COMP7705 Project

Detailed Project Proposal

Project Title:	Cyber Sentinel: Cyber Security Intelligence through Localized Cyber-Security focused Large Language Model with RAG
Mentor:	Dr. Vivien PS Chan
Student 1 (Leader)	CHOW Sau Ho
Student 2	CHEUNG Sin Shing

Aim

Introduction

ChatGPT, Claude-3, Gemmi, the revolutionary Large Language Models (LLM) services have become omnipresent and major components in the daily life of the public. The power of LLMs not only limited to be capable of understanding and generating human-like text, but also being rapidly adopted across numerous industries, such as in Health Care, Content Creator, Media etc, which unlocking new possibilities, revolutionizing the way we approach tasks and problems and bringing convenient and efficient to the public. However, when it comes to the cybersecurity field, the benefit of LLM may be hindered by the nature of cybersecurity field and the limitation of the existing LLM applications.

Problem Statement

The current issues of the LLMs to the cybersecurity fields are mainly related to the Confidentiality and Privacy Concerns, Outdated Knowledge and Requirements of Domain Specific Knowledge.

1. Confidentiality and Privacy Concerns

The cybersecurity domain heavily relies on confidential and sensitive data, especially during the Incident Response phase. Leaking such information can have catastrophic consequences, exposing internal vulnerabilities, compromising the organization's reputation, and eroding public trust. However, when cybersecurity experts try to use the existing online LLM for analysis, the necessary data traffic will introduce the potential risk of data leakage, particularly when confidential and sensitive information is included in the prompt. For example, the Samsung Data Leakage in ChatGPT incident is one of the examples showing the risk of sensitive data leakage in using LLMs service. While other industry can easily avoid including sensitive information in their queries, cybersecurity experts often require access to this confidential data to gain a comprehensive understanding of security incidents which possess challenges for using the LLM services.

2. Outdated Knowledge

Language models like ChatGPT have a knowledge cutoff, which means their training data is limited to a specific timeframe, such as ChatGPT's knowledge being restricted to information before January 2022. However, the cybersecurity domain requires up-to-date knowledge, especially when dealing with vulnerabilities and emerging threats.

For instance, zero-day vulnerabilities or vulnerabilities that remain unpatched immediately after their announcement can create disastrous opportunities for malicious attackers to gain unauthorized access to systems. The knowledge cutoff greatly hinders cybersecurity experts from catching up with the latest trend and techniques which may increase the risk of the system. The rapidly evolving nature of cybersecurity threats and the necessity for timely responses make the static knowledge of these language models a potential limitation in effectively addressing emerging cybersecurity challenges.

3. Requirement of Domain-Specific Knowledge

The majority of existing large language models (LLMs) are designed for general-purpose tasks and may lack in-depth knowledge specific to certain domains, such as cybersecurity. This limitation can hinder their effectiveness in addressing industry-specific challenges and grasping the nuances of cyber threats. Consider ChatGPT as an example: ChatGPT could provide basic guidelines and fundamental knowledge of Wireshark (Network Analysis Tool), but it did not learn sufficient knowledge on the attack patterns or suspicious patterns for network analysis that could be found when using Wireshark which the user may not get the forensic artifacts for investigation.

Objective

The project aims to address the aforementioned challenges in order to unlock the promising potential of LLMs in cyber security field. Due to the complex and comprehensive nature of cybersecurity knowledge, experts in this domain often face significant challenges in terms of human effort and time-consuming tasks required to stay up to date with the latest developments. They need to engage in tedious work and read through an enormous number of papers to acquire the most recent cybersecurity knowledge. With the power of LLMs, cybersecurity experts can efficiently keep pace with the latest and comprehensive cybersecurity knowledge and their burden could be alleviated with this tool.

The goals of this project are:
(i) Develop a localized Large Language Model (LLM) specifically tailored for cybersecurity, with domain-specific knowledge and expertise.

(ii) Integrate a routine web scraper and the Retriever Augmented Generation (RAG) framework to ensure the LLM has access to up-to-date cybersecurity news and information.

(iii) Build a comprehensive knowledge base for the LLM by incorporating a wide range of cybersecurity data from academic papers, blogs, reports, and news sources.

