A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

Aakash Aundhkar

Charan Reddy Bobbala

Davon Carvalho

Susmith Reddy Poreddy

Baba Akhil Sarvasiddhy

Vishnu Vasireddy

DAEN 690

Project Report

Vulnerabilities of LLM’s across different industries

**About the Cover**

Professor Berlin is an instructor at the George Mason University College of Engineering and Computing, Volgenau School of Engineering, MS Data Analytics Engineering (DAEN) program. He began working with the DAEN program as an adjunct faculty member in 2012 and became a fulltime faculty member in 2016. He is a passionate contributor to the program and a devoted mentor to his students.

His passion for new value creation is built on over 50 years of professional experience – innovating and advocating for innovators applying leading-edge digital solutions to mission challenges. He has served with outstanding teams in various roles, including senior strategy executive, consultant, and mentor; applied information and systems technologist; collaborative leader; computer scientist, and public policy entrepreneur.

He serves as a strategy advisor and mentor to public and private sector innovators and entrepreneurs and as a public speaker (emerging challenges, innovation opportunities, and ethics). His core interests include public policy, high-performance computing, cyber, emerging big data, health informatics, and digital economy and governance challenges.

In addition to teaching and mentoring, Professor Berlin seeks new engagements with high-quality, core-value-centered innovation teams – collaborating to address societal and market challenges with cyber-physical and policy innovation. Specifically, sustainable solutions can be delivered at the intersection of innovative value creation, human aspiration, and strategic vision.

Professor Berlin is a graduate of the US Air Force Academy with a Bachelor of Science in Computer Science and Mathematics as well as a graduate of the University of Texas at Austin with a Master of Arts in Computer Science. He is also attended the USAF Air Command and Staff College for leadership and strategy training.

Contents

Table of Contents

[Section 1: Problem Definition 3](#_Toc165667753)

[1.1 Background 3](#_Toc165667754)

[1.2 Problem Space 5](#_Toc165667755)

[1.3 Research 6](#_Toc165667756)

[1.4 Solution Space 6](#_Toc165667757)

[1.5 Project Objectives 7](#_Toc165667758)

[1.6 Primary User Stories 8](#_Toc165667759)

[Section 2: Datasets 9](#_Toc165667760)

[2.1 Overview 9](#_Toc165667761)

[2.1.1 VERIS Database: 9](#_Toc165667762)

[2.1.2 CVE Database: 9](#_Toc165667763)

[2.1.3 MITRE ATLAS Database 10](#_Toc165667764)

[2.1.4 Census Dataset 11](#_Toc165667765)

[2.2 Field Descriptions 11](#_Toc165667766)

[2.2.1 VERIS Data: 11](#_Toc165667767)

[2.2.2 CVE Data: 12](#_Toc165667768)

[2.3 Data Context 13](#_Toc165667769)

[2.3.1 VERIS Data 14](#_Toc165667770)

[2.3.2 CVE Data 16](#_Toc165667771)

[2.4 Data Conditioning 16](#_Toc165667772)

[2.4.1 VERIS Data 16](#_Toc165667773)

[2.4.2 CVE Data 17](#_Toc165667774)

[2.5 Data Quality Assessment 18](#_Toc165667775)

[2.5.1 VERIS Data: 18](#_Toc165667776)

[2.5.2 CVE Data: 18](#_Toc165667777)

[2.6 Other Data Sources 18](#_Toc165667778)

[2.7 Storage Medium 19](#_Toc165667779)

[2.7.1 VERIS database: 19](#_Toc165667780)

[2.7.2 CVE database: 20](#_Toc165667781)

[Section 3: Algorithms & Analysis / ML Model Exploration & Selection 21](#_Toc165667782)

[3.1 Solution Approach 21](#_Toc165667783)

[3.1.1 Systems Architecture 21](#_Toc165667784)

[3.1.2 Systems Security 21](#_Toc165667785)

[3.1.3 Systems Data Flows 22](#_Toc165667786)

[3.1.4 Algorithms & Analysis 23](#_Toc165667787)

[3.2 Web Scrapping 25](#_Toc165667788)

[3.3 Risk Assessment 27](#_Toc165667789)

[3.3.1 Main Components of Risk Assessment: 29](#_Toc165667790)

[Section 4: Risk Assessment and Markov Chain 30](#_Toc165667791)

[4.1 Project Architecture 30](#_Toc165667792)

[4.1.1 LLM Attack Filtering 32](#_Toc165667793)

[4.1.2 Quantifying impact 32](#_Toc165667794)

[4.1.3 Risk Assessment Hospital Data Analysis 34](#_Toc165667795)

[4.2 Risk Analysis-Threat 35](#_Toc165667796)

[4.2.1 Mapping VERIS Attack Techniques with MITRE ATLAS Attack Techniques 35](#_Toc165667797)

[4.2.2 Categorizing LLM Attacks into different types 36](#_Toc165667798)

[4.2.3 Probabilities Threat Calculation 41](#_Toc165667799)

[4.3 Risk Analysis-Vulnerability (Markov Chain Analysis) 42](#_Toc165667800)

[4.3.1 Analyzing Sectors: Identifying Trends and Patterns 42](#_Toc165667801)

[4.4 Risk Analysis-Consequence 45](#_Toc165667802)

[4.5 Final Risk Assessment 46](#_Toc165667803)

[Section 5: Calculating Normalized Risk Factor on Different Sectors 47](#_Toc165667804)

[Section 6: Findings 54](#_Toc165667805)

[Section 7: Summary 54](#_Toc165667806)

[Section 8: Future Findings 55](#_Toc165667807)

[Table 1: Glossary Table 54](#_Toc165668854)

[Table 2: Sprint 1 Risks 56](#_Toc165668855)

[Table 3: Sprint 2 Risks 57](#_Toc165668856)

[Table 4: Sprint 3 Risks 57](#_Toc165668857)

[Table 5: Sprint 4 Risks 58](#_Toc165668858)

[Table 6: Sprint 5 Risks 59](#_Toc165668859)

[**Figure 1: Snapshot of CVE database** 10](#_Toc165671415)

[**Figure 2: Data conditioning** 16](#_Toc165671416)

[**Figure 3: The above figure shows no of attacks over the time** 25](#_Toc165671417)

[**Figure 4: Risk assessment matrix** 27](#_Toc165671418)

[**Figure 5: Project Architecture** 30](#_Toc165671419)

[**Figure 6: Sample of Count, percentage, and score calculation for health care sector** 32](#_Toc165671420)

[**Figure 7: Mapping of LLM Attacks** 35](#_Toc165671421)

[**Figure 8: Mapping of LLM Attacks** 36](#_Toc165671422)

[**Figure 9: Visual shows the threats across different sectors** 38](#_Toc165671423)

[**Figure 10: Visual shows the count of each attack in different sectors** 39](#_Toc165671424)

[**Figure 11: Threat Score Calculation** 40](#_Toc165671425)

[**Figure 12: Transition Matrix for Education Sector** 42](#_Toc165671426)

[**Figure 13: Transition Matrix for Information Sector** 42](#_Toc165671427)

[**Figure 14: Transition Matrix for Information Sector** 43](#_Toc165671428)

[**Figure 15: Transition Matrix for Professional Sector** 43](#_Toc165671429)

[**Figure 16: Transition Matrix for Retail Sector** 43](#_Toc165671430)

[**Figure 17: Transition Matrix for Transportation sector** 44](#_Toc165671431)

[**Figure 18: Consequence Calculation for LLMs** 45](#_Toc165671432)

[**Figure 19: Final Risk Assessment** 45](#_Toc165671433)

[**Figure 20: The above map indicates the Normalized Risk factor in the healthcare sector across US** 48](#_Toc165671434)

[**Figure 21: The above map indicates the Normalized Risk factor in the finance sector across US** 48](#_Toc165671435)

[**Figure 22: The above map indicates the Normalized Risk factor in the retail sector across US** 49](#_Toc165671436)

[**Figure 23: The above map indicates the Normalized Risk factor in the professional sector across US** 50](#_Toc165671437)

[**Figure 24: The above map indicates the Normalized Risk factor in the public sector across US** 50](#_Toc165671438)

[**Figure 25: The above map indicates the Normalized Risk factor in the manufacturing sector across US** 51](#_Toc165671439)

**This page intentionally left blank.**

Abstract

Abstract

The deployment of Large Language Models (LLMs) expands across industries, yet their security vulnerabilities emerge as critical challenges for adoption. The study aims to address the vulnerabilities of LLMs by integrating the MITRE Corporation's Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) attack patterns with the Verizon Vocabulary for Event Recording and Incident Sharing (VERIS) Community Database. Additionally, incorporating North American Industry Classification System (NAICS) codes from census data allows for a sector-specific risk assessment, enhancing the precision of our analysis. Mapping MITRE ATLAS attack techniques to VERIS database attacks and integrating NAICS codes, the study provides an industry-tailored analysis that underscores the specific vulnerabilities of LLMs. Employing Markov Chain modeling, the study delves into VERIS threat data to identify patterns and trends in LLM attacks, facilitating an approach to threat prediction and sector-specific vulnerability identification. This methodology not only aids in understanding the likelihood of attack transitions but also in pinpointing industries at increased risk of persistent threats. The centerpiece of the research is the development of a comprehensive risk assessment framework. This framework assesses the vulnerability of LLMs in different industries and the consequent impacts on businesses, such as financial losses and reputational damage. The study culminates in an innovative assessment tool designed for businesses to evaluate their LLM usage risks judiciously. This tool leverages sophisticated data extraction methods to amalgamate insights on cybersecurity risks, offering a granular view of industry specific LLM vulnerabilities. This enhanced framework and tool empowers companies with the knowledge to navigate the complex security landscape of LLMs, contributing a safer and more informed utilization of this transformative technology.

**This page intentionally left blank**

Report

# Problem Definition

## Background

A large language model (LLM) refers to a language model equipped with many parameters, which undergoes pre-training tasks like masked language modeling and autoregressive prediction. These tasks enable the LLM to comprehend and process human language by capturing contextualized text semantics and probabilities from vast amounts of text data. A proficient LLM should possess four essential features: (i) a deep understanding of natural language context, (ii) the capability to generate text resembling human-like language, (iii) contextual awareness, particularly in knowledge-intensive domains, and (iv) a strong ability to follow instructions, which is beneficial for problem-solving and decision-making processes [1], [2].

In 2023, several LLMs were developed and released, gaining significant popularity. Notable examples include OpenAI’s ChatGPT, Meta AI’s LLaMA, and Databricks’ Dolly 2.0. For instance, ChatGPT alone boasts a user base of over 180 million. These LLMs offer a wide array of versatile applications across various domains. They not only provide technical support to domains related to language processing, such as search engines, customer support, and translation but also find utility in more general scenarios like code generation, healthcare, finance, and education. This versatility highlights their adaptability and potential to streamline language-related tasks across diverse industries and contexts [3].

**Understanding LLMs**

LLMs are, at their foundation, robust statistical models that acquire knowledge from extensive corpora of textual data. By utilizing methods such as deep learning, they acquire comprehension of subtleties in language, which empowers them to emulate how humans speak and produce outputs that are logical and cognizant of context. There are numerous LLM architectures, each with its own advantages and disadvantages. Common examples consist of:

Transformers: Transformers, such as GPT-3 and Jurassic-1 Jumbo, demonstrate exceptional proficiency in producing imaginative and fluid text, rendering them indispensable for applications such as dialogue systems and content generation.

BERT and its variants: With a primary emphasis on textual comprehension, these models prove to be valuable in various domains such as question answering, text summarization, and sentiment analysis.

**Rise of LLM and integration Across Sectors**

Large language models (LLMs) have surfaced as a groundbreaking technological advancement, swiftly altering how we engage with language and information. Trained on massive datasets of code and text, these complex AI models can generate text that resembles humans, translating languages, composing a variety of creative content, and providing informative responses to queries. The perpetual development of LLM capabilities pushes the limits of what is feasible and initiates substantial integration across diverse industries.

The potential of LLMs has attracted various industries, leading to rapid integration across sectors:

Healthcare: LLMs help in various healthcare domains, including drug discovery through analysis of medical literature, personalized health report generation, and chatbot-based mental health support.

Finance: LLMs are streamlining financial processes by automating duties such as loan approvals and fraud detection, in addition to analyzing financial news and generating trading signals.

Customer Service: Chatbots powered by LLMs respond to customer inquiries, provide personalized recommendations, and efficiently resolve issues, thereby enhancing the customer experience.

Education: LLMs can support and personalize the learning process by translating languages, creating individualized learning materials, and providing written feedback.

Media and Entertainment: LLMs produce various creative works, including news articles, poems, scripts, and even musical compositions. Additionally, they contribute to the generation of personalized recommendations and advertisements.[10]

|  |  |  |
| --- | --- | --- |
| **Industry** | **Uses of LLM** | **Drawbacks of LLM** |
| healthcare | Diagnostics, patient data analysis, personalized medicine, virtual patient assistance | Privacy concerns, reliability in critical diagnostics |
| Finance | Fraud detection, personalized financial advice, risk management, customer service | Security vulnerabilities, algorithmic biases |
| Customer Service | Automated responses, personalized interaction, 24/7 service availability | Misinterpretation of queries, privacy breaches |
| Education | Personalized learning, grading and feedback, content creation | Content accuracy, bias in educational materials |
| Legal | Document analysis, legal research assistance, contract review | Confidentiality risks, accuracy of legal advice |

Table 1. Comparison of different industrial sectors in LLM aspect

**Challenges of LLMs**

* Bias and Fairness: LLMs educated using biased data can reinforce and magnify existing biases within society.
* Security and Privacy: There are ongoing concerns pertaining to potential privacy violations, data security intrusions, and the manipulation of LLM outputs for malicious intent.
* Explainability and Transparency: With the increasing intricacy of LLMs, it becomes imperative to comprehend their decision-making procedures to establish confidence and responsibility.
* Job Displacement: Concerns arise regarding the potential of LLMs to automate tasks currently performed by humans, leading to job losses.

**The Future of LLMs**

The evolution of LLM development is brisk, and the incorporation of LLMs across industries is likely to continue expanding. It is essential to address the obstacles to guarantee ethical and responsible development. The establishment of ethical guidelines for LLM usage, the development of tools to detect bias, and the implementation of explainability mechanisms will be crucial in determining the trajectory of a future in which this technology contributes positively to society.[15]

## Problem Space

The Looming Vulnerabilities of Large Language Models Across Sectors

Large language models (LLMs) are undergoing accelerated development and integration across multiple industries, providing numerous advantages including personalization, accessibility, and efficiency. Nevertheless, during this advancement, a somber shadow arises—an increasing apprehension regarding the susceptibilities of LLMs and their capacity to inflict damage throughout various industries.

The problem space comprises an extensive range of potential vulnerabilities that LLMs may demonstrate, the vast spectrum of industries they may affect, and the complex repercussions that may result from their exploitation. It is essential to examine the particulars to comprehend the situation's true gravity.[12]

Types of LLM Vulnerabilities:

* Prompt Injection: The prompts provided to LLMs can be manipulated by malicious actors to trick them into producing detrimental or unintended outputs. Envision a healthcare chatbot that erroneously delivers medical advice in response to a perceptively designed prompt.[2]
* Data Poisoning: Saturating LLM outputs with skewed or detrimental training data is known as "data poisoning." A finance LLM educated on biased data, for instance, could contribute to the perpetuation of discriminatory lending practices.
* Model Theft: Misuse may result from unauthorized access to or replication of the LLM or its parameters. This could potentially entail the unauthorized acquisition of sensitive data or the malevolent manipulation of the model.
* Adversarial Examples: Constructing inputs with the intention of inducing erroneous or deceptive outputs from the LLM. This may encompass the production of fabricated news articles with the intention of swaying public sentiment or the development of phishing schemes that mimic the appearance of authentic ones.
* Privacy Concerns: The use and instruction of LLMs may cause privacy concerns because they potentially store and process sensitive data.[6]

Sector-Specific Impact:

The consequences of LLM vulnerabilities vary depending on the sector.

* Healthcare: Breach of privacy, misdiagnosis, and tampering with medical records constitute substantial threats to patient welfare and confidence in the healthcare system.
* Finance: Insider trading, financial fraud, and market manipulation have the potential to disrupt financial markets and undermine consumer trust.
* Customer Service: Marketing misinformation, phishing attacks, and manipulation of customer behavior can result in financial losses and reputational harm for a brand.
* Social media: The dissemination of hate speech, disinformation, and the manipulation of public opinion via social media can result in significant societal ramifications.[8]

## Research

Literature review:

A thorough examination of security vulnerabilities, particularly Remote Code Execution (RCE), in applications that are integrated with Large Language Models (LLMs), is addressed in the paper "Demystifying RCE Vulnerabilities in LLM-Integrated Apps" by Tong Liu et al. The study presents LLMSMITH, an innovative instrument that utilizes prompt-based testing and static analysis to identify RCE vulnerabilities in frameworks and applications integrated with LLM. Thirteen vulnerabilities are discovered across six frameworks from the research, which generates seven CVE IDs. The testing of 51 applications revealed vulnerabilities in 17 of them, thereby exposing substantial security risks. This paper makes a valuable contribution to the field of security awareness by presenting mitigation strategies that developers can implement to effectively safeguard against these vulnerabilities.[6]

The article titled "A Survey on Large Language Model (LLM) Security and Privacy: The Positives, Negatives, and Unfavorable" offers an exhaustive examination of the security and privacy implications associated with LLMs. The impact of LLMs is classified into three distinct segments: advantageous applications that bolster security measures, detrimental applications that present risks, and weaknesses present within LLMs in conjunction with potential defenses. The authors shed light on the dual potential of LLMs in cybersecurity by presenting findings on their use in offensive cyber activities, their role in code and data security, and inherent vulnerabilities that require further investigation, as determined by a comprehensive literature review.[7]

This paper presents a methodical examination of the security obstacles encountered by applications that incorporate Large Language Models (LLMs). The text delineates the extent of the risks encountered by these applications, presenting an all-encompassing threat framework and analysis of the attack surface. The study's body of work presents a proposed defense mechanism called Shield, which is described in detail, including its essential characteristics, an exhaustive evaluation of its efficacy, and sections devoted to its description. The paper additionally explores related research by providing a comparative analysis of current security protocols utilized in applications that integrate LLM. In addition to a discussion of the implications and consequences of the findings, the conclusion includes a statement of reproducibility and an ethical statement to emphasize the research's integrity and openness. The discussion is further expanded with supplementary materials that address threat evaluations. These materials consist of evaluation prompts and queries, a cost analysis of security measures, comprehensive attack scenarios, and further details on the cryptographic foundations and security properties proof of the defense design. These resources contribute to a deeper understanding of the proposed defense mechanism's resilience against identified threats.[8]

## Solution Space

In the phase of our project, we start by exploring Large Language Models (LLMs) to understand their structure and uses better. This includes conducting a review of research papers, academic studies and industry reports. By diving into the details of LLM technology we aim to build a foundation of knowledge about how it works and its potential applications.

