A black rectangle with a black background

Description automatically generated with low confidenceLogo, company name

Description automatically generated

DAEN 690

Project Report

Kanganut Jiradejdamrong

Matin Jahani Jirsaraei

Sai Unnathi Nandula

Bharat Sundar Rajan

Rahul Shivkumar

Suyog Patgave

Spring 2024

Security Risk of Large Language Models (LLMs)

**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 7](#_Toc165653440)

[1.1 Background 7](#_Toc165653441)

[1.2 Problem Space 9](#_Toc165653442)

[1.3 Research 10](#_Toc165653443)

[1.3.1 Phase 1 10](#_Toc165653444)

[1.3.2 Phase 2 12](#_Toc165653445)

[1.3.3 Phase 3 13](#_Toc165653446)

[1.4 Solution Space 19](#_Toc165653447)

[1.5 Project Objectives 20](#_Toc165653448)

[1.6 Primary User Stories 20](#_Toc165653449)

[1.7 Product Vision 21](#_Toc165653450)

[1.7.1 Scenario #1: Enhancing Security in Financial Institutions 21](#_Toc165653451)

[1.7.2 Scenario #2: Strengthening Healthcare Data Security 21](#_Toc165653452)

[Section 2: Datasets 22](#_Toc165653453)

[2.1 Overview 22](#_Toc165653454)

[2.2 Field Descriptions 23](#_Toc165653455)

[2.2.1 NAICS Code 23](#_Toc165653456)

[2.2.2 VERIS Dataset 24](#_Toc165653457)

[2.2.2.1 VERIS Data Exploration 25](#_Toc165653458)

[2.2.2.2 Classifying Ambiguous Industry Labels into Defined Categories 29](#_Toc165653459)

[2.2.3 Statistics of U.S. Businesses (SUSB) dataset 31](#_Toc165653460)

[2.2.4 MITRE ATLAS dataset 32](#_Toc165653461)

[**2.2.5** Web Scraping 34](#_Toc165653462)

[2.2.6 Synthetic Data 36](#_Toc165653463)

[2.2.6.1 What is Synthetic Data? 37](#_Toc165653464)

[2.2.6.2 Can Synthetic Data Replace Real Data? 37](#_Toc165653465)

[2.2.6.3 Implementation of Synthetic Data. 38](#_Toc165653466)

[2.2.6.4 List of Potential Use Cases of Synthetic Data 38](#_Toc165653467)

[2.2.6.5 Utility, Fidelity, and Privacy of Synthetic Data 39](#_Toc165653468)

[2.3 Data Context 39](#_Toc165653469)

[2.4 Data Conditioning 40](#_Toc165653470)

[2.5 Data Quality Assessment 40](#_Toc165653471)

[2.6 Storage Medium 41](#_Toc165653472)

[2.7 Storage Security 41](#_Toc165653473)

[2.8 Storage Costs 42](#_Toc165653474)

[Section 3: Algorithms & Analysis 42](#_Toc165653475)

[3.1 Solution Approach 42](#_Toc165653476)

[3.1.1 Systems Architecture 43](#_Toc165653477)

[3.1.2 Systems Security 45](#_Toc165653478)

[3.1.3 Systems Data Flows 45](#_Toc165653479)

[3.1.4 Algorithms & Analysis 46](#_Toc165653480)

[3.1.4.1 SMB Cyber profile Method 46](#_Toc165653481)

[3.1.4.2 Risk Framework Analysis 48](#_Toc165653482)

[3.1.4.3 Mapping VERIS Dataset with MITRE ATLAS 52](#_Toc165653483)

[3.1.4.4 Application of Lorenze Curves in Risk Profiling 61](#_Toc165653484)

[3.1.4.4.1 Lorenze Curve and Gini Coefficient 61](#_Toc165653485)

[3.1.4.4.2 Risk Calculation Based on the Frequency 64](#_Toc165653486)

[Section 4: Visualizations 73](#_Toc165653487)

[4.1 Overview 73](#_Toc165653488)

[4.2 Visualizations 74](#_Toc165653489)

[4.2.1 Risk Model 1 - Lorenz Curve Risk Model 74](#_Toc165653490)