Brief Literature Review

Artificial Intelligence (AI) has become a cornerstone technology in recent years, with its potential profoundly realized in the field of cybersecurity. AI's capacity to automate and enhance security measures against cyber threats is well-documented, with techniques such as Machine Learning (ML) and Deep Learning (DL) leading to significant advancements in threat identification and response(Sarker et al., 2021).

Natural Language Processing (NLP), a critical AI domain, plays a vital role in cybersecurity(Motlagh et al., 2024). By enabling machines to understand and interpret human language, NLP facilitates more effective monitoring and analysis of cyber threats communicated through text, enhancing the detection and prevention of attacks.

Large Language Models (LLMs), a subset of NLP, have transformed threat analysis in cybersecurity. Trained on vast datasets, LLMs can understand subtle linguistic cues, allowing them to generate and identify potential cyber threats amidst large data volumes (Tann et al., 2023). By integrating LLMs into their toolsets, cybersecurity experts can discern complex patterns in communications and establish advanced defences to preempt cyber adversaries. LLMs also streamline security operations, reducing the time and resources needed for threat identification and remediation.

However, employing online LLM services for sensitive data processing raises serious privacy concerns. Sharing data with external platforms for analysis exposes it to potential breaches. In Incident Response, where data sensitivity peaks, any unauthorized disclosure can have catastrophic consequences, revealing vulnerabilities and eroding trust (Gupta et al., 2023).

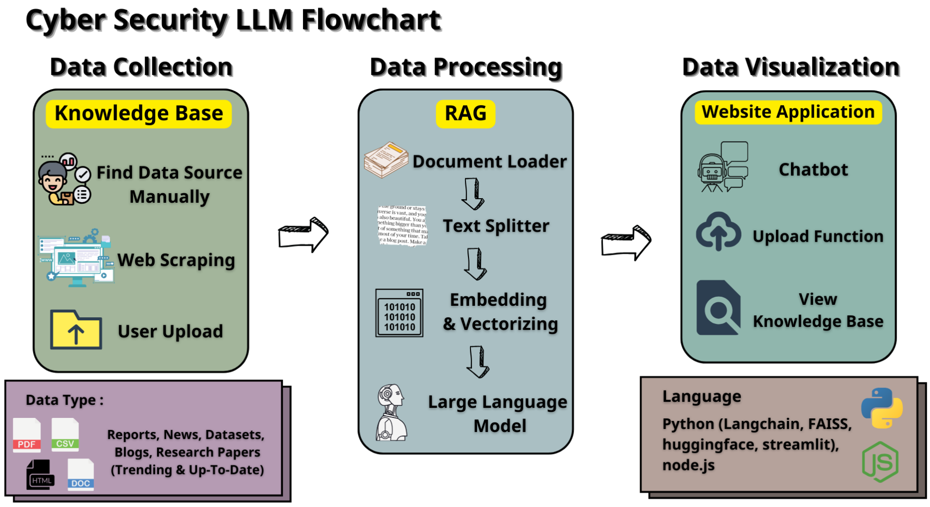
The incident involving the Samsung Data Leakage through ChatGPT serves as a stark reminder of these risks, illustrating how sensitive data can inadvertently be exposed through the use of LLM services (Gupta et al., 2023). While industries outside of cybersecurity may avoid this risk by excluding sensitive information from their LLM queries, cybersecurity professionals do not have the luxury of such omissions. Their need for a thorough understanding of security incidents necessitates direct interaction with confidential data, thus complicating their use of LLM services. This challenge underscores the need for a secure and private approach to employing LLMs in cybersecurity, ensuring that the advantages of AI do not come at the expense of critical data privacy.

Recognizing the importance of data protection, the cybersecurity industry must also confront the need for timely and up-to-date information. Cyber threats evolve swiftly, making it essential for the data used in defense strategies to be current. As Takahashi and colleagues (Takahashi et al., 2018) have highlighted, the dynamic nature of cyber threats requires a responsive approach to threat intelligence, where having access to the most recent information can determine whether a cyber attack is successfully repelled or not.

The Retriever Augmented Generation (RAG) framework presents a solution to this challenge by enabling Language Models (LMs) to retrieve the most current information from external sources in real-time. When combined with LMs, RAG serves as a conduit between the model's vast yet static knowledge base and the constantly updated data streams on the internet (Lewis et al., 2020). This integration allows LMs to generate responses that are up to date with the latest cybersecurity threats and countermeasures.