As we progress our attention turns towards identifying ways in which LLMs could be vulnerable to attacks. Using web scraping methods, we gather information from a variety of sources such as research papers, online forums and cybersecurity websites. This comprehensive approach helps us create a list of attack avenues. We specifically investigate techniques like model inversion attacks, adversarial attacks and input manipulation to shed light on the threats that could compromise the integrity of LLMs.

The next step involves evaluating how feasible it is for these identified attack methods to be carried out successfully. This assessment includes analyzing the level of knowledge and resources needed for each type of attack. By applying AI/ML methodologies and studying data, on LLM attacks we aim to identify patterns that can help us understand what level of expertise an attacker might require to exploit vulnerabilities effectively.To deepen our comprehension further we explore the outcomes of attacks, on Large Language Models (LLMs). We go beyond the effects by looking at how these attacks can affect confidentiality, integrity and availability. This thorough analysis helps us understand the implications of manipulated outputs from LLMs in scenarios.

With an understanding of LLM vulnerabilities we move on to selecting industries or activities for a more detailed examination. By focusing on sectors like healthcare and finance we aim to connect our research to real world situations, ensuring that our findings remain relevant and specific.

Next, we meticulously compare LLM vulnerabilities across sectors and activities. Using the information gathered we create a list of vulnerabilities for each sector. This comparative study allows us to identify vulnerabilities well as unique challenges specific to different industries.

To organize and present our findings effectively we developed a risk assessment method. This includes establishing a scoring system and criteria for assessing and comparing the risk of LLM attacks across industries. Factors such as vulnerability severity, potential impact and likelihood of exploitation are considered in this process. By creating this assessment framework, we ensure a consistent approach that enhances our understanding of LLM security risks, in contexts.

## Project Objectives

Comprehensive Vulnerability Mapping:

* Determine and classify the complete range of LLM vulnerabilities present in diverse industries, including healthcare, finance, customer service, and social media.
* Evaluate the magnitude and potential ramifications of every vulnerability present in every sector, considering both immediate and indirect effects.
* Construct a vulnerability matrix or map that visually represents the interconnections among impacted sectors, vulnerability categories, and LLM types.

Sector-Specific Risk Assessment:

* Perform comprehensive case studies to gain a deeper understanding of practical situations where in-sector LLM vulnerabilities were exploited.
* Using pertinent data and metrics, quantify the potential economic, social, and ethical costs associated with LLM vulnerabilities in each sector.
* Construct risk assessment frameworks that are sector-specific, delineating critical vulnerabilities, strategies for mitigating them, and responsible practices for implementing LLM.

Mitigation Strategy Exploration:

* Assess current approaches to mitigating LLM vulnerabilities, including adversarial training, explainability mechanisms, and data security practices.
* Identify emerging research and development initiatives, such as ethical frameworks and innovative security solutions, that are geared toward addressing LLM vulnerabilities.
* Suggest innovative approaches for mitigating specific LLM vulnerabilities that are pertinent to a wide range of industries.

## Primary User Stories

* To safeguard patient data and mitigate risks, I wish to comprehend the specific vulnerabilities of LLMs utilized in medical chatbots as a security engineer for a healthcare organization.
* As a financial analyst with a vested interest in market manipulation, I require an exhaustive synopsis of the ways in which LLMs might be employed to exploit vulnerabilities in trading systems and disseminate false information.
* As a customer service manager with the objective of safeguarding my clientele against fraudulent activities, I seek pragmatic and implementable approaches to thwart phishing attempts by malevolent actors utilizing LLMs.
* As a developer of social media platforms dedicated to responsible artificial intelligence, I desire transparent protocols and optimal strategies for constructing LLMs that thwart the dissemination of hate speech and disinformation.
* I desire, as a policymaker concerned with the ethical ramifications of artificial intelligence, a framework for assessing the potential social and economic costs of LLM vulnerabilities across multiple sectors.
* As an investigator into the future of artificial intelligence, I desire a collaborative platform where I can develop inventive solutions to LLM vulnerabilities and share knowledge.

# Datasets

## Overview

### VERIS Database:

Implementing the VERIS (Vocabulary for Event Recording and Incident Sharing) framework entails systematically documenting and classifying security violations and incidents. It was created as part of the Verizon Risk Team's initiative to enhance the organization's comprehension of incidents that have a global impact. The objective of VERIS is to establish a standardized set of metrics and a common vernacular for the consistent and repeatable description of security incidents.

Listed below are several essential characteristics of the VERIS database:

* Structured Data Collection: VERIS facilitates the systematic collection and analysis of data pertaining to security incidents for organizations. This comprises details regarding the breach's method of occurrence, the compromised data type, the actions taken upon its discovery and response, as well as the incident's repercussions.
* A comprehensive taxonomy is incorporated into the framework to facilitate the classification of security incidents. The taxonomy encompasses a wide range of incident-related elements, including the partners, external, or internal perpetrators, the actions that precipitated the incident (e.g., social engineering, malware, hacking), the assets that were compromised, and the repercussions experienced by the organization.
* VERIS is a collaborative and open framework that fosters participation and utilization from a wide range of organizations. This transparency encourages data sharing and collaboration regarding security incidents, which can contribute to the advancement of cybersecurity knowledge and defense strategies.
* Data Analysis and Sharing: The information gathered via VERIS may be disseminated in an anonymous fashion to the public, thereby contributing to the yearly Verizon Data Breach Investigations Report (DBIR). By analyzing security incidents and breach trends, this report provides significant insights into the present state of the threat landscape.
* Assistance with Tools: Various tools are accessible to aid organizations in the integration of VERIS into their incident recording and analysis procedures. These tools facilitate the standardization of incident data capture, thereby streamlining the process of trend analysis and information sharing.

Enabling organizations to gain insights from one another's experiences, the VERIS database is an indispensable tool in advancing the comprehension of security incidents. By standardizing the reporting and analysis of incidents, VERIS contributes to the enhancement of the international community's security posture.

### CVE Database:

A compendium of identified security vulnerabilities and exposures pertaining to a wide range of hardware and software is comprised of the Common Vulnerabilities and Exposures (CVE) database, which is accessible to the public. The CVE database, which is overseen by the MITRE Corporation, is of utmost importance in the cybersecurity ecosystem as it assigns a unique and standardized identifier (CVE ID) to each vulnerability or exposure. The subsequent synopsis elucidates the fundamental characteristics and significance of the CVE database:

Structured Identification of Vulnerabilities

* CVE IDs: A distinct identifier, referred to as a CVE ID, is allocated to every entry in the CVE database. This system enables convenient reference and discourse regarding a specific vulnerability across various platforms and tools.
* Standardization: CVE IDs facilitate the identification of vulnerabilities in a consistent manner, assisting cybersecurity professionals in promptly comprehending the characteristics and particulars of a security concern devoid of any uncertainty or confusion.

Comprehensive Descriptions of Vulnerabilities

* Detailed Records: The database comprises comprehensive descriptions of vulnerabilities, encompassing their potential impacts, methods of exploitation, and established remedies or workarounds. This information is of the utmost importance for remediation and vulnerability assessment.
* Reference Links: The CVE database frequently incorporates citations to additional security advisories, patches, and reports, thereby furnishing a holistic perspective of the vulnerability derived from various sources.

The facilitation of security administration

* Interoperability is facilitated by the standardized format of CVE IDs, which enables security information and event management (SIEM) systems, vulnerability analyzers, and intrusion detection systems to operate in concert.
* Severity Evaluation: CVE IDs are frequently accompanied by the Common Vulnerability Scoring System (CVSS), which provides a standardized method for evaluating the severity of vulnerabilities. This facilitates the prioritization of response and mitigation efforts by organizations according to the potential impact.

A screenshot of a computer

Description automatically generated

**Figure 1: Snapshot of CVE database**

### MITRE ATLAS Database

The MITRE Corporation oversees MITRE ATLAS, a globally accessible data storage system that contains a wealth of information on the strategies and techniques adversaries employ in actual AI system attacks. With a focus on Linguistic Layer Models (LLM), this database provides cybersecurity professionals with essential knowledge regarding the many methods that hackers employ to target AI systems. With about 75 LLM attack routes meticulously recorded in its files, MITRE ATLAS allows experts to precisely examine the tactics and methods utilized during these targeted attacks. For cybersecurity professionals trying to fortify their defenses against LLM attacks and other emerging cyberthreats, MITRE ATLAS is essentially an indispensable tool. Experts can identify and reduce risks more effectively by studying the specifics of adversary behavior and attack methods when developing improved defenses for AI systems. The global accessibility of the repository fosters collaboration and information sharing within international cybersecurity communities. This enables cooperative efforts to fortify cybersecurity postures across all industries and stay abreast of evolving threats.

With the help of MITRE ATLAS, cybersecurity professionals have a systematized way to understand LLM attacks in detail. This structure is useful for finding possible weaknesses in AI systems and it also helps organizations to put countermeasures into place ahead of time so that risks can be reduced effectively. The knowledge from MITRE ATLAS allows those working in cybersecurity to stay ahead of the changing nature of complex computer risks. This aids in protecting important infrastructure and maintaining the value of AI systems within an ever more connected digital environment.

### Census Dataset

The Census dataset is a comprehensive compilation of economic and demographic information compiled by government organizations. It provides us with crucial insights into all facets of society and the operations of enterprises. The North American Industry Classification System (NAICS) code for each industry is a significant component of this collection. These codes aid in the identification of distinct economic sectors and enable accurate categorization and analysis of companies based on their primary functions. Because they make it easier to determine which industrial sector a company entity is in, the NAICS numbers are significant. These codes are used to structure the Census statistics, making it possible to compare and analyze various sectors in great depth. This enables scholars and decision-makers to precisely comprehend changes in the industrial sector, labor force composition, and economic trends.

The Vulnerability Coordination and Disclosure Barometer (VCDB), a crucial instrument in cybersecurity, is used to gather and arrange data regarding cyberattacks. One unique feature of VCDB is its ability to classify attacks based on industry NAICS codes. These NAICS numbers associated with companies that are the subject of cyberattacks are used by cybersecurity specialists to analyze attack trends and patterns in a variety of industries. This makes it possible for them to create targeted responses to threats, encouraging reactive rather than passive defense tactics.

It is possible to analyze Linguistic Layer Model (LLM) attacks occurring in certain industrial sectors with a strong structure by incorporating NAICS codes into the Census information and applying them in VCDB. When attack data is compared to NAICS codes, cybersecurity professionals have a better knowledge of the unique challenges and vulnerabilities that each industry faces. This aids in the development of cybersecurity policies and techniques that are unique to each industry. The collaborative analysis of data enhances the efficacy of cybersecurity measures, enabling entities to effectively manage risks and safeguard critical infrastructure in a dynamic threat landscape.

## Field Descriptions

### VERIS Data:

* Incident ID (Type: string) - For every reported incident, a universally unique identifier (UUID) is assigned. This ID enables the straightforward and unambiguous location of a record. The purpose of the ID is to guarantee that each incident can be uniquely identified, regardless of the record's status, as it is transferred between the entities responsible for reporting, addressing, or investigating the incident.
* Incident Title (Type: string)- A designation or title assigned to an incident, typically denoting its nature or a particular facet of the occurrence (e.g., "Data Breach in the Cloud Environment"). It is essential that the name or headline of the event be formatted and described in a clear manner, regardless of any variations in nature, type, or industry.
* Victim Information (Type: string)- An array of textual identities, along with other comparable notations, that represent the class-based and individual entities identified as users of the compromised systems. However, it is possible for the field to be empty when data imputation fails to provide sufficient information for detection.
* Incident Summary (Type: text) - Presents a guided summary of the incident that was reported, generally formulated from a retrospective perspective. From a worldview perspective, its documentation would typically include the relationship and sequence of the environment, area, or impact in question.
* Date and Time of Occurrence (Type: datetime)-Embedded notifications indicate when perceived outcomes or observable space were impacted by a particular protocol. The formation of notations under this content ID is permissible via UTC (YYYY-MM-DD or YYYY-MM-DD). The purpose of these information outcomes is to similarly subjugate all organizational and individual actors.
* Threat Actor Organization (String: type) - Victimization indicated by neighboring correspondents who may have multiple receivers, as evidenced by identification. This exposes the trace to the potential for knowledge fragmentation, online-offline impact, or an isolated technological transaction, thereby indicating a temporal association or correlation with a solitary organization or network of susceptible objects.
* Tactics, techniques, and procedures (TTPs) (Type: text): This section typically contains information gathered through observation by organizations or data handlers. It ensues after an analytical communiqué detailing the means, attack activities, preparation, infiltration or entry period, samples, and event support forms.
* Discovery Information (JSON Type) - Conservatory in nature, this communicates the identification and administration of processes categorized, the company's victim, and the idiosyncratic strategy that reintroduces or self-maps the flaw to identifiable game theory or indicates the approach of failure.
* Implementation of Controls and Outcomes (Data Type: string) - This data record associates a value or metric with a digital resource after a compromise, loss, or an interpretive period of abuse attempts. In essence, it should be digitized to enable genuine quantum supervision and assistance, thereby mitigating the risks associated with delays or outsourced security access.
* Economic Effects (Type: float) - A compound-time tincture or hold-hold tithe of an unknown or family member that motivates actions that have a financial repercussion. In a narrative sense, it originates from or approaches a grading system that is indicative of intellectual, public, or product-oriented attire.

For the sake of international comprehension, each primary domain of the VERIS data is modelled to inform, share, and quantify on a global scale in accordance with perspectives. The correspondence with the prevalent diversity of elements such as a comprehensive timeline, time, or current applicable figures is not an exact one-to-one match. The VERIS system assigns a significant reflection, control, and procedural impetus to each note to facilitate a formalized community and industry stance regarding health assessment and service advancements, as intended by the framework.

### CVE Data:

The Common Vulnerabilities and Exposures (CVE) database is a standardized repository developed by the MITRE Corporation to catalog and describe cybersecurity vulnerabilities across various software and hardware systems. It serves as a comprehensive reference for identifying, tracking, and mitigating security flaws that could be exploited by threat actors. The CVE database plays a vital role in our project associated with Large Language Models (LLMs) and attack vectors by providing a structured framework to document and analyze vulnerabilities relevant to LLM technologies. Leveraging CVE data enables our team to identify common vulnerabilities, understand attack patterns, and prioritize mitigation strategies effectively. By utilizing the CVE dataset, we can enhance our understanding of cybersecurity risks associated with LLMs across different sectors and bolster our defenses against potential threats.

* Vulnerability ID (Type: string):

A unique identifier assigned to each reported vulnerability, allowing for easy tracking and reference across different entities involved in reporting, analyzing, or mitigating the vulnerability.

* Vulnerability Title (Type: string):

A descriptive title assigned to the vulnerability, summarizing its nature or key aspects (e.g., "Remote Code Execution in Software X"). Clarity and consistency in naming are essential for effective communication.

* Affected Products (Type: string):

An array of textual descriptions identifying the software or hardware products impacted by the vulnerability. This field may be empty if insufficient data is available for precise identification.

* Vulnerability Description (Type: text):

Provides a detailed summary of the vulnerability, often presented retrospectively. It outlines the nature of the vulnerability, its potential impact, and any relevant technical details for understanding and addressing the issue.

* Date of Discovery (Type: datetime):

Indicates when the vulnerability was first identified or reported. Timestamps follow UTC format (YYYY-MM-DD or YYYY-MM-DD) to ensure consistency across different time zones.

* Attack Vector (Type: string):

Describes the path or method through which the vulnerability can be exploited, focusing on intermediate assets or services targeted during an attack rather than the final victim asset.

* Exploited CVEs (Type: text):

Records any Common Vulnerabilities and Exposures (CVEs) that were exploited as part of the attack. This field helps in understanding specific vulnerabilities leveraged by threat actors during exploitation.

* Impact Assessment (Type: float):

Quantifies the severity or impact of the vulnerability on affected systems, often measured on a scale to indicate the potential consequences of exploitation in terms of confidentiality, integrity, and availability.

* Mitigation Strategies (Data Type: string):

Details the recommended controls or actions to address and mitigate the identified vulnerability post-exploitation. This information aids in developing effective remediation plans to reduce risks associated with vulnerability.

* Financial Implications (Type: float):

Represents any financial repercussions associated with exploiting the vulnerability, providing insights into potential costs incurred due to security breaches or system compromises.

## Data Context

### VERIS Data

Incident Description:

This segment presents the incident narrative in a format that is more conducive to trending and analysis, transforming the "who did what to what (or whom) with what result" into a more manageable format. VERIS accomplishes this by implementing the A4 threat model that Verizon's RISK Team has developed. The A4 model defines an incident as a sequence of occurrences that have a detrimental impact on the information assets of a given organization. The subsequent components (the 4 A's) comprise each event and serve as the overarching framework for metrics in this section.

Actors: Whose actions affected the asset?

Actions: What actions affected the asset?

Assets: Which assets were affected?

Attributes: How was the asset affected?

The VERIS framework, which stands for Vocabulary for Event Recording and Incident Sharing, classifies and analyzes security incidents in a methodical fashion. The "4A" classification, which divides the components of a security incident into four principal categories—Assets, Actors, Attributes, and Assets—is a fundamental concept of VERIS. The purpose of these categories is to furnish a thorough synopsis of the incident, encompassing the perpetrators, the methods employed, the intended target, and the resultant consequence. Each is briefly described as follows:

**Actors:**

This classification designates the entity culpable for the occurrence. Actors are conventionally classified according to their operational methodology, motivation (such as vengeance, espionage, or financial gain), or organizational affiliation.

The following are subcategories:

* External entities include hackers, criminal organizations, and state-sponsored organizations, among others.
* Internal: Entities comprising personnel, contractors, or partners who could initiate a security breach inadvertently or intentionally.
* Partner: Entities that are affiliated with the organization in some capacity for conducting business and via which the incident transpires or is enabled.

**Actions:**

This category delineates the techniques employed by the assailant to manipulate the asset. Instances of malicious software installation, social engineering, and physical assaults are all possible.

The following are subcategories:

* Malware encompasses any application designed to interfere with computer processes, steal confidential data, or infiltrate private computer systems.
* Exploitation of a vulnerability in a computer system or network constitutes hacking.
* Social: Methods employed to deceive individuals into circumventing standard security protocols.
* Physical: Behaviors that require physical contact, such as theft, hardware device injury, or tampering.

**Assets**

Assets are the objects that fall victim to security incidents. This category provides information regarding the elements that were compromised or impacted throughout the incident, including hardware, software, or data.

The following are subcategories:

* Database servers, message servers, and web servers are all examples of servers.
* User Devices: Mobile devices, desktop computers, and laptops.
* Networks: Comprising both internal and external infrastructure for networks.

**Attributes**

Attributes delineate the consequences or ramifications that an incident has on an organization. This contains information on the financial and reputational repercussions, in addition to the data's confidentiality, availability, and integrity.