[4.2.2 Risk Model 2 – Risk Framework 77](#_Toc165653491)

[Section 5: Findings 78](#_Toc165653492)

[Section 6: Summary 79](#_Toc165653493)

[Section 7: Future work 80](#_Toc165653494)

[Appendix A: Glossary 81](#_Toc165653495)

[Appendix B: GitHub Repository 82](#_Toc165653496)

[Appendix C: Risks 83](#_Toc165653497)

[Appendix D: Agile Development 87](#_Toc165653498)

[Works Cited 91](#_Toc165653499)

Table of Figures

[Figure 1: Order and pathways of the system. 11](#_Toc165653500)

[Figure 2: LLM attacks. 11](#_Toc165653501)

[Figure 3: Developers of LLM. 16](#_Toc165653502)

[Figure 4: Frequency of attacks on various sectors. 26](#_Toc165653503)

[Figure 5: Frequency of different types of attacks on various sectors. 27](#_Toc165653504)

[Figure 6: Victim Industry names before classification. 30](#_Toc165653505)

[Figure 7: Victim Industry names after classification. 31](#_Toc165653506)

[Figure 8: MITRE ATLAS Dataset. 33](#_Toc165653507)

[Figure 9: Heatmap of Timeline Events. 34](#_Toc165653508)

[Figure 10: Number of articles published per month. 35](#_Toc165653509)

[Figure 11: Top 10 published by number of articles. 36](#_Toc165653510)

[Figure 12: Risk Assessment Framework. 43](#_Toc165653511)

[Figure 13: Solution Architecture. 44](#_Toc165653512)

[Figure 14: SMB profiling method. 47](#_Toc165653513)

[Figure 15: Mapping Motives of Actors for Various Industries. 48](#_Toc165653514)

[Figure 16: Number of attacks on various industries. 50](#_Toc165653515)

[Figure 17: Frequency of different types of consequences for each industry. 52](#_Toc165653516)

[Figure 18: Lorenze cure description. 61](#_Toc165653517)

[Figure 19: Lorenz Curve for Incidents Frequencies of Attack Techniques within the Banking Sector. 63](#_Toc165653518)

[Figure 20: Lorenz Curve for Incidents Frequencies among Industry Sectors considering the DDos mode of Attack technique. 64](#_Toc165653519)

[Figure 21: Risk score based on the frequency of each impact in financial sector 65](#_Toc165653520)

[Figure 22: Heatmap showing the frequency of incidents per each asset and action type. 66](#_Toc165653521)

[Figure 23: Risk framework features. 68](#_Toc165653522)

[Figure 24: Power BI data schema and dataset relationships. 72](#_Toc165653523)

[Figure 25: Features for assessing risk score. 74](#_Toc165653524)

[Figure 26: Risk Model 1, Geographical Visualization –LLM Prompt Injection in Financial Sector. 75](#_Toc165653525)

[Figure 27: Risk Model 1, Geographical Visualization –LLM Prompt Injection in Health Care Sector. 76](#_Toc165653526)

[Figure 28: Risk Score 2, Geographical Visualization – Financial Sector. 77](#_Toc165653527)

[Figure 29: Risk Score 2, Geographical Visualization – Health Care Sector. 78](#_Toc165653528)

[Figure 30: Sprint project dates. 87](#_Toc165653529)

Table of Tables

[Table 1: Percentage of LLMs’ response on simple safety tests [16]. 12](#_Toc165653530)

[Table 2: Overview of examined models. 18](#_Toc165653531)

[Table 3: NAICS code with definitions. 24](#_Toc165653532)

[Table 4: VERIS Dataset Fields Description. 25](#_Toc165653533)

[Table 5: Frequency of attacks on various industries. 25](#_Toc165653534)

[Table 6: Reduced data frame with information for the attacks. 26](#_Toc165653535)

[Table 7: Additional research and data exploratory analysis. (1971-2023). 28](#_Toc165653536)

[Table 8: Statistics of U.S. Businesses (SUSB) Annual Datasets. 32](#_Toc165653537)