LOCALINTEL showcases the application of RAG in cybersecurity intelligence (Mitra et al., 2024). Contrasting with our project's focus on real-time data, LOCALINTEL leverages RAG to adapt global threat intelligence to an organization's unique environment, factoring in network architecture, hardware, and business objectives. This enables tailored threat intelligence that aligns with specific security needs. Our project, in comparison, emphasizes the immediacy of RAG, striving to deliver threat intelligence that is both contextually relevant and up to date, thus equipping organizations to counter emerging cyber threats effectively.

Proposed Methodology



The above figure shows the methodology we are going to implement for our project. Our project could be classified into THREE stages: Data Collection, Data Processing and Data Visualization.

The first stage is constructing the knowledge base as the reference for the localized LLM. The knowledge base will be constructed through (i) manual effort to pre-load some documents related to cybersecurity (ii) routine web scraper to extract up-to-date information (iii) user can upload their own sensitive information. As the RAG and LLM service will be totally running locally, users no longer need to worry about data leakage issues like those using online LLM services. This stage mainly uses Python (selenium and beautifulsoup) and node.js to do so.

The second stage will be the Retrieval-Augmented Generation stage which utilizes Python (Langchain, FAISS, huggingface embedding). During prepartion, the program will use different document loaders to extract texts from the source documents and split the texts into different chunks. Then, the program will embed and vectorize the chunked texts to vectors such as the FAISS vectors and store it locally. When the user types the prompt and queries our program, the program will first be doing a retriever search such as using the similarity search to compare the similarity of the keywords in prompt with those vectors. Then the program will return the top K relevant documents (k = pre-defined number) and pass the contents (vectors) of the top K relevant documents to the LLM as a retriever content for the LLM to refer on them. Then the LLM could provide a more accurate and domain specific answer based on the references extracted from the retriever.

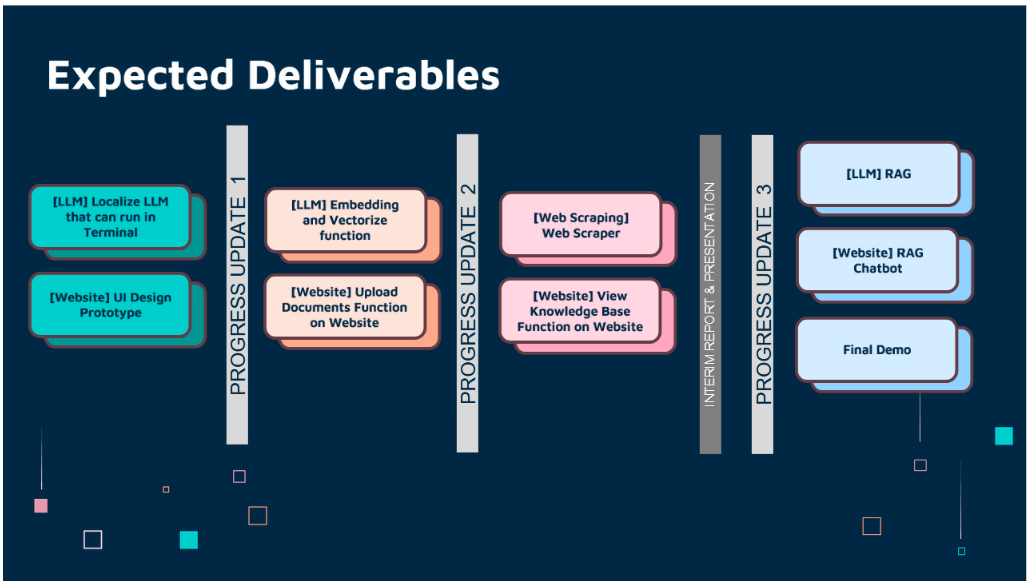
The reason why we choose to use RAG for improving the output of LLM instead of fine-tuning is due to the rapidly evolving nature of cybersecurity threats and the necessity for timely responses. Compared to fine-tuning, RAG could be much more cost-efficient in terms of time and computing power as it do not need to change the parameters originally set in the LLMs.