The following are subcategories:

* Confidentiality refers to circumstances in which information is disclosed without authorization.
* Integrity refers to occurrences in which data is altered.
* Availability: Occurrences that compromise the accessibility of services or information, including denial of service attacks.

Components of VERIS Data Context:

* Threat intelligence comprises details pertaining to the capabilities, methodologies, and motivations of the entities posing a threat. This encompasses information regarding established attack patterns, utilized tools, and any indicators of compromise (IoCs) that may be utilized to discern and ascribe attacks.
* Vulnerability information pertains to specifics regarding hardware or software weaknesses exploited throughout the incident. This encompasses data pertaining to unpatched systems, identified yet unmitigated vulnerabilities, and the methods employed to exploit said vulnerabilities.
* Impact Analysis: In addition to the immediate ramifications delineated in the Attributes section, Data Context examines the more extensive evasion of an incident's effects. This encompasses enduring financial consequences, harm to reputation, legal and regulatory ramifications, and stakeholder effects.
* Security Controls: Details about the efficacy of the security measures implemented during the incident.mpass information regarding controls that were unsuccessful, in addition to controls that effectively alleviated elements of the incident.
* Incident Discovery and Response: An examination of the incident detection process, the response timeline, and the efficacy of incident response measures. In addition, this section contains information regarding any obstacles faced during the response and any lessons learned.
* Business Context: Comprehending the impact of the incident on business processes, operations, and assets. This encompasses the impact of the incident on the operational capabilities of the organization, any interruptions to essential services, and the strategic significance of the assets that were impacted.
* External Environment: An analysis of the potential influence of external factors on the incident or its consequences. Potential factors to be considered encompass geopolitical dynamics, regulatory frameworks, market dynamics, and pertinent cyber threat patterns.

By incorporating Data Context into the VERIS framework incident log's fundamental components, organizations can acquire a more comprehensive comprehension of every incident. This holistic perspective facilitates the detection of systemic vulnerabilities, enhances the state of security, and formulates approaches to alleviate forthcoming hazards. Data Context facilitates a more comprehensive examination that surpasses the immediate technical intricacies of an occurrence, encompassing the wider-ranging organizational, environmental, and strategic ramifications.

### CVE Data

Incident Description:

This segment presents the incident narrative in a format conducive to analysis, transforming the details of identified vulnerabilities into a manageable framework. The CVE (Common Vulnerabilities and Exposures) framework accomplishes this by categorizing vulnerabilities and exposures, providing a standardized approach to tracking and sharing security vulnerabilities. Each vulnerability is assigned a unique identifier known as a CVE ID, facilitating communication and coordination among cybersecurity professionals and organizations.

Components of CVE Dataset:

* CVE ID: A unique identifier assigned to each security vulnerability or exposure, allowing for easy reference and communication.
* Vulnerability Description: This section provides a concise description of the identified vulnerability, including details such as affected software, potential impact, and mitigating factors.
* Attack Vectors: Describes the methods by which attackers can exploit the vulnerability, including remote or local exploitation, authentication requirements, and potential prerequisites.
* Affected Software: Lists the software or systems known to be affected by the vulnerability, including specific versions and configurations.
* Vulnerability Type: Classifies the vulnerability based on its nature, such as buffer overflow, SQL injection, cross-site scripting (XSS), or authentication bypass.
* Severity: Indicates the potential impact of the vulnerability on affected systems or data, typically categorized as low, medium, high, or critical.
* Mitigation: Provides guidance on mitigating vulnerability, including patches, workarounds, or configuration changes recommended by vendors or cybersecurity experts.
* Exploitability: Assesses the ease with which an attacker can exploit the vulnerability, considering factors such as available exploit code, complexity of exploitation, and likelihood of successful attacks.
* CVE References: Includes links to additional resources, such as vendor advisories, security bulletins, or technical documentation related to the vulnerability. By incorporating these components into the CVE dataset, cybersecurity professionals and organizations can effectively track, prioritize, and mitigate security vulnerabilities, enhancing the overall security posture of their systems and networks. This standardized approach to vulnerability management facilitates proactive risk mitigation and enables timely response to be emerging threats, ultimately reducing the likelihood and impact of cybersecurity incidents.

## Data Conditioning

### VERIS Data

Data conditioning of the VERIS dataset involves several crucial steps to ensure the accuracy and reliability of the analysis. Initially, the dataset is imported and configured using Python and Pandas, leveraging tools like VERISpy to extract JSON objects into a structured Pandas data frame. This process likely includes handling missing values, formatting inconsistencies, and standardizing data types to facilitate further analysis. Additionally, during the exploration phase, potential data anomalies and outliers are identified and addressed to prevent skewing the analysis results. Furthermore, filtering and cleaning operations are performed to focus on specific categories or incident types of interest, such as incidents caused by internal actors like developers. By conducting these data conditioning steps, the VERIS dataset can be prepared for comprehensive analysis, ensuring that the insights derived are accurate, reliable, and actionable for cybersecurity purposes. Here's a flowchart illustrating the data conditioning process for the VERIS dataset:

A diagram of standardize data types

Description automatically generated

**Figure 2: Data conditioning**

### CVE Data

Data conditioning of CVE data is crucial for cybersecurity stakeholders as it enables the identification and prioritization of vulnerabilities, ultimately enhancing overall cybersecurity preparedness. By ensuring the cleanliness, standardization, and removal of duplications within CVE datasets, security professionals can effectively identify and understand the specific vulnerabilities affecting their systems. Well-organized data empowers organizations to intelligently allocate resources by assessing risk scores, severity ratings, and exploitability metrics. This approach allows for the efficient allocation of resources to address the most critical security risks. Organizations can strategically utilize conditioned CVE data to proactively manage emerging threats, stay updated on patch availability, and mitigate potential cyberattacks, thereby strengthening their cybersecurity defenses and safeguarding their digital assets and reputation.

Gathering CVE data involves various vital procedures. Initially, sources like the National Vulnerability Database (NVD) and other aggregators are used to collect the data. Next, unfiltered information is analyzed to extract crucial elements like CVE ID, description, impacted products, and gravity. Following this, standardization guarantees uniformity in formats by adjusting date notations and merging product designations. Removing duplicate entries is essential for effective data management, as enriching the CVE information with insights from multiple sources, such as vendor patches and exploit assessments, is critical. Using taxonomies like CWE simplifies the identification of vulnerabilities. Risk assessment utilizes systems like CVSS to determine severity levels based on potential impact and exploit likelihood. The shift in temporal patterns allows monitoring of vulnerability progression over time. By incorporating CVE data feeds into security systems, real-time tracking and analysis can be achieved. Consistent updates and thorough quality assurance measures guarantee the precision and comprehensiveness of the curated CVE dataset, facilitating security experts and entities in proficiently overseeing and minimizing potential security hazards.

Diverse sources such as vendor-specific repositories and security advisories contribute to a comprehensive dataset beyond the National Vulnerability Database (NVD). Advanced algorithms, along with natural language processing techniques - parsing methods in particular - extract crucial details from heterogeneous data sources; they successfully overcome challenges presented by format variations and language nuances. Efforts towards normalization address inconsistencies within data formats and severity ratings: concurrently, enrichment strategies employ automated web scraping plus API integrations to enhance CVE entries with contextual information sourced broadly. The Common Weakness Enumeration (CWE), a categorization framework, structures vulnerabilities for prioritization and analysis: it's an aid. As emerging threat landscapes evolve, risk scoring methodologies - aligning with organizational risk tolerance levels – also adapt; this demonstrates their dynamic nature. Through temporal analysis--which reveals patterns in vulnerability emergence and patch adoption--proactive response strategies receive crucial information.

## Data Quality Assessment

### VERIS Data:

The VERIS dataset undergoes a rigorous evaluation across various dimensions to ensure its quality and reliability. Completeness is ensured by capturing comprehensive incident narratives that encompass all relevant aspects, including actors, actions, assets, and attributes, providing a thorough understanding of each security incident. Uniqueness is maintained through the absence of duplicate records, ensuring that each entry represents a distinct occurrence. Accuracy is upheld by verifying incident details against reliable sources, ensuring precise and reliable information regarding the impact on organizational assets. Atomicity ensures that each incident entry represents a single, indivisible occurrence, facilitating granular analysis and understanding. Conformity to predefined standards and conventions guarantees consistency and comparability across incidents, enhancing the dataset's usability and reliability. Overall quality is continually monitored and improved through regular validation processes, ensuring that the dataset provides a comprehensive, accurate, and reliable depiction of security incidents, ultimately supporting effective trending, analysis, and proactive risk mitigation efforts.

### CVE Data:

The CVE dataset undergoes a meticulous data quality assessment across various dimensions to ensure its reliability and utility. Completeness is guaranteed by including detailed information about each identified vulnerability, covering aspects such as CVE ID, vulnerability description, attack vectors, affected software, vulnerability type, severity, mitigation, exploitability, and CVE references. Uniqueness is maintained through the assignment of unique CVE IDs to each vulnerability, ensuring clear identification and differentiation. Accuracy is upheld by verifying vulnerability details against authoritative sources, ensuring precise and reliable information regarding the nature and impact of each vulnerability. Atomicity ensures that each vulnerability entry represents a single, distinct security issue, allowing for granular analysis and understanding. Conformity to established CVE standards and conventions ensures consistency and comparability across vulnerability entries, enhancing the dataset's usability and reliability. Overall quality is continually monitored and enhanced through regular validation processes, ensuring that the CVE dataset provides a comprehensive, accurate, and reliable resource for tracking and mitigating security vulnerabilities, ultimately supporting effective risk management and cybersecurity efforts.

## Other Data Sources

When collecting and analyzing data on industry-specific assaults on Large Language Models (LLMs), it is possible to investigate a range of data sources and methodologies. Although certain sources may appear promising at first glance, they may prove unsuitable for implementation due to various factors. The rationale behind the exclusion of specific methodologies, including web crawling from Azure Cloud Services and employing Bing Chat v7 for data collection, from the final analysis is as follows:

**Conducting web scraping using Azure Cloud Services and Bing Chat v7**

* Data Relevance: The relevance of the collected data may have constituted the principal obstacle. Although a considerable quantity of websites was scraped, the information obtained may not have comprehensively examined the intricacies of attacks targeting LLMs across diverse industries. The absence of precise details may diminish the concentration necessary for a focused examination of LLM vulnerabilities.
* Quality and Accuracy: Web crawling frequently generates vast quantities of data whose quality and precision can vary considerably. The acquired information may have lacked specificity, been obsolete, or been excessively general, thereby impeding the ability to derive meaningful conclusions regarding LLM attacks.
* Legal and ethical considerations may be brought into the forefront when web skimming occurs, particularly in situations involving copyrighted content or data that is safeguarded by terms of service agreements. The potential legal ramifications could have played a role in the decision to abstain from utilizing collected data from specific sources.
* The management and processing of an excessive volume of collected data can be a burdensome and resource-intensive task. Potential benefits may have been outweighed by the effort required to cleanse, preprocess, and analyze the data, particularly if the data had no direct bearing on LLM assaults.

**Sentiment Analysis and Word Frequency**

* Insufficient Contextual Depth: Although sentiment analysis and frequency analysis of words can yield valuable insights regarding overall trends and attitudes, they may fall short in providing the necessary depth to discern particular security vulnerabilities and attack vectors that impact LLMs. These approaches might prove inadequate in capturing the intricate technical details and contextual subtleties of cybersecurity threats.
* The task of interpreting the outcomes of word frequency and sentiment analysis within the domain of cybersecurity presents certain difficulties. Understanding the sentiment surrounding security incidents is frequently intricate and necessitates a nuanced comprehension of the industry-specific and technical context, neither of which may be immediately apparent from the data.

**Conclusion:**

Factors such as data relevance, quality, legal and ethical ramifications, and the difficulties linked to data overflow and interpretation ultimately influenced the determination to abstain from utilizing specific data sources. By concentrating on more specific, reputable sources, such as the VERIS database and CVE entries, an examination of assaults on LLMs across various sectors can be rendered more accurate and significant.

## Storage Medium

The strategic choice was made to utilize the VERIS (Vocabulary for Event Recording and Incident Sharing) and CVE (Common Vulnerabilities and Exposures) databases as the primary data storage mediums for the project, which aimed to analyze attacks on Large Language Models (LLMs) across multiple industries. Every database provides distinct benefits and attributes that are in perfect harmony with the project's goals. The rationale for selecting these databases as the repository for the project dataset is as follows:

### VERIS database:

The VERIS database contains a structured and standardized framework designed to facilitate the documentation and categorization of security incidents. This framework is of the utmost importance for the methodical examination and juxtaposition of incidents spanning various sectors.

Exhaustive Incident Particulars: VERIS amasses an extensive array of incident particulars, encompassing the assault methods employed, the assets compromised, and the repercussions experienced by the organization. Such comprehensive information is of the utmost importance in comprehending the susceptibilities and avenues of attack that target LLMs.

The open and collaborative nature of VERIS fosters an environment that promotes the exchange of ideas and cooperation between organizations. The transparency and availability of a greater variety and quantity of incident data for analysis contribute to the strength and reliability of the project's results.

### CVE database:

Standardized Vulnerability Information from the CVE Database: The CVE database provides a standardized system for identifying vulnerabilities, enabling discussion and reference that are consistent. The establishment of this standard is critical to precisely detect and assess vulnerabilities that may impact LLMs.

Comprehensive Vulnerability Descriptions: The CVE database comprises entries that furnish elaborate explanations of vulnerabilities, in addition to citations for available remedies and workarounds. This comprehensive data facilitates thorough examination of potential risks to LLMs and aids in the formulation of strategies to mitigate them.

Extensive Usage and Acceptance: Consensus and utilization throughout the cybersecurity community establish the CVE database as a dependable repository of vulnerability data. Furthermore, the extensive implementation of this technology enables seamless integration with various security tools and databases, thereby augmenting the project's repository of data.

Remaining Storage Moderate Opposition Aspects of Consideration

Reliability and Accessibility: The management of the VERIS and CVE databases is entrusted to reputable organizations, which guarantees the data's dependability and availability. Securing this level of dependability is essential for preserving the integrity of the project's results.

The databases' emphasis on security incidents and vulnerabilities renders them exceedingly pertinent to the objectives of the project in terms of both quality and relevance. The data's exceptional quality and specificity enable a nuanced examination of the risks faced by LLMs.

Assistance with Analysis and Reporting: The streamlined structure of both databases facilitates effective data reporting and analysis. The exhaustive trend analysis made possible by the availability of detailed incident and vulnerability information contributes to the project's goals of improving comprehension and implementing countermeasures against LLM-related threats.

In brief, the VERIS and CVE databases were chosen to house the project dataset on account of their collaborative and transparent structure, extensive and detailed incident information, and systematic data collection. Due to these attributes, they are highly suitable for facilitating comprehensive examinations of security breaches and susceptibilities associated with Large Language Models.

# Algorithms & Analysis / ML Model Exploration & Selection

## 3.1 Solution Approach

### Systems Architecture

**Data Ingestion**: We collect and ingest raw data from both the VERIS and CVE datasets: this includes textual descriptions of cybersecurity incidents; vulnerability information; timestamps--which provide a chronological context for these events--, severity scores–that quantify the intensity or potential harm caused by an incident–and other pertinent attributes. After initially ingesting the data, we subject it to critical transformation and cleaning processes for ensuring proper formatting and consistency: these steps include removing duplicates; handling missing values - a crucial task – as well as standardizing data formats across various sources.

**Model Training Infrastructure:** Efficient language model training necessitates robust infrastructure: specifically, distributed computing resources like GPUs or TPUs. These tools are indispensable for grappling with the size and complexity of both VERIS - Vocabulary for Event Recording and Incident Sharing – as well as CVE - Common Vulnerabilities and Exposures – datasets.

**Distributed Computing**: We utilize distributed computing frameworks and approaches to parallelize the training workload across multiple nodes or clusters; this strategy enables us not only for faster model training--but also scalability in handling large datasets.

**Training Frameworks**: Developers employ frameworks like TensorFlow or PyTorch to construct and train the language model architecture; these powerful tools and APIs facilitate the development of intricate neural network models--optimizing training performance in turn.

**Model Deployment**: Inference Deployment: After training the language model, we must deploy it for inference. This process usually requires us to establish APIs or other interfaces; their function is to handle incoming requests and manage data processing - either in real-time or through batch operations.

The deployment architecture employs a horizontal scalability design to accommodate the escalating data from both VERIS and CVE datasets, along with user demand. It implements load balancing algorithms for an even distribution of incoming requests across numerous instances of the deployed model; this optimizes resource utilization--a crucial factor in robust system performance.

To ensure reliable system performance during fluctuating loads, the architecture integrates fault tolerance mechanisms. These include redundancy; failover strategies--which are automated recovery mechanisms designed to minimize downtime and guarantee uninterrupted service.

Implementing these enhanced components into the system architecture enables organizations to process and analyze cybersecurity event data efficiently from both the VERIS and CVE datasets. This implementation also prioritizes scalability, reliability, and security.

### Systems Security

**Data Encryption**:

Implement encryption mechanisms for End-to-End Encryption; this ensures that data remains encrypted from its inception - whether captured or generated - until consumption or storage. The process involves two crucial steps: encrypting the data in transit and at rest--employing robust encryption algorithms along with secure protocols.

Establish robust practices in key management: securely generate, store, and rotate encryption keys. This practice guarantees that only authorized entities possess access to decryption keys; consequently, it thwarts any unauthorized attempts at accessing sensitive data.

Utilize techniques like data masking and tokenization for an advanced level of sensitive information protection; this ensures only authorized users can access the original data.

**Access Control**:

Implement Role-Based Access Control (RBAC) policies to establish role definitions and delineate permissions for system access. This implementation guarantees that users exclusively gain entry to data and functionalities indispensable within their assigned roles, thereby aligning with their specific responsibilities.

Enforce Multi-Factor Authentication (MFA) mechanisms: This action adds an additional security layer during the authentication process; it serves to deter unauthorized access, particularly in cases where user credentials might be compromised.

Implementing Privileged Access Management (PAM) solutions: this allows for a tight control and meticulous monitoring of sensitive data– particularly critical system resources; it provides an effective guard, predominantly over privileged users or administrators.

**Threat Detection**:

Deploying Intrusion Detection and Prevention Systems (IDPS) solutions allows for the vigilant monitoring of network traffic, system logs, and user activities: these measures aim to uncover signs of unauthorized access or malicious behavior. Such systems boast an automatic capability--they not only detect security threats in real time but also respond to them promptly.

We utilize Security Information and Event Management (SIEM) platforms to aggregate and analyze security event logs from diverse system sources; this approach facilitates proactive threat detection, incident response - and even forensic analysis.

Implement anomaly detection algorithms to identify abnormal patterns or deviations from normal system behavior; this proactive approach assists in detecting sophisticated cyber-attacks--or insider threats--that might elude traditional security measures.

Conducting simulated incident response drills and tabletop exercises: this strategy tests the effectiveness of security incident procedures. It not only identifies gaps in detecting, responding to, or mitigating an issue - but also allows for proactive improvements; indeed, a crucial approach towards enhancing capabilities.