[Table 9: Attack on Industries by Different Actor Groups for 2023. 49](#_Toc165653538)

[Table 10: The number of attack incident. 49](#_Toc165653539)

[Table 11: The number of integrities, confidentiality, and availability for each industry. 51](#_Toc165653540)

[Table 12: Identified LLM attack techniques from MITRE ATLAS framework. 53](#_Toc165653541)

[Table 13: Results of mapping MITRE ATLAS LLM techniques to VERIS Cyberattack techniques. 61](#_Toc165653542)

[Table 14: An Example of Incident Frequencies by Attack Techniques Across Industry Sectors. 62](#_Toc165653543)

[Table 15: The Cumulative Percentages computed from Attack Technique incident frequencies in the Banking Sector. 62](#_Toc165653544)

[Table 16: The Cumulative Percentages computed from Industry sector incident frequencies considering the DDos mode of Attack Technique. 63](#_Toc165653545)

[Table 18: The probabilities of the most occurred hacking type for each industry. 67](#_Toc165653546)

[Table 19: The probabilities of all action type for each industry. 67](#_Toc165653547)

[Table 20: relation between incident ID, actions, assets, and attributes. 68](#_Toc165653548)

[Table 21: Converting impact as categorical feature to numerical. 69](#_Toc165653549)

[Table 22: The probabilities of all types of actions, assets, and attributes for different industries. 70](#_Toc165653550)

[Table 23: Composite risk score for each industrial sector. 71](#_Toc165653551)

[Table 24: Example of final table (combining VERIS dataset with LLMs mapped, NAICS code and SUSB dataset). 73](#_Toc165653552)

[Table 25: Example of final risk calculation. 73](#_Toc165653553)

[Table 26: Glossary Table. 81](#_Toc165653554)

[Table 27: Sprint 1 Risks 83](#_Toc165653555)

[Table 28: Sprint 2 Risks 84](#_Toc165653556)

[Table 29: Sprint 3 Risks 84](#_Toc165653557)

[Table 30: Sprint 4 Risks 85](#_Toc165653558)

[Table 31: Sprint 5 Risks 86](#_Toc165653559)

Abstract

Abstract

Cyberattacks have been prevalent for many decades, spanning various industries such as finance, retail, education, and many more. These attacks target digital assets across industries, resulting in a wide range of damages and costing $10 billion in 2022, a surge of 50% when compared with $6.9 billion in 2021. The number of complaints the FBI received reached 3.26 million in 2022. The adoption of Large Language Models (LLMs) has been rapidly increasing across all industries to improve operation efficiency. However, since the technology is advanced and complex, it is necessary to secure LLMs to avoid severe consequences such as data breaches, financial losses, theft, and other cyberattacks. To address this problem, this study focuses on assessing the risk and providing an analysis of LLM attacks across industries. Analyzing datasets such as Vocabulary for Event Recording and Incident Sharing (VERIS), Statistics of US Businesses (SUSB), and framework such as MITRE Corporation Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS), the study aims to develop risk measures to address vulnerabilities across industries. The study gathers information on attack actors, assets affected, the actions undertaken, and the impact of the attack. The study presents an advanced methodology for assessing risk using Lorenz's curve. The result of the assessment provides foresight into possible LLM attacks and related vulnerabilities. The result of this study is an essential factor in guiding future research on analyzing and predicting LLM attacks with a detailed approach, potentially helping in mitigating and diverting attacks.

**This page intentionally left blank**

Report

# Problem Definition

## Background

A large language model is the language model with massive parameters that undergoes pretraining tasks (e.g., masked language modeling and autoregressive prediction) to understand and process human language, by modeling the contextualized text semantics and probabilities from large amounts of text data. A capable LLM should have four key features [1] : (i) profound comprehension of natural language context; (ii) ability to generate human-like text; (iii) contextual awareness, especially in knowledge-intensive domains; (iv) strong instruction-following ability which is useful for problem-solving and decision-making.

