Project (CS3IP16)

Department of Computer Science
University of Reading

Project Initiation Document

PID Sign-Off

Student No.	**30021591**
Student Name	**Ruben Lopes**
Email	**Il021591@reading.ac.uk**
Degree programme (BSc CS/BSc CSwIY)	**BSc Computer Science**
Supervisor Name *(Consultation with supervisor is mandatory)*	**Dr . Xiaomin Chen**
	Supervisor to sign PID form on Bb (grade centre)
Date	**17/10/2024**

SECTION 1 – General Information

Project Identification

1sl	**Project Title**
	Exploring the Use of Large Language Models to Perform Automated Cyber Attacks
1.2	**Please describe the project with key-phrases (max 5)**
	- How LLMs could be used to Automated Cyber Attacks. - Comparison with Traditional automated Hacking tools. - Building a Fine-tuning Large Language Models for offensive purposes. - Testing this in a Closed Network system and using - Ethical Considerations and Security Mitigation Techniques
	E-logbook maintenance agreed with supervisor *Use Google doc, OneDrive, or any mobile App whereby you will be able to generate a PDF copy*
	https://rubenlopes.uk/fyp/
1.4	**GitLab link for maintain source code and research data** *Any change in GitLab link and Source code repository MUST be explicitly mention in final report*
	https://csgitlab.reading.ac.uk/il021591/fyp https://github.com/ru4en/fyp/ (Mirror)

SECTION 2 – Project Description

2.1	**Summarise the project's background in terms of research field /application domain (max 100 words).**
	The recent rapid progression of large language models has raised growing concerns about their potential impact on privacy and security. This project aims to explore how LLMs could be used to automate cyber-attacks, offering an alternative to traditional automated tools, which typically rely on predefined rule-based scripts to exploit vulnerabilities. The project challenges the current consensus by fine-tuning an LLM specifically for offensive purposes. The project will also explore other advanced techniques, such as using LLMs for vulnerability discovery and exploitation.

2.2 **Summarise the project aims, objectives and outputs (max 250 words).**
These aims, objectives, and outputs should appear as the tasks, milestones and deliverables in your project plan (fill out Section 3).

The project aims to provide a deeper understanding of the impact of using Large Language Models to Perform Automated Cyber Attacks. It will try to provide a demonstration on how LLMs could be fine-tuned for offensive security purposes, going beyond traditional rule-based hacking tools. The project will also explore the ethical considerations and security implications of using AI-driven models for cyber-attacks.

Objectives:

- Develop a fine-tuned LLM capable of performing cyber-attacks using two key tools: a Linux shell and a browser.
- Successfully carry out automated attacks on a target machine within a controlled virtual environment (closed network system).
- Compare the performance of LLM-driven attacks with traditional automated hacking tools.
- Provide insights into the ethical implications and potential security mitigation techniques required to defend against LLM-based attacks.

Deliverables:

- **Theoretical Document:** A comprehensive report titled “Exploring the Use of Large Language Models to Perform Automated Cyber Attacks”. This document will discuss the potential impact of LLMs on offensive security, compare them with existing rule-based scripts, explain the data used, model training methods, and address the ethics of AI-driven cyber-attacks.
- **Fine-tuned Model and Application Code:** The actual LLM fine-tuned for cyber-attacks, along with the codebase to reproduce and run the model.
- **Data Used:** Dataset for training and testing the LLM, including any modifications made.
- **Demo Video:** A video demonstration showcasing the model executing attacks on a target machine. (terminal output, network traffic, impact on the target machine)

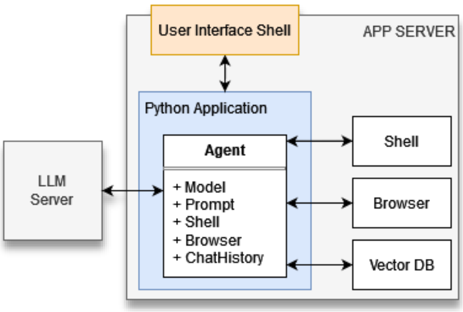
By the end of the project, we should have a fully trained LLM capable of executing cyber-attacks, along with theoretical and practical insights into its real-world applications and risks.

2.3 **Initial project specification – roughly indicate key features and functions of your finished program/application. Indicate possible method, data source, technology etc. (max 400 words)**
(Sensible and relevant Charts, Table, and Figures can be used)

Application

The Application with consist of:

A shell Interface where the user can monitor and converse and interact with the agent.