### Systems Data Flows

**Data Cleansing**: The VERIS dataset, in conjunction with the Common Vulnerabilities and Exposures (CVE) dataset, might contain inconsistencies, errors or vacant values. Techniques such as outlier detection, imputation and error correction come into play to guarantee data quality and consistency; this guarantees the reliability and accuracy of analytical data.

**Normalization**: The VERIS and CVE datasets might encapsulate data in diverse formats or units; however, employing normalization techniques—specifically scaling or standardization—can metamorphose the data from both sets into a mutual scale. Such standardization: an act that fosters comparison and analysis across varying attributes, amplifies interpretability--and consequently deepens our understanding of the presented information.

**Tokenization, Stemming, and Stop-word Removal:** The VERIS dataset and the CVE dataset apply text processing techniques, such as tokenization, stemming, and stop-word removal to their textual descriptions of cybersecurity incidents and vulnerabilities. This structured transformation proves essential for analysis and modeling: it is particularly suitable for tackling natural language processing tasks.

**Feature Engineering**: In the training of models, we select pertinent features from both the VERIS and CVE datasets. These features encapsulate critical information about cybersecurity incidents and vulnerabilities; they play a significant role in predicting or comprehending their occurrence and impact. To enhance model performance and efficiency: techniques such as Principal Component Analysis (PCA) – or feature embedding–are applied for reducing dimensionality of the feature space while retaining crucially important data.

We extract temporal features such as timestamps, incident durations, and vulnerability discovery dates from both datasets. These attributes grasp the dynamics of cybersecurity incidents and vulnerabilities over time by capturing temporal patterns and trends; thus providing valuable insights into their evolution.

**Model Training**: We select suitable machine learning algorithms to train the language model, utilizing features extracted from both VERIS and CVE datasets. These might encompass deep learning models like recurrent neural networks, transformers, decision trees; as well as ensemble methods—depending on the problem's nature and data characteristics.

Techniques like grid search or random search optimize the model's hyper parameters, aiming to enhance performance and generalization ability with an optimal configuration. To evaluate the generalization performance of a model and mitigate overfitting, we use cross-validation methods such as k-fold cross-validation.

**Inference:** Models deployed in real-time process incoming data from both the VERIS and CVE datasets, providing immediate responses or predictions. This critical functionality caters to applications that demand swift decision-making and response towards cybersecurity incidents as well as vulnerabilities.

We utilize batch inference mode for offline analysis or batch processing of large data volumes from both datasets. This deployment enables efficient data processing, allowing us to examine historical incident data and extensive vulnerability datasets extensively.

The model generates results or predictions during inference; we interpret these, performing analysis to derive insights about cybersecurity incidents and vulnerabilities. To make the outcomes interpretable and actionable, we apply post-processing steps--thresholding, aggregation, and visualization.

### Algorithms & Analysis

**Natural Language Processing (NLP) Techniques:** We tokenize textual descriptions from both the VERIS and CVE datasets, splitting them into individual words or tokens; this crucial process facilitates subsequent analysis and feature extraction. In both datasets, we stem the words to their root or base form; for instance, "attacked", "attacking", and "attacks" are all simplified as 'attack'. This process of stemming simplifies text analysis: it treats identical words similarly--a key advantage.

**POS Tagging:** It identifies the grammatical parts of speech in textual descriptions from both datasets; this process aids in extracting specific types of information--such as affected assets or undertaken actions--from the text.

**Statistical Analysis**: By analyzing the frequency distribution of various types of cybersecurity incidents from the VERIS dataset and vulnerabilities from the CVE dataset, we gain insights into: Graduate-level punctuation—such as colons; semi-colons; and dashes—are used here to enhance readability. These provide a clear understanding on what types are most encountered by organizations in terms of security threats. Temporal patterns and trends in cybersecurity incidents and vulnerabilities' occurrence undergo examination over diverse time intervals. This analysis actively identifies seasonal trends or long-term alterations within the threat landscape: an essential tool for trend analysis over time.

**Anomaly Detection**: We apply statistical techniques—such as z-score analysis or time-series anomaly detection algorithms—to identify anomalous patterns, outliers in both datasets; this aids us in detecting unusual cybersecurity incidents or vulnerabilities that deviate from normal behavior.

**Machine Learning Models**: Supervised machine learning models, trained on both VERIS and CVE datasets, use incident classification to categorize cybersecurity incidents into predefined groups such as incident types and impact levels. This automation accelerates the efficiency of incident management by parsing textual descriptions along with other attributes.

Unsupervised learning algorithms, trained on both datasets, cluster similar cybersecurity incidents and vulnerabilities according to their textual descriptions or other features; this aids in the comprehension of common attack patterns or vectors: a crucial aspect of understanding cyber threats.

Predictive models, developed from historical data in both datasets, forecast future trends in cybersecurity incidents and vulnerabilities. This development equips organizations to mitigate risks proactively by foreseeing potential security threats.

**Visualization Techniques**:

**Natural Language Processing (NLP) Techniques**: We tokenize textual descriptions from both the VERIS and CVE datasets, splitting them into individual words or tokens; this crucial process facilitates subsequent analysis and feature extraction. In both datasets, we stem the words to their root or base form; for instance, "attacked", "attacking", and "attacks" are all simplified as 'attack'. This process of stemming simplifies text analysis: it treats identical words similarly--a key advantage.

**POS Tagging:** It identifies the grammatical parts of speech in textual descriptions from both datasets; this process aids in extracting specific types of information--such as affected assets or undertaken actions--from the text.

**Statistical Analysis:** By analyzing the frequency distribution of various types of cybersecurity incidents from the VERIS dataset and vulnerabilities from the CVE dataset, we gain insights into: Graduate-level punctuation—such as colons; semi-colons; and dashes—are used here to enhance readability. These provide a clear understanding on what types are most encountered by organizations in terms of security threats.

Temporal patterns and trends in cybersecurity incidents and vulnerabilities' occurrence undergo examination over diverse time intervals. This analysis actively identifies seasonal trends or long-term alterations within the threat landscape: an essential tool for trend analysis over time.

**Anomaly Detection**: We apply statistical techniques—such as z-score analysis or time-series anomaly detection algorithms—to identify anomalous patterns, outliers in both datasets; this aids us in detecting unusual cybersecurity incidents or vulnerabilities that deviate from normal behavior.

**Machine Learning Models**: Supervised machine learning models, trained on both VERIS and CVE datasets, use incident classification to categorize cybersecurity incidents into predefined groups such as incident types and impact levels. This automation accelerates the efficiency of incident management by parsing textual descriptions along with other attributes.

Unsupervised learning algorithms, trained on both datasets, cluster similar cybersecurity incidents and vulnerabilities according to their textual descriptions or other features; this aids in the comprehension of common attack patterns or vectors: a crucial aspect of understanding cyber threats.

Predictive models, developed from historical data in both datasets, forecast future trends in cybersecurity incidents and vulnerabilities. This development equips organizations to mitigate risks proactively by foreseeing potential security threats.

**Visualization Techniques:**

**Bar Charts and Pie Charts**: These visual representations swiftly depict the distribution of incident types or severity levels from both datasets; they offer an immediate overview of prevalent cybersecurity incidents and vulnerabilities.

**Time-Series Plots**: Time-series plots visualize temporal trends and patterns in cybersecurity incidents and vulnerabilities, facilitating the identification of seasonal variations, pervasive trends, and anomalies within datasets.

**Heat maps:** These visualizations illuminate the correlation matrix of incident attributes in each dataset; they provide an understanding of relationships between a multitude of variables - shedding light on factors that influence cybersecurity incidents and vulnerabilities.

Organizations, by harnessing these algorithms and analysis techniques with the VERIS and CVE datasets: can glean valuable insights into their cybersecurity landscape; identify potential threats--all while making informed decisions to fortify their security posture. Effective visualization of analytical results not only amplifies understanding but also facilitates insightful communication among all stakeholders.: These visual representations swiftly depict the distribution of incident types or severity levels from both datasets; they offer an immediate overview of prevalent cybersecurity incidents and vulnerabilities.

## 3.2 Web Scrapping

After we finish extracting data from news articles, which we got using web scraping through Bing API and Google RSS, the data that has been extracted goes through a very important step called structuring. In this process, all the information obtained is carefully organized into an understandable and planned format to make it easier for later analysis and exploration.

The data that has been gathered covers different important information like the name of article, its description, when it was published, URL and who published it. This information is acquired by employing web scraping techniques through the Bing API, as well as Google RSS feeds. Every single data point from these sources is parsed and allocated to appropriate columns in the dataset for systematic organization of data.

The "article name" column is for the news article headlines or titles. These are short and describe what the news story is about. The "description" column contains small summaries from each of these articles, giving more context and insight into them. The goal of the "date published" column is to record when these news articles were made public, which allows us to study them chronologically and identify trends over time.

Additionally, the "URL" column holds web links of articles. This is helpful for quickly finding and reviewing original sources. The last column named "publisher" records who published that information or where it came from, adding reliability to our gathered data.

When the data is structured and put in a neat form, we save it into a CSV file called "LLM\_articles.csv" for keeping and later use. This CSV file works as a single place where all the news articles are gathered, making it easier to study them deeply or examine thoroughly.

This dataset's structure allows researchers and analysts to investigate how often, what kind, and the effect of Large Language Model (LLM) attacks as they appear in news articles. With help from different analysis methods and tools, they can dig into the changing world of LLM attacks for better comprehension which aids strategies related to cybersecurity and lessening risks.

Also, the well-organized data can be put in different formats or databases other than a CSV file. This is done to match particular analysis needs and choices. No matter what kind of storage style it gets saved in, the well-structured data acts as important material for making better cybersecurity actions and handling new dangers caused by LLM attacks.

A graph with a line

Description automatically generated

**Figure 3: The above figure shows no of attacks over the time**

After performing web scraping on news articles using the Bing API and Google RSS, we can observe a significant rise in the reporting of Large Language Model (LLM) attacks after November 2022. This surge corresponds to the bigger use and popularity of ChatGPT, which is among top generative AI tools based on LLM technology. As ChatGPT started to get more recognition and was inserted into different applications and platforms, it unintentionally increased the potential of LLMs making them easier to reach everywhere.

The rise in LLM attacks after November 2022 highlights how much ChatGPT's increasing use is affecting the world of cybersecurity. Because of its high-level skills in understanding natural language, ChatGPT has made it easy for people with different knowledge levels to produce text, making content creation more equal. Yet, this even-handedness has also given wrongdoers the ability to misuse LLM technology for bad intentions such as creating misleading material or false data and aiming focused attacks.

Furthermore, the rise in LLM attacks shows how cyber threats are getting more complex and flexible. These changes are happening because of improvements made in AI technology. As ChatGPT and other tools based on LLM keep getting better, people who create threats also come up with new tactics and methods to counteract them. The rise in reported LLM attacks after November 2022 is a clear indication that we need to be more watchful and take preventive actions when dealing with new dangers.

Moreover, the increase in LLM attacks shows that ongoing supervision, study and adjustment of cybersecurity plans is necessary to diminish changing risks suitably. Businesses need to be watchful and take initiative in recognizing and handling vulnerabilities linked with LLM technology. This includes setting up strong systems for detecting threats, improving security measures, as well as providing comprehensive training programs for employees.  
  
The observed rise of LLM assaults after November 2022, in essence, highlights the intricate relationship between cybersecurity issues, malevolent intent, and technological improvements. As the use of ChatGPT and other generative AI tools spreads, it is critical for stakeholders in various industries to work together and come up with new ideas to strengthen defenses against LLM-posed cyber threats.

## 3.3 Risk Assessment

The process of systematically recognizing, assessing, and ranking possible hazards is known as risk assessment. Organizations seeking to comprehend and handle the intricacies involved in utilizing LLM technology must adopt this methodical approach. To provide a clear and simple overview of the detected hazards and their related severity levels, visual representations such as risk assessment matrices are essential to this process. These risk matrices allow stakeholders to properly prioritize mitigation activities and distribute resources where they are most required by classifying threats according to their likelihood and impact.

There are many aspects that contribute to the overall risk environment when it comes to LLM implementation. Risks, weaknesses, and outcomes are the essential elements that must be considered and analyzed. Threats can take many different forms, including cyberattacks, data breaches, and disinformation operations, which are possible sources of unintended events. To effectively create mitigation measures and implement adequate protections against possible harm, it is imperative to comprehend the nature and scope of these dangers.

Also, vulnerabilities are like weak spots or spaces in LLM implementation that could be taken advantage of by threats to compromise system's strength or data's secret keeping. These vulnerabilities might come from different places such as lack of security methods, lack of training and technical limits in the LLM framework. Recognizing and solving these weaknesses is very important for increasing total security status and lowering the chance for successful attacks.

The consequences, on the other hand, are about what might happen if something bad occurs inside an organization. The outcomes can take various shapes such as money loss, harm to reputation or not following rules and regulations along with legal obligations. When we check possible consequences of risks, it helps organizations comprehend how serious security incidents could be and do something ahead to lessen their effect.

Matrices for assessing risk offer an all-around visual picture of the risks found, putting them into categories according to how likely and severe they are. The situations with high risk, where both the chance of happening and possible results are considerable, are given priority for instant attention and reduction. However, situations with less risk might not need urgent action but still require observation and handling to avoid increasing danger.

A diagram of a risk assessment matrix

Description automatically generated

**Figure 4: Risk assessment matrix**

Risk assessment is more important than just identifying and reducing possible hazards when deploying the Large Language Model (LLM). It is a proactive approach to managing trust and reputation, compliance, prevention, and resource allocation, among other important areas.

**Proactive Prevention**: By using risk assessment, companies can see possible dangers and weak points related to the use of LLM before they become real incidents. Organizations may take proactive steps to stop safety breaches, data leaks, and other unfavorable events by carrying out comprehensive risk assessments. By adopting a proactive approach, organizations can minimize possible harm to their systems, data, and reputation by staying ahead of new threats and mitigating risks before they become serious issues.  
  
**Compliance:** Compliance is crucial for maintaining data security, privacy, and integrity in many businesses. Regulatory standards and industry best practices must be followed. For firms to identify and handle compliance obligations pertaining to the implementation of life cycle management (LLM), risk assessment is a critical tool. Risk assessments are a useful tool for enterprises to make sure that their LLM systems and procedures comply with applicable legal frameworks, like GDPR, HIPAA, and ISO standards. Stakeholder credibility and confidence are increased as well as legal and financial risks are reduced by adhering to these regulations.

**Resource Allocation**: To maximize safety measures and return on investment in LLM technology, efficient resource allocation is crucial. Organizations can more effectively arrange and direct resources—budget, manpower, and technology, for example—to the areas that most require them by using risk assessment. Organizations can strategically deploy resources to reinforce security controls, set up essential safeguards, and improve overall resilience against cyber threats by identifying high-risk areas and possible weaknesses. By allocating resources effectively, businesses can manage risks more effectively and make the best use of their limited resources.

**Trust and Reputation Management:** In today's digital landscape, where data privacy and security are major concerns for both individuals and enterprises, trust and reputation management are crucial. By displaying a dedication to cybersecurity and risk management, risk assessment is essential to preserving confidence and protecting reputation. Organizations may demonstrate to stakeholders—including partners, consumers, and regulators—that they take cybersecurity seriously and have put strong safeguards in place to protect confidential data by undertaking open and thorough risk assessments. In the end, stronger relationships with stakeholders are fostered and long-term success is ensured by this proactive approach to risk management, which also improves trust, credibility, and reputation.

### 3.3.1 Main Components of Risk Assessment:

Risk assessment is a complex process that includes analyzing several parts to find, assess, and reduce possible hazards to a system or organization. The three primary elements of risk assessment are consequences, vulnerabilities, and threats.

**Threats:** Threats are possible sources of unintended events that have the power to damage an organization or system. These can include a broad variety of situations, such as cyberattacks, natural catastrophes, mistakes made by people, or malevolent acts by people or organizations. Organizations can reduce risks and protect their assets and operations by being aware of and responding quickly to possible threats.

**Vulnerabilities:** vulnerabilities are made up of weak points or gaps in security measures that could be used by attackers to obtain unauthorized access or do harm. These vulnerabilities may be caused by out-of-date software, insufficient physical security measures, poor security procedures, or a lack of personnel training, among other things. To lessen the possibility that possible threats may take advantage of a vulnerability, companies must properly detect and resolve vulnerabilities.

**Consequences:** The influence or result of an event materializing that affects the organization's goals or resources is known as a consequence. These repercussions may take many different forms, such as monetary loss, harm to one's reputation, interruptions to business operations, legal ramifications, or injury to specific people. Organizations can reduce the effect of unfavorable occurrences by allocating resources and prioritizing mitigation activities based on their understanding of the probable repercussions of risks.

# Risk Assessment and Markov Chain

## Project Architecture

**LLM Techniques Mapped To VERIS**

The dataset "llm\_techniques\_mapped\_to\_VERIS.txt" offers a structured mapping between various cybersecurity techniques and the VERIS framework. This mapping is valuable because it clarifies how specific actions undertaken during cyber incidents, such as hacking, malware usage, physical attacks, and access misuse, correspond to broader categories of threat techniques.

The dataset systematically organises these actions under distinct technique names. Each entry includes a rationale explaining the mapping and quantifies the occurrences of these actions within the dataset, referred to as "VERIS Action Count". The data is structured into columns with headers like Technique Name, Action Name, Rationale, and VERIS Action Count. Each row represents a specific cybersecurity technique, the corresponding action, a brief explanation for the association, and the frequency of this action's appearance in the dataset. For instance, it details actions like "Acquire Infrastructure" achieved through various means like web application hacking, remote access misuse, direct malware installation, and physical theft, along with explanations and occurrence counts.

This dataset is particularly useful for cybersecurity professionals and researchers, especially those interested in the intersection of machine learning and cybersecurity. It offers a structured approach to understanding and analysing the diverse actions that can be employed in cyberattacks, how they connect to established cybersecurity techniques, and their prevalence within the dataset. This information can be instrumental in threat modelling, incident analysis, and developing defensive strategies against a wide range of cyber threats. The importance of this dataset is further emphasized by recent research on applying large language models (LLMs) to identify ATT&CK techniques within threat intelligence reports, as mentioned in a LinkedIn post by Jonathan Baker, Director & Co-founder of the Centre for Threat-Informed Defence. This research underscores the potential of AI and machine learning to enhance cybersecurity threat intelligence and analysis.  
  
  
A diagram of a system

Description automatically generated

**Figure 5: Project Architecture**

The diagram outlines a data mapping process involving various datasets and tools aimed at cybersecurity risk assessment. Key components of the flowchart include:

The use of different datasets like Census Data, VCDB Dataset and MITRE Atlas Dataset along with the sector-wise cybersecurity attacks has given a solid base for our exploration on the growing danger from Large Language Model (LLM) attacks.

Census Data is a very important part of our analysis, helping us understand industry classifications by using NAICS codes. These codes group business establishments based on their kind of economic activity. This allows us to put LLM attacks into context within certain sectors and comprehend how they might affect different industries.