To visualize and be more user-friendly, we aim to build a website application mainly with chatbot function by streamlit/gradio in python. The user can simply type the prompts and ask their questions related to cybersecurity field and supported by the RAG framework with the comprehensive, up-to-date knowledge base as reference. Moreover, users can review the source documents on the knowledge base as well as uploading their own documents on the knowledge base through the web UI.

Milestones

	Tasks	*Estimated completion time*	*Estimated number of learning hours*
1	**[Preparation] Literature Review on LLM Application in Cyber Security, RAG**	**16 – 03 - 2024**	**30 x 2**
2	**[Preparation] Web Scraping Prepation Work: Explore reliable website source**	**23 – 03 – 2024**	**10 x 2**
3	**[LLM] Localize LLM**	**06 – 04 –2024**	**40 x 2**
4	**[Website] UI Design**	**06– 04 - 2024**	**10 x 2**
	Project Progress Update 1	**13 – 04 - 2024**	
5	**[LLM] Embedding & Vectorize**	**20 – 04 - 2024**	**20 x 2**
6	**[Website] Upload Documents**	**27 – 04 - 2024**	**10 x 2**
	Project Progress Update 2	**04 – 05 - 2024**	
7	**[Web Scraping] One Time Web Scraper**	**18 – 5 - 2024**	**40 x 2**
8	**[Web Scraping] Scheduler and Routine Web Scraper**	**25 – 5 - 2024**	**30 x 2**
9	**[Website] View Knowledge Base**	**28 – 5 -2024**	**5 x 2**
	Interim Report and Presentation **Project Progress Update 3**	**01-06-2024** **22-06-2024**	
10	**[LLM] RAG**	**01 – 07 – 2024 (Start at May)**	**65 x 2**
11	**[Website] RAG**	**06 – 07 – 2024 (Start at May)**	**40 x 2**
12	**Enhancement of application (e.g. Agents, Fine-Tune LLM, extra tools)**	**EXTRA**	**EXTRA**
	Project Progress Update 4	**04 – 05 - 2024**	
			Total: 300

Deliverables



Expected Deliverables

[LLM] Localize LLM that can run in Terminal
[Website] UI Design Prototype
PROGRESS UPDATE 1
[LLM] Embedding and Vectorize function
[Website] Upload Documents Function on Website
PROGRESS UPDATE 2
[Web Scraping] Web Scraper
[Website] View Knowledge Base Function on Website
INTERIM REPORT & PRESENTATION
PROGRESS UPDATE 3
[LLM] RAG
[Website] RAG Chatbot
Final Demo

Items	
1	[Project] Project Progress Web Page
2	[LLM] Localize LLM that can run in Terminal
3	[Website] UI Design Prototype
4	[LLM] Embedding and Vectorize function
5	[Website] Upload Documents Function on Website
6	[Web Scraping] Web Scraper
7	[Website] View Knowledge Base Function on Website
8	[Project] Project Interim Report
9	[LLM] RAG
10	[Website] RAG Chatbot
11	[Project] Final Demo
12	[Project] Project Final Report

Reference:

Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. In *IEEE Access* (Vol. 11, pp. 80218–80245). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2023.3300381

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. http://arxiv.org/abs/2005.11401

Mitra, S., Neupane, S., Chakraborty, T., Mittal, S., Piplai, A., Gaur, M., & Rahimi, S. (2024). *LOCALINTEL: Generating Organizational Threat Intelligence from Global and Local Cyber Knowledge*. http://arxiv.org/abs/2401.10036

Motlagh, F. N., Hajizadeh, M., Majd, M., Najafi, P., Cheng, F., & Meinel, C. (2024). *Large Language Models in Cybersecurity: State-of-the-Art*. http://arxiv.org/abs/2402.00891

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. In *SN Computer Science* (Vol. 2, Issue 3). Springer. https://doi.org/10.1007/s42979-021-00557-0

Takahashi, T., Panta, B., Kadobayashi, Y., & Nakao, K. (2018). Web of cybersecurity: Linking, locating, and discovering structured cybersecurity information. *International Journal of Communication Systems*, *31*(3). https://doi.org/10.1002/dac.3470

Tann, W., Liu, Y., Sim, J. H., Seah, C. M., & Chang, E.-C. (2023). *Using Large Language Models for Cybersecurity Capture-The-Flag Challenges and Certification Questions*. http://arxiv.org/abs/2308.10443