The 2 tools that the agent will have access to is a Shell and a browser. The shell is where the commands that is required to execute the attack will be executed. For safety reasons the shell might be a container with host network access. The browser will be used in case the Agent requires updated documentation on specific tools.

Database

A Database will also be used to store history of previous attacks, conversation or methods. Using the conversation and methods/steps from previous attacks could potential used to finetune the agents since the quality of the data is very high.

LLM, Agent and Data

The finetuned LLM model will be running on a separate server which is expected to be using Llama 3.2 as its base model. This will be fine-tuned to perform a specific set of cyber-attacks, using shell commands for tasks like network reconnaissance. The Data that will be used to finetune the model will be a mixture of dataset about using a shell, kali tools and a high-quality dataset to learn hacking.

Possible Considerations and Changes:

Environment and Infrastructure

Currently, the plan is to run all application components in Docker containers on the same virtual machine (VM), apart from the LLM server. This setup aims to enhance security by protecting the host system from arbitrary commands that could potentially disrupt its functionality.

Database Selection

The two options that is currently being considered are a Traditional Relational Database, which is suitable for structured data, and a Vector Database, which is ideal for managing high-dimensional data, particularly for applications that require similarity searches and machine learning integration.

Data Set and Training

Another crucial decision involves exploring potential datasets for fine-tuning the LLM. This includes researching for high-quality dataset that has step by step commands to carry out an attack (with commands), Linux shell command usage, and documentation on relevant tools to ensure the agent is well-equipped for its intended tasks.

2.4 **Describe the social, legal and ethical issues that apply to your project. Does your project require ethical approval? (If your project requires a questionnaire/interview for conducting research and/or collecting data, you will need to apply for an ethical approval)**

YES, the project requires ethical approval due to its focus on potentially sensitive topics related to cyber-attacks. The application has the capacity to facilitate cyber-attacks, which could lead to negative consequences if misused. Therefore, there is a responsibility to ensure that it is used strictly for ethical purposes, such as penetration testing and cybersecurity education. To mitigate any potential risks, the source code has been made private for the time being.

Additionally, the project must comply with all relevant laws and regulations concerning cybersecurity, data protection, and privacy. This includes ensuring that the application does not enable illegal activities or violate any laws related to hacking and data breaches. To uphold these standards, all activities conducted using this application will be performed within a closed network system.

2.5 **Identify the items you may need to purchase for your project. A cost upto £200 can be applied (include VAT and shipping if known). You need to have consent of your supervisor. Your request will be assessed by the department.**

- Up to four virtual machines and two virtual networks for demonstration and testing purposes.
- Specifically, one VM with GPU capabilities (8 GB VRAM Recommended for Llama 3.2) will be essential for hosting, inferring, and training.
- Optionally, if a VM with GPU capabilities is not available, OpenAI/Groq credits may be required for training and inference. Additionally, if knowledge distillation is required to improve the model

For this project, I will utilise VirtualBox (Locally) or Proxmox to host the VMs and VNs, while the LLM server can be run on the university's existing hardware resources.

2.6 **State whether you need access to specific resources within the department or the University e.g. special devices and workshop**

I will need permission to use departmental servers or VM with GPU to effectively carry out the training and inference processes.

SECTION 3 - Project Plan

Please provide your project plan.

Below is an example project plan, you can use any tool or software to generate yours.

Project stage	**START DATE: 15/10/2024** **Project Weeks**												
	0-3	3-6	6-9	9-12	12-15	15-18	18-21	21-24	24-27	27-30	30-33	33-36	36-39
1 Research and Planning													
Theorical Planning	🟦												
Technical Planning	🟦	🟦											
Finding Dataset for LLM		🟦	🟦										
2 Analysis and Design			🟦										
Design Model and Application Architecture			🟦	🟦									
Design Controlled Virtual Environment				🟦									
3 Develop and prototyping													
Build Interface and application		🟥	🟥	🟥	🟥								
Train Finetuned Model				🟥	🟥	🟥	🟥						
Refine and iterate									🟥	🟥	🟥		
4 Testing/evaluation/validation													
Design Test and Benchmarks		🟧	🟧					🟧					
Testing in Controlled Environment						🟧	🟧	🟧	🟧				
Validation of Ethical Implications										🟧			
5 Assessments													
Theoretical Writeup		🟩	🟩	🟩	🟩								
Performance Assessment						🟩	🟩	🟩	🟩	🟩			
Ethical and Security Impact								🟩	🟩	🟩	🟩		
Final Deliverables											🟩	🟩	🟩

🟦 **PLANING AND DESGINING** 🟥 **DEVOLOPING/TRANING** 🟧 **TESTING** 🟩 **DOCUMENTATION**