The second source is the VCDB Dataset, also known as the VERIS Community Database. It is a collection that contains extensive records of security incidents. These have been categorized using a standardized model for describing and classifying cybersecurity events called VERIS Framework. This dataset provides us with information about these incidents in detail and enables our system to recognize LLM attacks from other kinds of security breaches alike.

At the heart of our inquiry is the study of cybersecurity attacks according to their sectors, which gives a concentrated view on incidents divided by industry sector. This allows us to investigate how often and where LLM attacks happen in certain areas, giving important understanding about vulnerabilities and risk characteristics specific to each sector.

Connected to the sector analysis is the MITRE Atlas Dataset. This dataset tags types of cyber-attacks like LLM attacks as Low, Likely or Moderate. MITRE, a non-profit organization aiding many U.S. government offices, offers this dataset to create consistent structure for comprehending and classifying cybersecurity hazards.

The merging of cybersecurity and machine learning (ML) has brought a fresh kind of hazard called LLM assaults. These attacks target weaknesses in machine learning systems, causing substantial danger in various fields. Our study aims to investigate how often and what outcomes are caused by LLM assaults. We plan on accomplishing this by filtering out LLM attacks from the VERIS dataset, studying where they happen most often across sectors or industries, measuring how much impact these assaults have had so far and discovering methods that could lessen their effects.

The datasets and frameworks we have examined are expected to offer useful understanding for stakeholders about how LLM attacks might change the safety environment. By studying the characteristics and frequency of these attacks in various areas, organizations can improve their planning and put strong cybersecurity actions into action to lessen risks related with LLM attacks.

### LLM Attack Filtering

We started our research by filtering LLM attacks from the large VERIS dataset. LLM attacks cover a wide range of strategies, such as adversarial examples, data poisoning, backdoor, Lan access, direct install, acquire infrastructure, Exploit vulnerability and model evasion etc., that are intended to compromise machine learning (ML) systems. Using the MITRE ATT&CK framework, which offers an extensive taxonomy of adversarial approaches, we were able to recognize and extract strategies and techniques relevant to LLM from the VERIS dataset.

**Analyzing Distribution Across Sectors**:

After compiling a list of LLM assaults, we looked at how they were distributed among various industries. Information technology, retail, public services, healthcare, and other industries were shown to be especially vulnerable to LLM attacks. These industries frequently depend significantly on machine learning (ML) systems for vital tasks, which makes them attractive targets for attackers looking to sabotage operations or obtain private information.

We looked at the frequency of each assault inside each sector to gain further insight on the distribution of LLM attacks. This detailed investigation offered insightful information about the risks that various industries face. Attacks on machine learning infrastructure and artifacts, for instance, were common in many industries, showing how widespread these vulnerabilities are. On the other hand, industries with distinct features, including public services and healthcare, had to deal with dangers that were particular to their operational environments.

### Quantifying impact

We measured the frequency of each attack and computed its percentage in relation to the overall number of the first 25 attacks to evaluate the effect of LLM attacks. We were able to rank attacks according to how common and important they were in the overall danger landscape thanks to these criteria. Furthermore, we gave each attack a score determined by its proportion, and this gives each attack a numerical representation of its relative impact.

Percentage formula

Percentage= (Total Count of All Attacks/Count of Specific Attack)

Risk score formula

Score= ((Total Count of All Attacks/Count of Specific Attack) \*percentage) \*100

A screenshot of a spreadsheet

Description automatically generated

**Figure 6: Sample of Count, percentage, and score calculation for health care sector**

**Finding Mitigation Strategies**:

Equipped with knowledge about the frequency and consequences of LLM attacks, we focused on possible mitigation techniques. A comprehensive defensive strategy must include proactive defense measures including strong authentication mechanisms, anomaly detection systems, and adversarial training methods. Through the incorporation of these safeguards into ML development and deployment workflows, entities can enhance their resistance against LLM assaults and lessen their consequences.

Moreover, cooperation and information exchange are essential in thwarting LLM attacks. To keep one step ahead of competitors, stakeholders can share best practices, insights, and threat intelligence by establishing partnerships across government agencies, businesses, and academia. Moreover, industry standards and regulatory frameworks operate as stimulants to promote cybersecurity investments and guarantee adherence to ML security best practices.

To effectively counter the constantly changing threat landscape provided by LLM attacks against machine learning (ML) systems across multiple industries, our goal is to develop a strong risk assessment tool. Our risk assessment tool is based on the important insights we have gained about the frequency, distribution, and impact of these attacks using the VERIS dataset and the MITRE ATT&CK architecture. Through the integration of the comprehensive taxonomy of adversarial techniques provided by the MITRE ATT&CK framework with data from the VERIS dataset, which records real-world security incidents, our risk assessment tool will provide a comprehensive understanding of LLM attack vectors and their potential implications. Through this connection, enterprises may proactively assess and manage risks by identifying patterns, trends, and emerging threats within the ML threat environment.

**Risk Assessment Tool:** Connected to "Mapping LLM Attacks," this indicates that the diagram is a representation of the process for assessing risks associated with these attacks, employing a specialized tool.

**Mapping LLM Attacks to Risk Assessment Tool:** This connection demonstrates a sequential flow of analysis from the categorization of cyberattacks by sector, leading to their mapping, and culminating in the assessment of associated risks using a dedicated tool.

### Risk Assessment Hospital Data Analysis

The data gives a thorough look into possible dangers of cybersecurity in hospitals, highlighting Large Language Model (LLM) attacks. The dataset shows many attack vectors and incidents, even detailing certain occurrences tied to machine learning (ML) models.

The dataset has notable entries that show different methods of attack like "Acquire Public Infrastructure", "Exploit vuln" (vulnerability), "Phishing", and "Pretexting". Each entry includes data points, probably representing the number of times it happens or a linked risk measurement. For example, the phrase 'Acquire Public Infrastructure' is noted as having 1,590 instances. This suggests a high occurrence or risk value for this specific attack method.

There are certain dangers related to ML models, such as "Discover ML Model Family" or "Erode ML Model Integrity". These titles suggest possible weak spots in the hospital's operations that rely on machine learning, emphasizing how important it is to have strong cybersecurity protections for protecting delicate health data and maintaining the trustworthiness of ML systems.

In general, the data shows us how crucial it is for hospitals to do risk assessment and protect themselves against LLM attacks. By finding out about possible weaknesses and dangers, hospitals can improve their cybersecurity position and decrease the chances of data breaches or other security incidents. Sheet1 Overview:

**NAICS:** Indicates the code "622" which relates to the health care and social assistance sector, specifically hospitals.

**SECTOR:** Labeled as "HOSPITALS", confirming that the data pertains to this sector.

ESTIMATED NO. OF BUSINESS: Lists 7,465 estimated businesses in this sector.

TOTAL NO. OF BREACHES: Documents 2,329 total breaches.

Acquire Public Infrastructure: 1,590 instances.

Exploit vuln: 12 instances.

Phishing: 84 instances

Pretexting: 9 instances

This data suggests a focus on various attack methods, possibly indicating the frequency or the identified risk of these methods within the hospital sector.

## Risk Analysis-Threat

### Mapping VERIS Attack Techniques with MITRE ATLAS Attack Techniques

As the investigation progressed, it became apparent that the Veris database was deficient in precise information concerning attacks classified as Large Language Models (LLM). To rectify this shortcoming, a process was devised to associate each attack identified in the MITRE ATLAS framework with corresponding attack vectors recorded in the Reciprocal Veris Community Database.

The MITRE ATLAS framework offers a standardized language for delineating adversary behavior and provides a systematic taxonomy of cyber threat tactics and techniques. On the other hand, the Reciprocal Veris Community Database functions as a repository for incident data, which includes a multitude of characteristics including assault vectors, demographics of victims, and metrics of impact.

A bridge mechanism was implemented to address the lack of LLM-specific data in Veris by establishing correspondences between MITRE ATLAS techniques and corresponding attack vectors in the Veris database. Through the utilization of the all-encompassing framework furnished by MITRE ATLAS and the empirical data obtained from the Veris database, this mapping exercise promotes a more intricate comprehension of the cyber threats presented by Large Language Models. In essence, this methodology empowers scholars and professionals to acquire knowledge regarding the frequency and attributes of breaches involving Large Language Models, even considering the intrinsic constraints of specific data sources. Furthermore, it emphasizes the significance of incorporating varied datasets to develop a comprehensive understanding of the cyber threat environment; this improves the efficiency of endeavors related to detecting, analyzing, and mitigating threats.

A diagram of a computer system

Description automatically generated

**Figure 7: Mapping of LLM Attacks**

The MITRE ATLAS Methods:

Black-Box Transfer: Emphasized as a technique through which malicious actors can compromise systems without requiring an exhaustive comprehension of their inner workings. By identifying and exploiting concealed vulnerabilities, this method gains access to a system or network.

Mapping to VERIS Attacks:

* Vector for mapping VERIS attacks to action.hacking.Desktop software for sharing: This pertains to the utilization of software that enables malicious remote control over computing systems.
* action. variety of malware. Backdoor or C2: This pertains to the covert implementation of backdoor access mechanisms into a system, which enable unauthorized manipulation or larceny of data.
* action. variety of malware. The backdoor technique is designed to create hidden points of access into a system or network in order to circumvent standard security measures.

Objective of Mapping:

By correlating specific techniques described in MITRE ATLAS with actual attack vectors recorded in VERIS, the mapping function is performed. This methodology enhances comprehension regarding the practical implementation of theoretical techniques in the context of intrusions.

As a result, by utilizing an integrated data-driven framework, this methodical approach improves the capacity to anticipate, identify, and mitigate cyber threats with greater precision and insight into adversary actions.

### Categorizing LLM Attacks into different types

LLM Security Threats are categorized into five groups based on the analysis, with each group reflecting a distinct attack vector and method:

A diagram of a computer program

Description automatically generated

**Figure 8: Mapping of LLM Attacks**

**1. Attack Preparation and Development**

This category includes techniques that involve the preparatory steps for conducting more targeted attacks, including the development of capabilities and the acquisition of necessary tools or data.

* Acquire Infrastructure
* Acquire Public ML Artifacts
* Develop Capabilities
* Discover ML Artifacts
* Discover ML Model Family
* Discover ML Model Ontology
* ML Artifact Collection
* ML Development Workspaces
* Obtain Capabilities
* Train Proxy via Gathered ML Artifacts
* Train Proxy via Replication
* Use Pre-Trained Model

**2. Data and Model Targeting Attacks**

These attacks directly target data integrity or manipulate ML models to corrupt, exploit, or steal data and model functionality.

* Adversarial ML Attacks
* Adversarial ML Attack Implementations
* Backdoor ML Model
* Black-Box Transfer
* Craft Adversarial Data
* Create Proxy ML Model
* Data
* Data from Information Repositories
* Data from Local System
* Datasets
* Erode ML Model Integrity
* Evade ML Model
* Extract ML Model
* Full ML Model Access
* Infer Training Data Membership
* Invert ML Model
* Poison ML Model
* Poison Training Data
* Publish Poisoned Datasets

**3. Resource and Service Disruption**

This category focuses on attacks that primarily aim to disrupt services or exhaust resources, whether computational or financial.

* Black-Box Optimization
* Cost Harvesting
* Denial of ML Service
* GPU Hardware
* Spamming ML System with Chaff Data
* White-Box Optimization

**4. Exploitation and Exfiltration**

These techniques exploit system vulnerabilities to perform unauthorized actions or exfiltrate data, including financial and intellectual property theft.

* Command and Scripting Interpreter
* Exploit Public-Facing Application
* Exfiltration via Cyber Means
* Exfiltration via ML Inference API
* Financial Harm
* ML Intellectual Property Theft
* ML Model Inference API Access
* ML Software
* ML Supply Chain Compromise

**5. Deception and Intrusion**

These attacks involve deception tactics to gain unauthorized access or gather information, often using social engineering or phishing.

* Consumer Hardware
* Establish Accounts
* Manual Modification
* Phishing
* Physical Environment Access
* Search Application Repositories
* Search for Publicly Available Adversarial Vulnerability Analysis
* Search for Publicly Available Adversarial Vulnerability Research
* Search for Victim's Publicly Available Research and Development
* Search for Victim's Publicly Available Research Materials
* Search Victim-Owned Websites
* Spear phishing via Social Engineering LLM
* Unsecured Credentials
* User Execution
* Valid Accounts

**A table of heatmap of lm threats across sectors

Description automatically generated**

**Figure 9: Visual shows the threats across different sectors**

**Analysis by Sector**:

This dataset delves deeper into the hazards associated with several sectors, such as finance, public, educational, retail, information, manufacturing, administrative, professional, and transportation. The frequency and types of LLM security threats that are encountered are indicative of the distinct challenges and vulnerabilities that each sector faces.

Important Results in different sectors:

The financial sector highlights the need for increased security measures to protect sensitive information and transactions. Similarly, the public sphere and government systems face numerous hazards, emphasizing the importance of fortifying public data. The education realm also confronts challenges in defending against planned attacks, highlighting the critical obligation to safeguard educational assets. The retail industry is vulnerable to attacks that defile models and steal data, emphasizing the need for strict security procedures. In the healthcare sector, threats to data credibility and patient confidentiality are prevalent, requiring resilient strategies to protect sensitive medical information and maintain the reliability of healthcare systems. The transportation sector faces dangers related to infrastructure vulnerabilities and operational disruptions, underscoring the importance of comprehensive security measures to ensure the integrity and trustworthiness of transportation networks. The professional and administrative industries face various threats to sensitive information and valuable resources. It is crucial to prioritize the protection of organizational assets and maintain operational continuity amidst the constantly evolving cyber risks.

A graph showing the number of attacks

Description automatically generated

**Figure 10: Visual shows the count of each attack in different sectors**

The incidence of LLM (Large Language Model) attacks varies significantly throughout sectors, as our data shows. The public sector has the highest number of instances, followed by education and finance, while transportation shows the least vulnerability. This discrepancy highlights the infrastructure that may be vulnerable in government and educational institutions, as well as the ongoing threats that financial institutions must contend with. The transportation sector may have a lower reliance on digital infrastructure, or effective measures as seen by the relatively low occurrence of LLM attacks. However, to lessen the impact of LLM threats and protect vital services, data, and infrastructure, proactive steps are necessary in all industries. Effective collaboration and information exchange among sectors are essential for enhancing resilience and managing changing threats.

### Probabilities Threat Calculation

Following the classification of the 83 Large Language Model (LLM) assaults into five discrete categories, the frequency of these five distinct threats within each sector was computed. The process entailed calculating the frequency of each hazard category in different sectors of our dataset.

Following that, the probability of each sector being impacted by a specific LLM threat category was calculated. The derivation of this calculation involved the division of the count of a particular threat category identified within a sector by the overall count of LLM attacks detected in that sector. By employing this methodology, we were able to evaluate not only the frequency of LLM attacks across various industries but also the probability of confronting categories of threats within each sector. Through the process of quantifying these probabilities, we acquired valuable insights regarding the comparative risk presented by distinct LLM threat categories in multiple sectors. This enabled us to make more informed decisions and prioritize security measures more effectively.

A screenshot of a data table

Description automatically generated

**Figure 11: Threat Score Calculation**

After ascertaining the likelihood of interacting with each LLM threat category in a specific sector, we generated a composite metric known as the "threat score" for that sector. The threat score functions as a quantitative indicator of the comprehensive peril presented by LLM attacks in a particular industry, considering the likelihoods of encountering different threat classifications.

To determine the hazard score for a specific sector, the subsequent formula was utilized:

Threat Score=1−i=1∏n​(1−Probabilityi)

* Probabilityi represents the probability of encountering theith LLM threat category within the sector.
* n denotes the total number of distinct LLM threat categories.

The formula utilizes the complement of the probability that none of the threat categories materialize to determine the probability of at least one instance of any LLM threat category occurring within the sector. Through an iterative process of multiplying the complement probabilities of specific threat categories, the cumulative probability of confronting at least one LLM threat within the sector can be determined.

The threat score that is generated offers a comprehensive evaluation of the sector's vulnerability to LLM assaults, considering the combined influence of various threat categories. Increased threat scores signify a more severe susceptibility, which necessitates heightened awareness and focused mitigation approaches within the impacted industry. On the other hand, threat scores that are lower indicate a comparatively diminished level of risk, although they still require continuous surveillance and proactive security protocols to uphold resilience in the face of potential threats.

## Risk Analysis-Vulnerability (Markov Chain Analysis)

The analysis of event sequences where the probability of changing from one state to another depends only on the present state and is unaffected by prior states is possible with Markov Chain modeling, a probabilistic framework. Essentially, it depicts a random procedure that is distinguished by a collection of conditions and the likelihood of transitions among them. Markov Chains are extensively used in many disciplines, such as statistics, engineering, economics, mathematics, and most importantly, the study of sequential data like time series and behavior patterns. By allowing for predictions, simulations, and decision-making based on the probabilistic nature of state transitions, this modeling technique is useful for comprehending the dynamics of systems that evolve over time.

**Markov Chain Modeling in Cybersecurity:**

Cybersecurity requires an understanding of the dynamics of LLM (Living off the Land) related assaults. These attacks, which are frequently covert and hard to identify, can seriously hurt companies in a variety of industries. Using the VERIS dataset, we can examine how LLM attacks shift over time between various states and sectors by utilizing Markov Chain modeling.

Markov Chain modeling, a potent mathematical framework for researching stochastic processes, is at the core of our approach. Markov Chains play a crucial role in the modeling of sequential events in cybersecurity, since they allow us to understand how LLM attacks evolve over time. In these models, the probability of transitioning to a future state is exclusively dependent on the current state.

The VERIS dataset, a goldmine of cybersecurity event reports covering different industries and time periods, is where our trip starts. We carefully sorted through this data, classifying threats associated to LLM according to their industries and states. State 0 (1-2) assaults, State 1 (3-4) attacks, State 2 (5-6) attacks, and State 3 (7-9) attacks are the states that are defined by the length of the attacks in this case. This fine-grained classification serves as the basis for our next Markov Chain investigation.

**Building the Markov Chain: Probability Transition Matrices**   
  
After carefully organizing our data, we start building the transition probability matrices for every sector. The likelihood that LLM attacks will shift between states in a particular year is captured by these matrices. These matrices provide insight into the dynamic nature of cyber threats within industries, with each entry denoting the probability of an assault transitioning from one state to another.

### Analyzing Sectors: Identifying Trends and Patterns

We now focus on the sector-specific transition probability matrices, where we identify trends and decipher complex patterns in the evolution of LLM attacks over time. Through examining the likelihoods of states changing, we uncover important information about how attackers in various sectors operate. Furthermore, pinpointing industries with high likelihoods of changing into states linked to extended attack times reveals weaknesses that can be fixed.

