WCD Capstone Project Proposal 3.2 **Ammar Battah**

Vulnerability Analysis and reporting tool

1 . Overview

The cybersecurity domain is a constant race where experts are always building up and fortifying defenses against attackers and malicious actors. The cyber field is diverse with several subdomains that are exposed to vulnerabilities, such as, software, network, hardware, application, access control, data and human factors. Software vulnerabilities are particularly foundational, as they often underpin other types of vulnerabilities and play a critical role in the security of other subdomains. Therefore, it is crucial to diligently reinforce the software being built, and minimize their exposure to vulnerabilities to build a robust cybersecurity defense.

2. Context of Data

In the cybersecurity domain, vulnerabilities are the results of weaknesses in the aforementioned subdomains, which attackers can exploit to gain access, misuse, or damage the target. One of the most known methods to track such vulnerabilities is a database known as the Common Vulnerabilities and Exposures (CVE) system, which documents such vulnerabilities.

3. Data source background and Purpose

The need for a standardized approach to document and share vulnerabilities has given rise to the CVE identifiers. The CVE system is maintained by the MITRE Corporation, and provides details, assessment and references for publicly known vulnerabilities and exposures. Each entry contains an identifier, a description, and references to reports and advisories by vendors or third-parties. Furthermore, several entities, such as the National Vulnerability Database, provide further insight based on the CVE entries.

The main purpose of using CVE data is to facilitate a shared knowledge source for the different security tools and services. By standardizing the identification of vulnerabilities, which is the source of information, it will aid in smooth integration and communication between different security tools and services, leading to better detection, analysis, responses, and mitigation of security threats.

4. Data Source

https://nvd.nist.gov/vuln/search

Other secondary sources:
https://www.cvedetails.com/
https://cve.mitre.org/
https://www.rapid7.com/db/

5. Application

Data collected during the previous phase of the project, which focused on CVEs, will be the foundation for this use case with LLMs(Large Language Models). By leveraging this data, an LLM tailored to cybersecurity applications can be utilized to understand and analyze vulnerability information at a proficient level. Such a model would be able to provide insights and recommendations, detect potential threats, manage vulnerabilities, and respond to incidents.

We will leverage a Retrieval-Augmented Generation (RAG) powered pre-trained model to generate reports that experts can rely on. To be able to execute that, we use Qdrant as our vector database and langchain to integrate and build our application.