**Education Sector:**  
The number of attacks seen in the data is used to illustrate the probability of the education sector shifting between states. The likelihood of changing to a certain subsequent state is represented by each column in the matrix, while each row denotes the current state. When looking at the first row, for example, there is no chance of moving from State 0 (which represents 0 to 1 attacks) to any other state (all transition probabilities are 0.0). Approximately 31.25% of people will remain in State 1, 6.25% will move to State 2 (representing 6 to 10 attacks), and 62.5% will move to State 3 (representing 11 attacks). The second row shows the probabilities of moving from State 1 (representing 1 to 5 attacks) to other states.

A number in a row

Description automatically generated with medium confidence

**Figure 12: Transition Matrix for Education Sector**

**Information Sector:**

Based on the number of attacks, the information sector shows the likelihood of changing states. Every row indicates the present state and the probability that it will change. For instance, there is a 25% probability of moving from State 0 (which represents 0 attacks) to State 1, and a 75% chance of moving from State 0 to State 3 (which represents 11 or more assaults). In a similar vein, the probabilities in other rows depict the information sector's transition dynamics, offering perceptions of attack trends and vulnerabilities unique to this industry.

A number on a white background

Description automatically generated

**Figure 13: Transition Matrix for Information Sector**

**Manufacturing Sector**:

Based on the number of attacks, the manufacturing sector's transition matrix shows the likelihood of moving between states. A present condition is indicated by each row, along with the probability that it will change. There is a 100% probability, for example, of moving from State 0 (which represents 0 attacks) to State 3 (which represents 11 or more attacks). On the other hand, there are three potential outcomes from State 1 (which in turn represents 1 to 5 attacks): 20% chance of remaining in State 1, 8% chance of traveling to State 2 (which represents 6 to 10 attacks), and 72% possibility of visiting State 3. By illuminating potential weaknesses and attack patterns unique to the manufacturing sector, these probabilities provide insights into the processes of transition within that sector.

A number of numbers and symbols

Description automatically generated with medium confidence

**Figure 14: Transition Matrix for Information Sector**

**Professional Sector:**  
The professional sector based on the number of attacks. The states are categorized based on the range of attacks they represent: 0 attacks for State 0, 1 to 5 attacks for State 1, 6 to 10 attacks for State 2, and 11 or more attacks for State 3. The likelihood of moving between these states is shown in the transition matrix. For example, there is a 100% possibility of going from State 0 to State 3 (which represents 11 or more attacks). On the other hand, from State 1, there's roughly a 16.67% chance of remaining in State 1, a 20.83% chance of going to State 2, and a 62.5% chance of going to State 3.

A number with numbers on it

Description automatically generated with medium confidence

**Figure 15: Transition Matrix for Professional Sector**

**Retail Sector**

Based on attack counts in the retail industry, various attack count ranges define states: Whereas State 0 denotes no attacks, States 1 through 3 denote attacks from 1 to 5, States 2 and 3 denote attacks from 6 to 10. These states' odds of changing are shown in the resulting transition matrix. For example, there's a 50% possibility of going from State 0 to State 3 (which represents 11 or more attacks). In contrast, there are roughly 31.82% of State 1 residents who choose to remain in State 1, 4.55% of State 1 residents who want to go to State 2, and 63.64% of State 1 residents who choose to move to State 3.

A number with numbers on it

Description automatically generated with medium confidence

**Figure 16: Transition Matrix for Retail Sector**

**Transportation Sector:**

Based on attack counts in the transportation sector, where several attack count ranges are used to classify states: States 0 through 3 denote no assaults, States 1 through 5 denote one to five attacks, States 2 through 10 denote six to ten attacks, and States 3 denote eleven or more attacks. The odds of changing between these states are depicted in the resulting transition matrix. For instance, there is a 100% probability of moving from State 0 to State 3 (which represents 11 or more attacks). On the other hand, from State 1, there is a 25% likelihood of remaining in State 1, an 18.75% possibility of advancing to State 2, and a 56.25% chance of reaching State 3.

A number and numbers on a white background

Description automatically generated

**Figure 17: Transition Matrix for Transportation sector**

**Interpreting Findings: Strengthening Cyber Defense**   
  
Organizations may strengthen their cyber security with accuracy and predictability if they have the insights from our Markov Chain study. Those sectors that are more likely to experience long-term LLM attacks are considered vulnerable and should be the focus of specific security improvements. Moreover, the identification of changes in transition probabilities signals the arrival of novel threats and facilitates the development of proactive mitigation tactics that adapt to changing assault environments.

When it comes to LLM-related threats, Markov Chain modeling is a lighthouse, providing a methodical way to break down their complexity. Organizations can increase their resistance to cyberattacks, identify new vulnerabilities, and steer toward effective cyber defense plans by utilizing the VERIS dataset and figuring out transition probabilities. To keep ahead of attackers and protect important assets in the constantly moving field of cybersecurity, constant awareness and analysis are essential.

## Risk Analysis-Consequence

In this stage of the risk assessment, it was intended to quantify the impact of cyber threats on various industries in terms of data confidentiality, availability, and integrity (CIA). Comprehending the potential harm and disturbance induced by cyber incidents necessitates cognizance of these three fundamental elements. The information utilized in this analysis was obtained from the VERIS database, which classifies cyber incidents according to their influence on availability, confidentiality, and integrity. The rationale behind selecting these metrics is that they serve as all-encompassing indicators of the extent of damage that a cyber threat can cause.

Normalization and computation:

* The CIA conducted a count of incidents that compromised confidentiality, availability, and integrity for each sector that was represented in the VERIS database.
* Normalization: The quantities were subsequently normalized for each metric within each sector using a scale ranging from 0 to 1. By means of this normalization procedure, the metrics are rendered comparable across sectors, irrespective of the magnitude or frequency of occurrences.
* The final consequence score for each sector was determined by averaging the three normalized scores, namely confidentiality, integrity, and availability. The average yields a solitary consequence score which symbolizes the collective potential ramifications of cyber threats on every sector.

A screenshot of a graph

Description automatically generated

**Figure 18: Consequence Calculation for LLMs**

## Final Risk Assessment

A chart with numbers and a number of indicators

Description automatically generated with medium confidence

**Figure 19: Final Risk Assessment**

Elements comprising the Risk Score:

The risk assessment instrument evaluates the risk across multiple sectors by integrating three primary scores:

* The Threat Score quantifies the probability that an attack will transpire within a given sector. It is computed by employing the formula. The probability of the ith form of attack is denoted by P(Attack i).
* Vulnerability Score: This score indicates, considering the current security measures in place, the degree to which a sector is vulnerable to cyber threats. The value is obtained by summing the probabilities of vulnerabilities present in states that are more prone to exploitation, such as State 2 and State 3 within the given methodology.
* The consequence score evaluates the potential ramifications of an attack on a specific sector, with a specific emphasis on data availability, confidentiality, and integrity. To derive the consequence score for the sector, each of these factors is quantified, normalized on a scale of zero to one, and then averaged.

Method of Calculating the Final Risk Score

An average of the threat, vulnerability, and consequence scores is used to calculate the final risk score for each sector. This methodology offers a comprehensive assessment of the risk environment by integrating the probability of an intrusion, the vulnerability of the industry to harm, and the potential gravity of the repercussions.

Insights Derived from the Data

* Different industries, including finance, healthcare, and information, exhibit diverse threats, vulnerability, and consequence scores, which are indicative of their distinct risk profiles.
* Notwithstanding its moderate consequence score, the financial sector exhibits a high vulnerability score, which culminates in a comparatively elevated ultimate risk score.
* The healthcare industry demonstrates the greatest consequence score, which has a substantial impact on its final risk score, which is among the highest of all industries.

This approach facilitates a thorough evaluation of risks by incorporating various facets of cybersecurity threats. It aids in the prioritization of sectors according to their distinct risk profiles and facilitates targeted risk management strategies.

# Calculating Normalized Risk Factor on Different Sectors

**Analyzing the data**   
Data analysis starts with a thorough review of two main datasets: one from the VERIS dataset, which provides information about how various sectors are distributed across each state, and another that includes event reports from a wide range of industries. This comprehensive examination forms the basis of the risk assessment process. Through a thorough analysis of these datasets, researchers can obtain important insights into historical event records and industry architecture. These kinds of insights are essential for making well-informed decisions and creating focused risk-reduction plans. By means of this meticulous analytical approach, interested parties can obtain an intricate understanding of the intricacies of sectoral infrastructure and identify trends in historical event patterns, establishing the foundation for subsequent research projects and risk mitigation campaigns.

**Determining the Risk Aspect**   
A crucial step in the risk appraisal process is creating a thorough risk factor. This feature aims to summarize the most important aspects of risk, such as the frequency of occurrences and the density of possible targets. A formula is used to quantify risk:

The risk factor can be calculated as follows:

Risk Factor= (Number of Incidents/Total Incidents) × (Number of Firms/Total Firm) ×Weighting Factor

By dividing the total number of organizations and the frequency of events by their respective totals, this method ensures a fair evaluation of risk. In addition, stakeholders can modify the relative significance of different risk components according to their own objectives and settings thanks to the flexibility offered by the weighting factor.

**Data Processing**   
To start computing risk scores, the preparation of data is crucial for getting accurate and dependable analysis. Any disputes or inconsistencies in datasets, like differences in state names or values that are not present, need thorough handling before beginning the computation process. The results of analysis depend on how well data quality problems are solved and formats made uniform. By processing data with great care, researchers can make sure that the analysis is done using dependable and steady information. This increases the trustworthiness of final risk assessment.

**Determining Risk Scores**   
The risk factor algorithm that we made is used to calculate the risk scores for each state. This number shows how much risk there is in a particular area, and it's based on different factors like density of population, climate conditions or economic activity. The use of this method gives a clear numerical value which can be compared with other states' risk levels – making it easier to understand and compare risks across locations. When the data from both datasets are combined using this algorithm, researchers can get an idea about how probable incidents are in every state as well as their comparative susceptibility. Risk scores are useful in decision-making processes. They help stakeholders decide where to put their resources first and carry out specific measures for reducing risks in areas that need it the most.

**Interpretation and Analysis**

After getting the risk ratings from calculations, it is important to interpret and analyze them deeply. This means looking at how risk is spread across different places, and understanding trends or patterns that are hidden inside. Through this interpretation process, people who have interest in the subject can find areas with more danger and make suitable plans for reducing risks depending on specific situations. In addition, when stakeholders deeply examine the results of analysis, they might understand what causes certain risks and where there are chances for actions. This can improve how well risk management works.

**Recommendations**

From the analysis and practical advice, stakeholders create good strategies to reduce risk. These suggestions might include taking specific actions like improving how incidents are responded to, strengthening cybersecurity methods, or boosting health care structures. By dealing with recognized risks and weak points in advance, stakeholders can improve their ability to endure difficulties while lessening possible harm from bad happenings. Furthermore, with continual observation and assessment, stakeholders can modify their risk management tactics as dangers change and new difficulties arise. This helps in promoting an environment that constantly enhances itself and stays ready.

**Visualizations**  
Risk scores are regionally represented using visual mapping approaches to aid in the comprehension and dissemination of risk assessment results. Stakeholders can identify areas that need more attention or intervention by clearly comprehending the spatial distribution of risk when risk factors are visualized on a map.

**Healthcare Sector**

The process of risk assessment in the healthcare sector is methodical and involves several steps, such as reviewing data, creating risk factors, preparing data, computing risk scores, analyzing, and creating suggestions. By employing methodical techniques and making use of easily available data, stakeholders can gain a substantial comprehension of the risks and vulnerabilities present within the healthcare system.A map of the united states

Description automatically generated

**Figure 20: The above map indicates the Normalized Risk factor in the healthcare sector across US**

**Finance Sector**

The finance industry uses a methodical approach for risk assessment, which includes data preparation, risk factor formulation, analysis, interpretation, and recommendation generation in addition to comprehensive data analysis. Through the careful implementation of these approaches and the utilization of readily available data, interested parties can obtain important understandings of possible risks and weaknesses in the banking system.

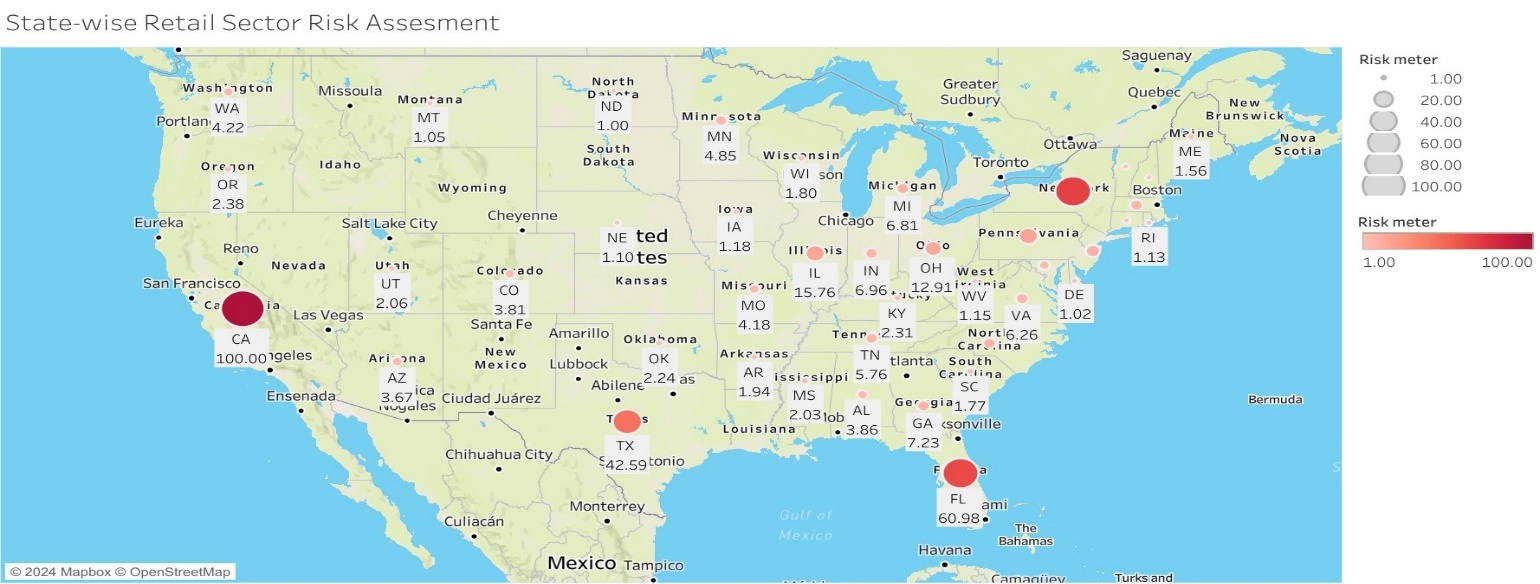
A map of the united states

Description automatically generated

**Figure 21: The above map indicates the Normalized Risk factor in the finance sector across US**

The above map indicates the Normalized Risk factor in Finance sector across US.

**Retail Sector:**  
Risk assessment in the retail industry is carried out using a systematic methodology that includes several crucial phases, such as data analysis, risk factor identification, data preparation, risk score computation, analysis, interpretation, and recommendation generation. Through the methodical use of these approaches and the utilization of accessible data, stakeholders can acquire a thorough comprehension of potential risks and weaknesses present in the retail environment.



**Figure 22: The above map indicates the Normalized Risk factor in the retail sector across US**

**The above map indicates the Normalized Risk factor in Retail sector across US**

**Professional Sector:**

Risk assessment is a systematic procedure used in the professional industry that entails several sequential processes, such as data analysis, risk factor identification, data preparation, risk score computation, analysis, interpretation, and proposal formulation. By employing systematic approaches and leveraging pertinent information sources, stakeholders can gain a sophisticated comprehension of possible risks and weaknesses in the workplace. Using efficient risk assessment methods is essential to improving readiness, flexibility, and preventive risk management plans in the face of changing market conditions, legal requirements, and professional services environments.

A map of the united states

Description automatically generated

**Figure 23: The above map indicates the Normalized Risk factor in the professional sector across US**

**The above map indicates the Normalized Risk factor in professional sector across US**

**Public Sector:**

In the public sector, risk assessment is carried out methodically using a set of clearly defined procedures, which include data analysis, risk factor identification, data preparation, risk score computation, analysis, interpretation, and recommendation formulation. Through adherence to established processes and utilization of accessible data sources, stakeholders can get significant insights on potential risks and vulnerabilities that are inherent in operations and services provided by the public sector. To improve government agencies and institutions' resilience, responsiveness, and proactive risk management, effective risk assessment procedures are essential.

A map of the united states

Description automatically generated

**Figure 24: The above map indicates the Normalized Risk factor in the public sector across US**

**The above map indicates the Normalized Risk factor in public sector across US Manufacturing Sector:**

Risk assessment in the manufacturing industry is done in a methodical manner using a number of systematic stages, such as data analysis, risk factor identification, data preparation, risk score computation, analysis, interpretation, and suggestion formulation. Stakeholders can obtain important insights into potential risks and weaknesses present in manufacturing processes and operations by following established approaches and utilizing accessible data sources. In manufacturing facilities and supply chains, resilience, efficiency, and proactive risk management are improved with effective risk assessment techniques.

A map of the united states

Description automatically generated

**Figure 25: The above map indicates the Normalized Risk factor in the manufacturing sector across US**

**The above map indicates the Normalized Risk factor in Manufacturing sector across US**

California consistently has the highest risk of danger when it comes to cyber threats and attacks in different industries or geographical areas, according to the VERIS dataset. Following closely behind with high degrees of threat are New York and Texas. Conversely, Midwesterners regularly exhibit lower cyber risk characteristics. This variation draws attention to the complicated cybersecurity environment that exists in different places and is influenced by multiple factors:   
  
**Population Density**: States with higher densities, like Texas, California, and New York, are more vulnerable to attacks since there is a greater surface area and a greater number of potential targets. Conversely, fewer cyber-attacks occur in the Midwest due to its lower population density.

**Economic Activity**: Cybercriminals seeking to profit from financial systems, intellectual property, and sensitive data are paying heed to the robust economic activity in areas like Texas, California, and New York. Cyberattacks target the manufacturing and banking sectors. On the other hand, the Midwest's economic circumstances might make cybercriminals less interested in there, which might result in fewer cases occurring there.

**Regulatory Environment: Variations in cybersecurity laws and procedures could impact the frequency and gravity of cyberattacks.**

States with stringent cybersecurity regulations may see fewer incidents, whereas states with lax regulations may be more susceptible to them.   
  
**Technological Infrastructure**: States like California and New York may experience more cyber problems due to their active role in the progress of technology. This is a result of the widespread usage of digital platforms and new technology in these areas. However, the Midwest might not be as aware of technological advancements as other regions, which might result in fewer cyberattacks.

Midwest stakeholders can strategically prioritize fortifying cybersecurity measures to uphold their comparatively lower risk profiles. This may involve allocating resources towards cybersecurity infrastructure, enacting policies and training schemes, and cultivating partnerships with law enforcement and cybersecurity entities. Conversely, stakeholders in California, New York, and Texas must prioritize resilient risk mitigation techniques to combat heightened cyber threats. This could involve implementing cutting-edge cybersecurity solutions, conducting regular security evaluations and audits, and fostering a culture of awareness and preparedness for cyber-attacks. Additionally, promoting collaboration and information sharing among states using datasets such as VERIS can significantly bolster national readiness and resilience against cyber threats. States can enhance their cyber defenses and diminish the impact of cyber incidents by collectively sharing knowledge on cyber threats, attack vectors, and best practices for responding to such incidents.

**Understanding the Landscape of Large Language Model Security Threats: An Analysis Across Sectors**

Large Language Models (LLMs) are now regarded as highly effective natural language processing tools, offering up a wide range of applications, including text production and chatbots. Their extensive use, meanwhile, also prompts worries about possible abuse and weaknesses in security. To do a thorough evaluation of the LLM security threats present in various industries, we utilize a dataset obtained from the VERIS framework. This dataset offers insights on the complex nature of security risks related to these models by classifying LLM threats into five different groups.

# Findings

Markov chain analysis was utilized to reveal the intricate levels of transition probabilities or risks between different states within each sector. This methodological approach was facilitated by transition matrices, which provided a probabilistic framework for modeling state changes over time. By assessing the likelihood of transitioning between states that represent different conditions or situations within a sector, individuals gained valuable insights into the evolving nature of risks. Organizations have been empowered by this concept, enabling them to proactively manage risks, optimize resource allocation, and strategically intervene in accordance with specific risk categories within their respective industries. By employing Markov chain analysis and leveraging web scraping techniques on Bing API and Google RSS feeds, a comprehensive threat assessment was conducted. This endeavor revealed a concerning trend: the occurrence of Large Language Model (LLM) attacks surged in November 2022, coinciding with the launch of ChatGPT. To enhance threat assessment, we have categorized LLM attacks into five distinct groups based on their methods and impacts. This comprehensive classification enables interested parties to better understand the various manifestations of LLM attacks. They were able to develop targeted strategies for risk reduction and resource utilization, thanks to this development. By deeply comprehending the nuances of each attack group, organizations could enhance their defense mechanisms and proactively anticipate emerging threats.

Threat assessments for each state were very important because they showed which areas had more danger. California became a main place for LLM attacks, with high threat level of this state. Texas and New York followed closely in terms of risk profile. The differences in regions emphasized the need for managing risks at local levels and being more watchful in areas with high-risk possibilities.

The combination of Markov chain analysis, web scraping for threat intelligence, and detailed threat assessments has allowed those involved to obtain useful understanding about the changing risk environment in every sector. With help from these data-based understandings, organizations can make themselves more resistant towards new threats. This helps protect their possessions, image and ability to function amidst an unpredictable LLM attack scene.

# Summary

In the end of our research, we have explored the risks present in Large Language Model (LLM) attacks found in different sectors and areas. Our investigation involved a detailed study of many important datasets such as VERIS dataset for cybersecurity incidents, MITRE ATT&CK database to identify LLM specific attacks, and census dataset for industry classification. We tried to use these data sources so we can understand more about how complicated LLM attack patterns are and what they mean for various industries along with their geographic locations. The core part of our method was Markov chain analysis which is a very useful tool to understand how likely or risky it is moving from one state to another within each sector. Using transition matrices, we could create a model that shows how risk alters as time passes. This is very useful for stakeholders to understand changes in the threat environment. Also, the threat assessment was completed by combining Markov chain analysis with web scraping through Bing API and Google RSS feeds. This process found a significant rise in LLM attacks after November 2022, which matched up with when ChatGPT was introduced. The classification of LLM attacks into specific groups based on how they are done, and their outcomes helped stakeholders to understand better the different forms that LLM attacks can take. This sorting allowed the creation of specific plans for risk lessening and resource use, improving organizations' defense strength and ability to predict threats in advance. Moreover, we also included state wise threat assessments in our study. This is an important part because it helps to point out areas where the risk levels are increased. We tried to show this distribution of LLM attacks on a map-like graph without showing exact results from these assessments. By understanding the way LLM attacks spread across space, groups can adjust their strategies for managing risks and distribute resources more efficiently to lessen dangers in places with high risks.

The combination of Markov chain analysis, web scraping for threat intelligence and detailed threat assessments has provided stakeholders with useful understanding about LLM attacks' dynamics in different sectors and states. This knowledge helps organizations to protect themselves from new dangers. It's crucial because this kind of attack environment becomes more unstable over time, threatening their resources, image, as well as business flow.

# Future Findings

Further investigation into LLM assaults in various industries and states holds enormous potential to uncover new insights and enhance understanding of evolving cyber threats. Here are some potential areas of investigation and discovery:

**Temporal Trends in LLM Attacks: Assessing LLM attack temporal trends over longitudinal research could provide important information about how attack patterns change over time.** Through monitoring variations in assault frequency, techniques, and impact metrics, scholars might detect new patterns and predict potential attack avenues in the future. Furthermore, investigating seasonal differences or associations with noteworthy occurrences may offer a more profound understanding of the variables influencing LLM attack dynamics.

**Sector-specific Vulnerability Assessments**: Diving further into these assessments might reveal distinct risks and difficulties that certain industries confront. Researchers may discover vulnerabilities unique to a given sector and create mitigation measures that are particularly customized to it by looking at the technology, procedures, and data kinds that are common in that industry. By using this strategy, firms may grow more resilient and stronger against LLM threats that are specific to their industry.

**Geospatial Analysis of LLM Attack Hotspots**: The identification of LLM attack hotspots at a more granular level by geographical analysis could yield important information into regional differences in attack intensity and prevalence. Researchers can find geographic clusters of high-risk areas and investigate possible underlying reasons influencing regional differences in attack frequency and impact by projecting LLM assault trends onto geographic regions. To successfully reduce the risks of LLM attacks, targeted initiatives and methods for allocating resources can be informed by such a geographical perspective.

**Machine Learning-Based Threat identification**: Enhancing proactive cybersecurity measures may be possible by utilizing machine learning models for threat and anomaly identification in LLM attack data. Researchers can create predictive models that can recognize new threats and unusual trends suggestive of impending attacks by training machine learning models on previous LLM attack data. Organizations that adopt a proactive strategy to threat identification may be able to prevent major harm from being caused by LLM attacks by taking proactive measures to mitigate risks.

**Ethical Considerations and Policy Implications:** By analyzing the ethical implications of LLM attacks, governance structures and regulatory frameworks that minimize risks and protect society's interests can be improved. Researcher engagement in interdisciplinary research that considers ethical, legal, and societal aspects might help shape responsible AI practices and laws that regulate LLM technology use. By taking a comprehensive approach, cybersecurity initiatives are guaranteed to be in line with larger moral standards and cultural norms, encouraging transparency, accountability, and trust in the implementation and use of LLM.

Appendix

Appendix A: Glossary

|  |  |
| --- | --- |
| Term | Definition |
| Asset | An asset is a valuable resource or item belonging to an organization, including tangible assets like hardware and equipment, intangible assets like reputation, and digital assets like data and software |
| Exploits | Exploits are techniques or methods used by attackers to take advantage of vulnerabilities in a system or application, often resulting in unauthorized access, data compromise, or system takeover. |
| Risk management | Risk management is a comprehensive strategy for recognizing, evaluating, addressing, and managing potential risks within an organization with the goal of reducing harm and maximizing benefits. |
| Incident management | Incident management involves the timely and efficient handling of security incidents, including identification, evaluation, and resolution processes aimed at minimizing harm, restoring functionality, and preventing future incidents. |
| Mitigation | Risk mitigation involves implementing measures to reduce the probability or severity of identified hazards, such as implementing safeguards, applying protective updates, and using effective methods. |
| Threat Actor | A Threat Actor is a person, group, or entity that carries out or directs malicious activities or attacks, often motivated by financial gain, political agendas, espionage operations, or activism. |
| Risk assessment | Risk assessment involves identifying, analyzing, and evaluating potential hazards to an organization's assets, operations, and objectives to prioritize mitigation efforts and allocate resources accordingly. |
| Threat Intelligence | Active Intelligence Gathering involves gathering, analyzing, and distributing information about imminent or developing threats, including threat actors, strategies, methods, and indicators of infiltration. |
| Confidentiality | Keep Information Private the guarantee that private data will be shielded from unwanted access or exposure and that only those with permission will be able to access it. |
| Integrity | The guarantee that information or data stays true, whole, and unaltered, protecting against unwanted additions, deletions, or modifications by malevolent parties |
| Availability | The guarantee that resources, services, and systems will be available and functional when needed, providing authorized users with continuous access to vital information and functions. |
| Large Language Model | A kind of AI model that can comprehend and produce language that is like that of a human after being trained on enormous volumes of text data. With many parameters, LLMs may handle a variety of natural language processing (NLP) tasks, including sentiment analysis, translation, summarization, and text synthesis. |
| Model Inversion | Attackers might try to reverse-engineer LLMs to obtain private data or insightful information from their internal representations. Model inversion attacks use flaws in the model's construction or training procedure to reassemble inputs or deduce private information. |
| Data Poisoning | Attackers may introduce fraudulent or deceptive data into LLM training datasets to undermine the model's behavior or performance. Attacks using data poisoning have the potential to introduce biases, distortions, or manipulations that compromise the validity and dependability of LLMs. |
| Privacy Breaches | Sensitive or private information stored in training data or internal representations may be disclosed by LLMs. Vulnerabilities in model architectures, inference techniques, or data handling procedures can result in privacy breaches by exposing sensitive data without authorization. |
| Model Manipulation | Attackers can change the way LLMs behave, produce, or make decisions by tampering with or modifying them. Model manipulation attacks might include changing the model's parameters, adding backdoor functionality, or faking input data to accomplish the attacker's intended results. |
| Data Exfiltration | Attackers may obtain confidential information, intellectual property, or proprietary data processed or stored via language processing systems. Various channels, such as network traffic, email, or file transfers, may be used for this. |
| LLM injection | A sort of attack known as LLM injection targets big language models by inserting slanted or malicious information into input prompts to alter the model's results. Attackers can manipulate the generated text to disseminate false information, spread propaganda, or maintain negative stereotypes by carefully constructing the input prompts. To detect and stop harmful content from being injected into the model's input prompts, adversarial testing, model monitoring, and strong input validation are necessary for mitigating LLM injection attacks. |
| Web Scrapping | Using software tools called web scrapers or crawlers, web scraping is the automated process of extracting data from websites. These tools search through webpages, find pertinent data, and save it in an organized manner for further processing or analysis. It makes it possible to do things like compile job posts, extract news headlines, and aggregate product prices. Following the terms of the website and acting ethically are essential to avoiding legal problems. |
| Security controls | Security controls are safeguards, countermeasures, or protective measures put in place to mitigate the possibility of security breaches. These can include technical, administrative, and physical controls. |

Table 1: Glossary Table

Appendix B: GitHub Repository

Overview:

The GitHub repository hosts all the code, data, and documentation related to the "LLM Risk Navigator" project conducted by Team Patriots. This repository serves as a centralized hub for managing project resources, tracking changes, and facilitating collaboration among team members. The main objective of this repository is to provide transparency, version control, and easy access to project artifacts, enabling both replication and future improvements.

GitHub Repository Link: [LLM\_risk\_analysis](https://github.com/DavonCarvalho/LLM_risk_analysis)

GitHub Repository Contents:

* **Markov Chain Analysis:** This directory contains Markov Chain Analysis for each sector. It also has several LLM attacks for each sector each month.
* **News Scrapping: This directory includes the code and data we acquired from web scrapping of the news articles related to LLM attacks.**
* **State wise Threat Assessment:** This directory includes the state wise analysis of LLM attacks. This data is used for tableau visualization which visualizes which state is more vulnerable to each sector.
* **VERIS Dataset:** This is the VERIS Community Database (VCDB) which is a repository of cyber-attacks happening in the whole world. We are using this dataset for our LLM attacks mapping. Here is the link of actual dataset: [VERIS](https://github.com/vz-risk/VCDB)

Appendix C: Risks

**Sprint 1 Risks**

| **Risk** | **Description** | **Probability** | **Impact** | **Mitigation** |
| --- | --- | --- | --- | --- |
| **Literature Gaps** | Limited accessible studies on LLM threats and applications | High | Moderate | Systematic literature reviews and expert consultations to identify critical research areas. |
| **Project Timing** | Constrained timelines for critical research phases | High | High | Develop an agile project management plan with flexible milestones to accommodate research needs. |

Table 2: Sprint 1 Risks

In Sprint 1, we address two significant risks that are critical to the initial stages of our project involving the integration and deployment of Large Language Models (LLMs): Literature Gaps and Project Timing.

**Literature Gaps:** We acknowledge that there is a good chance that we will come across few easily accessible papers on LLM risks and applications. This could have a moderate influence on our ability to base our project on the most recent findings and understandings. We are undertaking thorough literature reviews and consulting with experts to reduce this risk. Using this approach, we can ensure that our project is in line with the most recent advancements and industry best practices in LLM technology and highlight important research topics.

**Project Timing:** Due to the extremely tight deadlines for our project's crucial research phases, there is a serious risk that could materially affect our capacity to provide results on time. We are creating an agile project management strategy with adjustable milestones to handle this. This strategy lessens the possibility of delays by enabling us to respond more quickly to new information and difficulties in the field.

**Sprint 2 Risks**

| **Risk** | **Description** | **Probability** | **Impact** | **Mitigation** |
| --- | --- | --- | --- | --- |
| **Project Scope** | Ambiguities in project goals and expected outcomes | Moderate | Moderate | Regular alignment meetings with all stakeholders to ensure clarity and mutual understanding of project scope. |
| **Expertise Gaps** | Deficient knowledge in specific areas of LLM technology | Moderate | High | Scheduled training and continuous professional development in cutting-edge LLM applications. |
| **Data Integrity** | Errors and discrepancies in initial data sets | Low | High | Implement strict data validation protocols and periodic audits to ensure data accuracy. |
| **Literature Availability** | Limited studies on LLM application security across sectors | High | High | Engage with academic institutions and utilize databases like MITRE ATLAS for comprehensive reviews. |

Table 3: Sprint 2 Risks

The project scope, expertise gaps, and data integrity are among the major risks that our project tackles in Sprint 2. These risks are necessary for the effective continuation and execution of our Large Language Model (LLM) activities.

**Project Scope:** There is a moderate chance that disagreements about the objectives and anticipated results will arise, which will have a relatively negative influence on the project's course. We hold frequent alignment meetings with all stakeholders to lessen this. These meetings ensure that everyone involved understands the project scope and that the deliverables and objectives are in line with stakeholder expectations.

**Expertise Gaps:** There is a moderate chance that the team will lack some expertise, especially in more specialized LLM technology areas. This risk has a significant impact since it may make it more difficult for us to innovate and successfully apply cutting-edge solutions. As part of our mitigation approach, we make sure that our team stays up to date on the newest LLM applications through planned training sessions and ongoing professional development initiatives.

**Data integrity:** Although errors and discrepancies in initial datasets are not likely to occur, they might have a significant influence and jeopardize the accuracy and dependability of our study. We enforce stringent data validation procedures and carry out recurring audits to fix this. Through the project, these safeguards are intended to preserve the accuracy of our data, enabling precise results and well-informed decision-making.

**Sprint 3 Risks**

| **Risk** | **Description** | **Probability** | **Impact** | **Mitigation** |
| --- | --- | --- | --- | --- |
| **Integration Complexity** | Challenges in integrating LLMs with existing IT environments | Moderate | High | Pilot integration projects in controlled environments before full deployment. |
| **Resource Management** | Inefficient allocation of resources during development | Moderate | Moderate | Utilize project management tools to track resource usage and needs accurately. |

Table 4: Sprint 3 Risks

In Sprint 3, our project encounters two primary risks: Integration Complexity and Resource Management. Both are pivotal as we advance our integration of Large Language Models (LLMs) into existing IT environments and ensure efficient resource use during development.

**Integration Complexity:**

This risk relates to the difficulties of integrating LLM technologies into current IT infrastructures, which are sometimes complicated by different system capacities and legacy limitations. Pilot integration initiatives in controlled contexts are the focus of our mitigation plan because of the risk's substantial impact and modest probability. By identifying and resolving possible problems prior to a full-scale deployment, this method helps to maximize integration success and minimize disruptions.

**Resource Management:**

Maintaining project budgets and timetables during the development phase requires effective resource management. There is a moderate chance that these resources will be misallocated, which will have a moderate effect on the project's efficiency and advancement. We are using cutting-edge project management systems to track resource allocation and utilization to lessen this. By preventing resource waste, this guarantees that every project component has enough resources to support efficient project execution and decision-making.

**Sprint 4 Risks**

| **Risk** | **Description** | **Probability** | **Impact** | **Mitigation** |
| --- | --- | --- | --- | --- |
| **Regulatory Changes** | Updates in regulations affecting LLM deployment | Moderate | High | Stay updated with regulatory changes and adjust project guidelines accordingly. |
| **Performance Bottlenecks** | Decrease in system performance due to scaling | Moderate | High | Optimize system performance continuously and conduct scalability testing regularly. |

Table 5: Sprint 4 Risks

We are still committed to improving our Large Language Model (LLM) deployment tactics as our project moves into Sprint 4, even as we take care of any obstacles pertaining to performance and regulations. This stage is critical as we get ready for more widespread integration and implementation, making sure that our solutions not only comply with regulations but also function at their best in larger proportions.

**Regulatory Changes:**

We run the risk of having the deployment of Large Language Models (LLMs) impacted by regulatory changes in Sprint 4. There is a moderate chance that changes to the law and industry standards will have a major impact on our project because the quickly changing tech ecosystem is prone to frequent revisions. Failure to abide by these new regulations may have serious consequences, including lost money, project delays, and legal problems. To reduce this risk, we have taken a proactive stance that involves constant observation of the regulatory landscape and frequent consultation with legal professionals to make sure that our deployment plans are up to date and compliant.

**Performance constraints:**

The possibility of performance bottlenecks as we increase our LLM apps is another significant risk at this period. Due to the possibility that earlier scalability studies might not accurately represent the strains encountered in actual use, this risk is somewhat plausible. Performance problems have a big impact because they can make the LLM solutions less effective overall and worsen user experience, which could make stakeholders unhappy. To solve this, we're dedicated to always improving system performance through performance tuning, effective resource management, and routine scalability testing to make sure our systems can withstand growing loads without compromising functionality.

**Sprint 5 Risks**

| **Risk** | **Description** | **Probability** | **Impact** | **Mitigation** |
| --- | --- | --- | --- | --- |
| **Project Alignment** | Ensuring the project outcomes align with client expectations | Moderate | High | Conduct final reviews with stakeholders to ensure project deliverables meet all agreed specifications. |
| **Final Deliverables Quality** | Ensuring the quality and completeness of final deliverables | Moderate | High | Implement thorough testing and quality assurance processes before final submission. |
| **Knowledge Transfer** | Effective transfer of knowledge and technology to client teams | Low | High | Organize comprehensive training sessions and document all relevant project details for client handover. |

Table 6: Sprint 5 Risks

During the latter stages of our capstone project, we turned our attention to making sure that every project delivery met the operational requirements and client expectations. The results of our thorough evaluation of Large Language Models (LLMs) and their integration across several industries were solidified in large part thanks to this sprint. Our aims were to guarantee a smooth transfer of knowledge, produce a final product of the highest caliber, and verify that our outcomes complemented the client's strategic goals.

**Aligning with Client Goals**

It was crucial that the project's results matched the client's expectations. To accomplish this, we held several strategic meetings with all parties involved to thoroughly go over the project deliverables. Through these discussions, we were able to get input and make the necessary changes to make sure that every project component was suited to the client's strategic and practical goals. We also went over the project's goals and scope again to make sure that all the predetermined objectives were appropriately fulfilled and reflected in the work that was completed.

**Ensuring Quality of Deliverables**

This last sprint was heavily reliant on quality assurance. To guarantee the reliability, security, and effectiveness of any technical solution, we put strict testing procedures and quality controls in place. To improve functionality and remove any flaws, the project team thoroughly reviewed all implementation methodologies, code, and documentation. We wanted to give the customer tools that were both scalable for future demands and compliant with existing standards, thus we prioritized the dependability and effectiveness of our offerings.

**Knowledge Transfer and Documentation**

The successful knowledge transfer to the client's staff was an essential part of this phase. Throughout this process, thorough training sessions and extensive documentation covering maintenance protocols, troubleshooting techniques, and operational procedures were created. The client's personnel will be empowered by these materials, created to help them manage and maximize the use of LLM technology without requiring outside assistance. The long-term success of the project depended on making sure the client team was ready to take over.

**Project Closure**

At a formal closure meeting, we gave the client the final report and deliverables for the project. In addition to discussing the project's accomplishments and lessons learned, this meeting reinforced our team's and the client's cooperative efforts. We also proposed avenues for future research and development, as well as modifications and refinements to the technology now in use.

To sum up, our capstone project's last sprint was a thorough attempt to match every result with the client's strategic objectives, guaranteeing that the project not only fulfilled but also surpassed expectations. We were able to complete a project that is a tribute to the potential and possibilities of LLM technology through careful planning, execution, and teamwork. ￼

Appendix D: Agile Development

Scrum Methodology

**Figure 20: Sprint project dates**

For our project, we use an Agile method of development. This means that there are many short periods called "sprints." Each sprint has a set time and clear goals. These sprints work like organized cycles allowing our team to concentrate intensely on handling important parts of the project within an agreed-upon length of time. The separation between sprint dates is evident from Sprint 1 commencing on January 17th until January 31st, 2024 through to Sprint 5 encompassing April 10th until April 24th,2024 which makes it simple to understand how far along we have come in the project's progress. In every sprint, we keep our daily scrum logs up to date with the progress and tasks for each day. This helps in creating constant communication and teamwork among team members. By maintaining this everyday rhythm, everyone stays in sync with each sprint's objectives, aiding solutions to problems quicker and encouraging openness within the group.

Moreover, we use You Track for managing projects in a central location. We carefully note and follow all duties and assignments on this platform. It acts as our storehouse of work, helping us to keep an eye on advancement, spot possible blockages, and distribute resources well. We record all our work in You Track, making a detailed history of each person's input along with task updates and overall sprint development. This central method of managing projects simplifies the way we handle our project management. It gives power to those involved in deciding what projects should be given priority and where resources must be focused on. The use of Agile development dates as an obvious schedule for key points in our project allows us to plan and carry out every sprint with accuracy, ensuring that we are as efficient as possible while also providing value at right times to all parties interested.

**Sprint 1 Analysis**

In Sprint 1, the team made important progress in comprehending and tackling the intricate problem of large language model attacks. They carefully observed the matter area, linking how these attacks came about and what they imply to wider context of using language models. Their thorough investigation included studying previous literature and methods, which acted as a foundation for their following work.

They showed a proactive response to the problems by searching for possible solutions and ways to lessen their impact. It is likely that they held brainstorming sessions, read through relevant literature and maybe even started creating defense methods. The setting of specific goals for the project, which were centered on finding weak points and making strong responses, gave a clear focus or purpose to what the team was doing.

Also, the splitting up of main user stories and expression of an attractive product vision not only helped to create a common comprehension among team players but also prepared for upcoming development stages. The way Sprint 1 was done considers all aspects of this project, making sure that following sprints can use this first analysis as their starting point and move forward in delivering a strong solution for dealing with big language model attacks effectively.

**Sprint 2 Analysis**

In Sprint 2, the team started a thorough study of datasets that are important for dealing with big language model attacks. They concentrated on VERIS and CVE databases. The goal of this analysis was to comprehend the intricate situation connected to security incidents and flaws, which is essential for creating successful defense plans. By carefully investigating field descriptions in these sets of data, the group gained understanding about exact features and conditions that relate to their project goals. This careful examination made it possible to organize data in a way that can be easily analyzed and modeled later, matching the project's objectives.

In addition, we worked hard to condition the datasets. We have made them more consistent, correct and suited with what our project needs. By doing deep data cleaning, normalization and transformation tasks, we prepared the datasets to be ready for analysis and understanding. Moreover, a complete review of data quality was done which looked into aspects like how complete the information is as well its accuracy and timeliness. The team's evaluation of the dataset didn't just give us a clear understanding of its reliability, but also helped guide decisions on whether it was suitable to use for making decisions and training models. Besides VERIS and CVE, we looked at extra data sources which added more information about big language model attacks. It made our analysis deeper too. In these ways, Sprint 2 built a strong base for using datasets to make smart understandings and plans in later project parts.

**Sprint 3 Analysis**

In Sprint 3, the team moved ahead to outline a detailed solution method for handling big language model (LLM) assaults. The plan included creating an architecture of strong systems that can handle intricacies related with LLMs while being adaptable in face of changing dangers. Concentration was also on system security and ensuring robust actions against weak spots and unauthorized entry. The aim of creating systematic data flows was to make the processing and analysis of information more efficient. This would help in quickly spotting LLM attacks and responding to them. Likewise, when dealing with algorithms and methods for analysis, the team is showing they are dedicated to developing advanced ways of identifying and managing risks associated with LLM.

At the same time, Sprint 3 had a serious work on risk evaluation. They carefully checked threats, weaknesses and likely results linked with LLM attacks. This wide-ranging method for risk assessment helped the team understand all aspects of risks related to LLMs, which then allowed them to focus on mitigating these risks appropriately. Moreover, using methods of web scraping from Bing API and Google RSS gave immediate understanding about how often LLM attacks have been happening since November 2022. This active method of keeping an eye on and studying new dangers emphasizes the team's dedication to being very prepared and making sure their ways of reducing risks are always effective in a constantly changing threat environment.

To sum up, Sprint 3 was a big progress in the project's development. There were main focuses on creating solid plan for solution and doing thorough risk evaluations. By using high-level methods to design systems, keep them safe, flow and study data; the team prepared well for making strong actions against LLM attacks. The link with real-time threat information by scraping web showed that the group is taking steps before problems happen - showing their dedication to be updated about changing difficulties in cybersecurity caused by LLMs.

**Sprint 4 Analysis**

Sprint 4 demonstrates an essential time in the project's progress, emphasizing on constructing a strong project architecture using datasets like VERIS, MITRE ATT&CK and Census. The team plans to combine these different datasets so they can understand more about threats in different areas such as health care industry or financial sector among others: transportation field; making things sector; public service domain (police departments etc); schools/universities and administration area. This method not only gives room for better study of attack surface but also aids in forming mitigation plans that fit each sector's needs and weaknesses.

In Sprint 4, the focus is on using VERIS dataset. This dataset has a lot of detailed information about security incidents like how much they impact, ways attackers use and what assets are affected by the attack. By making use of VERIS data, our team can get an important understanding about what types of threats and attack paths are common in various areas. At the same time when we integrate MITRE ATT&CK framework it gives us a standard way to classify adversary actions and strategies. This gives room for a methodical study of attacks in various fields, helping to recognize similar patterns and tendencies.

In addition to this, the use of Census dataset gives information about population and socioeconomic aspects that enhance the study on sector-specific weaknesses. Knowing demographic properties in various areas can help reveal probable targets or reasons behind attacks. Furthermore, socioeconomic data might guide evaluations of danger by pointing out areas which could be more vulnerable due to resource limitations or infrastructure weaknesses for kinds of attacks.

One important part of Sprint 4 is using Markov chain analysis to model the movement between states in attack lifecycle across sectors. This method helps us see how attacks change from one phase to another by breaking down the attack lifecycle into separate states and studying the likelihood that an assault will move from state 0 (before attack) to state 1 (initial access) or other stages. In this way, we can understand better how attacks happen within each area's environment. For example, it could be that in the healthcare sector there are many transitions happening often from state 0 (pre-attack) to state 1 (initial access). This might be because of frequent phishing assaults aimed at health professionals. Knowing these transition probabilities helps in determining which threats to focus on for mitigation and how to distribute resources.

In summary, Sprint 4 is a big step forward in the project. The team creates a full project architecture using different datasets and complex analysis methods. By linking VERIS, MITRE ATT&CK with Census data, they form an all-inclusive comprehension of threats across sectors. This helps to make specific strategies for lessening risks in every sector. Also adding Markov chain analysis gives understanding on how attacks work within each area which guides actions for reducing dangers and improving security strength generally across every part of the system. In Sprint 4, we can see a notable progress in the development of our project as we construct an entire architecture that benefits from various datasets and advanced techniques for analysis. By merging VERIS together with MITRE ATT&CK along Census data - it forms into one comprehensive understanding about threats across sectors which then allows us to develop tailored strategies towards risk reduction more effectively in all areas involved. The application of Markov chain analysis also brings out detailed information on how attacks are happening within each sector's boundaries; this knowledge assists us greatly when taking steps proactively against these dangers while strengthening cybersecurity resilience altogether throughout entire system parts."

**Sprint 5 Analysis**

Sprint 5 is an important part of the project; it concentrates on threat evaluation in single state and sector. Using a formula for risk factor, the team did thorough study to identify areas and sectors that can be most at risk from big language model attacks. The results show that California is at more danger, with Texas and New Jersey next in line. This points towards a requirement for increased watchfulness and focused lessening methods in these states.

About the sector, the threat assessment helps to show us how vulnerable each main section is. These sections include education, finance, manufacturing, transportation and administration as well as healthcare and public services. This kind of analysis helps people involved in making decisions decide where they should put their resources and efforts first according to what risks are related with specific sectors. For instance, the finance field might be at a higher risk because financial data is delicate and any breach can result in big monetary losses. At times this may not be a direct translation but more of an explanation or meaning behind what has been said. On the other hand, there might be a weakness in the healthcare area because of how important patient information is and how it could affect their care if a security issue happens.

Also, the threat assessment is useful to create specific strategies for mitigation that focus on unique vulnerabilities in different sectors. People involved can recognize the particular problems and risk elements related to each sector, helping them in taking focused actions to strengthen cybersecurity. This might involve setting up superior systems for threat detection, putting strong access controls into place, or carrying out frequent security checks and training sessions.

In general, Sprint 5 emphasizes how necessary it is to do threat assessment in order to lessen the effects of big language model attacks. By figuring out which states and sectors have high risks, people who are interested can focus their resources and actions on reducing dangers. This helps them build up security against cyber threats. Moreover, the discovered results from threat assessment work like a guide for making specific reduction plans that aim at dealing with vulnerabilities unique to each sector as well as improving overall toughness towards attacks by large language models.

References:  
  
Works Cited:

1. *Mitre AtlasTM*. MITRE ATLASTM. (n.d.). <https://atlas.mitre.org/>
2. ThoughtSpot, T. (2024b, January 11). *Top security risks and risk mitigation in an LLM*. <https://www.thoughtspot.com/data-trends/ai/security-risk-mitigation-llm>
3. Elemind.com. (n.d.). *VERIS the vocabulary for event recording and Incident Sharing*. The VERIS Framework. [https://VERISframework.org/](https://verisframework.org/)
4. S Barnum. 2008. Common Attack Pattern Enumera- tion and Classification (CAPEC) Schema Description. *Department of Homeland Security*.
5. Niels Brink, Yori Kamphuis, Yuri Maas, Gwen Jansen- Ferdinandus, Jip van Stijn, Bram Poppink, Puck de Haan, and Irina Chiscop. 2023. Adversarial AI in the Cyber Domain.
6. Miles Brundage, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike, and Pamela Mishkin. 2023. Lessons Learned on Language Model Safety and Misuse.
7. OWASP. 2023. OWASP Top 10 for LLM.
8. Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023, July 12). A Comprehensive Overview of Large Language Models. arXiv:2307.06435 [cs.CL]. Retrieved April 5, 2024, from <https://arxiv.org/abs/2307.06435>
9. National Institute of Standards and Technology. (n.d.). NVD - Vulnerability Search. Retrieved February 8, 2024, from <https://nvd.nist.gov/vuln/search>
10. VERIS Community. (n.d.). VERIS Framework. Retrieved February 3, 2024, from [https://verisframework.org](https://verisframework.org/)
11. NAICS Association, LLC. (n.d.). Search NAICS Codes by Industry. Retrieved February 30, 2024, from <https://www.naics.com/search-naics-codes-by-industry/>
12. The OWASP LLM AI Cybersecurity & Governance Checklist provides a comprehensive framework for securing LLM applications. <https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.1.pdf>
13. Another version of the OWASP LLM AI Security & Governance Checklist, covering risks, vulnerabilities, and mitigation strategies. <https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist.pdf>
14. Blog, D. C. (2024, March 30). *Top 5 risks of large language models*. Deepchecks. <https://deepchecks.com/top-5-risks-of-large-language-models/>
15. Hassan, N. (2024, March 6). *Explore mitigation strategies for 10 LLM vulnerabilities: TechTarget*. Enterprise AI. <https://www.techtarget.com/searchenterpriseai/tip/Explore-mitigation-strategies-for-LLM-vulnerabilities>
16. Tableau Software. (n.d.). Discover - Tableau Public. Retrieved April 18, 2024, from https://public.tableau.com/app/discover
17. This paper introduces LLMSecGuard, an open-source framework that enhances code security by integrating static code analyzers with LLMs. <https://arxiv.org/html/2405.01103v1>
18. A blog post outlining a framework for securely using LLMs in companies, categorizing risks based on use cases and deployment types. <https://boringappsec.substack.com/p/edition-21-a-framework-to-securely>
19. A. Chidukwani, S. Zander and P. Koutsakis, "A Survey on the Cyber Security of Small-to-Medium Businesses: Challenges, Research Focus and Recommendations," in IEEE Access, vol. 10, pp. 85701-85719, 2022, doi: 10.1109/ACCESS.2022.3197899.
20. Protect Governments, and organizations Infrastructure against Cyber Terrorism (Mitigation and Stop of Server Message Block (SMB) Remote Code Execution Attack). International Journal of Engineering Research and Technology. [ISSN 0974-3154 Volume 11, Number 2 (2018), pp. 261](https://www.researchgate.net/profile/Nachaat-Mohamed/publication/326972739_Protect_Governments_and_organizations_Infrastructure_against_Cyber_Terrorism_Mitigation_and_Stop_of_Server_Message_Block_SMB_Remote_Code_Execution_Attack/links/5b6e97c5a6fdcc87df718202/Protect-Governments-and-organizations-Infrastructure-against-Cyber-Terrorism-Mitigation-and-Stop-of-Server-Message-Block-SMB-Remote-Code-Execution-Attack.pdf)
21. Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P., Goldblum, M., Saha, A., Geiping, J., & Goldstein, T. (2023, September 4). Baseline defenses for adversarial attacks against aligned language models. <https://arxiv.org/abs/2309.00614>
22. Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2024, March 2). Prompt injection attack against LLM-Integrated Applications. <https://arxiv.org/abs/2306.05499>
23. Normalizing risk scoring across different methodologies | SimpleRisk GRC Software. (2023, January 11). SimpleRisk GRC Software. <https://www.simplerisk.com/blog/normalizing-risk-scoring-across-different-methodologies>
24. Ait Maalem Lahcen, R., Mohapatra, R., Kumar, M. (2018). Cybersecurity: A Survey of Vulnerability Analysis and Attack Graphs. In: Ghosh, D., Giri, D., Mohapatra, R., Sakurai, K., Savas, E., Som, T. (eds) Mathematics and Computing. ICMC 2018. Springer Proceedings in Mathematics & Statistics, vol 253. Springer, Singapore. <https://doi.org/10.1007/978-981-13-2095-8_>
25. Evaluation of CHATGPT model for vulnerability detection. (n.d.-b). <https://arxiv.org/pdf/2304.07232.pdf>
26. Miranda-Calle, J.D., Reddy C., V., Dhawan, P. and Churi, P. (2021), "Exploratory data analysis for cybersecurity", World Journal of Engineering, Vol. 18 No. 5, pp. 734-749. <https://doi.org/10.1108/WJE-11-2020-0560>
27. Harjinder Singh Lallie, Lynsay A. Shepherd, Jason R.C. Nurse, Arnau Erola, Gregory Epiphaniou, Carsten Maple, Xavier Bellekens, Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic, Computers & Security, Volume 105, 2021, 102248, ISSN 0167-4048, <https://doi.org/10.1016/j.cose.2021.102248>.
28. 2024, 29th January. (2024, January 29). Large language models: Risk mitigation best practices - fintech futures: Fintech News. FinTech Futures. <https://www.fintechfutures.com/2024/01/large-language-models-risk-mitigation-best-practices/>
29. ​Google. (n.d.). Introduction to large language models (LLMs). Retrieved April 15, 2024, from https://developers.google.com/machine-learning/resources/intro-llms
30. ​Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023, July 12). A Comprehensive Overview of Large Language Models. arXiv:2307.06435 [cs.CL]. Retrieved April 5, 2024, from <https://arxiv.org/abs/2307.06435>
31. Blog, D. C. (2024b, April 21). Top security risks of large language models. Deepchecks. <https://deepchecks.com/top-security-risks-of-large-language-models/>
32. What are Large Language Models? - LLM AI Explained - AWS. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/what-is/large-language-model/>
33. ​Microsoft. (n.d.). Bing Web Search API. Retrieved May 3, 2024, from https://www.microsoft.com/en-us/bing/apis/bing-web-search-api
34. ​Demnati, I. (2019, August 29). Extract RSS News Feeds using Python and Google Cloud Services. Analytics Vidhya. Retrieved May 3, 2024, from https://medium.com/analytics-vidhya/web-scraping-news-data-rss-feeds-python-and-google-cloud-platform-7a0df2bafe44​Demnati, I. (2019, August 29). Extract RSS News Feeds using Python and Google Cloud Services. Analytics Vidhya. Retrieved May 3, 2024, from <https://medium.com/analytics-vidhya/web-scraping-news-data-rss-feeds-python-and-google-cloud-platform-7a0df2bafe44>