There are several LLMs that were developed and released in 2023, gaining significant popularity. Notable examples include OpenAI’s ChatGPT [2], Meta AI’s LLaMA [3], and Databricks’ Dolly 2.0 [4]. For instance, ChatGPT alone boasts a user base of over 180 million [5]. LLMs now offer a wide range of versatile applications across various domains. Specifically, they not only provide technical support to domains directly related to language processing (e.g., search engines, customer support, translation) but also find utility in more general scenarios such as code generation, healthcare [6], finance [7], and education [8]. This showcases their adaptability and potential to streamline language-related tasks across diverse industries and contexts.

LLMs are gaining popularity within the security community. As of February 2023, a research study reported that GPT-3 uncovered 213 security vulnerabilities (only 4 turned out to be false positives) in a code repository. In contrast, one of the leading commercial tools in the market detected only 99 vulnerabilities. More recently, several LLM-powered security papers have emerged in prestigious conferences. For instance, in IEEE S&P 2023, Hammond Pearce et al. [9] conducted a comprehensive investigation employing various commercially available LLMs, evaluating them across synthetic, hand-crafted, and real-world security bug scenarios. The results are promising, as LLMs successfully addressed all synthetic and hand-crafted scenarios. In NDSS 2024, a tool named Fuzz4All [10] showcased the use of LLMs for input generation and mutation, accompanied by an innovative auto prompting technique and fuzzing loop.

LLMs represent an evolution from language models. Initially, language models were statistical in nature and laid the groundwork for computational linguistics. The advent of transformers has significantly increased their scale. This expansion, along with the use of extensive training corpora and advanced pre-training techniques is pivotal in areas such as AI for science, logical reasoning, and embodied AI. These models undergo extensive training on vast datasets to comprehend and produce text that closely mimics human language. Typically, LLMs are endowed with hundreds of billions, or even more, parameters, honed through the processing of massive textual data. They have spearheaded substantial advancements in the realm of Natural Language Processing (NLP) [11] and find applications in a multitude of fields (e.g., risk assessment, programming, vulnerability detection, medical text analysis, and search engine optimization).

Based on Yang’s study [1], an LLM should have at least four key features. First, an LLM should demonstrate a deep understanding and interpretation of natural language text, enabling it to extract information and perform various language-related tasks (e.g., translation). Second, it should have the capacity to generate human-like text (e.g., completing sentences, composing paragraphs, and even writing articles) when prompted. Third, LLMs should exhibit contextual awareness by considering factors such as domain expertise, a quality referred to as ‘‘Knowledge intensive’’. Fourth, these models should excel in problem-solving and decision-making, leveraging information within text passages to make them invaluable for tasks such as information retrieval and question-answering systems.

There is a diversity of providers for language models, including industry leaders such as OpenAI, Google, Meta AI, and emerging players such as Anthropic and Cohere. The release dates span from 2018 to 2023, showcasing the rapid development and evolution of language models in recent years. Newer models such as ‘‘gpt-4’’ have emerged in 2023, highlighting the ongoing innovation in this field. While most of the models are not open-source, it is interesting to note that models like BERT, T5, PaLM, LLaMA, and CTRL are open-source, which can facilitate community-driven development and applications. Larger models tend to have more parameters, potentially indicating increased capabilities but also greater computational demands. For example, ‘‘PaLM’’ stands out with a massive 540 billion parameters. It can also be observed that LLMs tend to have more parameters, potentially indicating increased capabilities but also greater computational demands. The ‘‘Tunability’’ column suggests whether these models can be fine-tuned for specific tasks. In other words, it is possible to take a large, pretrained language model and adjust its parameters and training on a smaller, domain-specific dataset to make it perform better on a particular task. For instance, with tunability, one can fine-tune BERT on a dataset of movie reviews to make it highly effective at sentiment analysis.

The widespread adoption of Large Language Models (LLMs) across diverse industries presents a double-edged sword: while LLMs offer unprecedented efficiency and capabilities, they also introduce significant security risks. As cyberattacks continue to escalate in frequency and sophistication, securing LLMs has become a paramount concern for organizations worldwide. This section delves into the intricate security landscape surrounding LLMs, exploring vulnerabilities, potential threats, and strategies for mitigating risks.

LLMs, with their massive parameter sizes and intricate architectures, are not immune to vulnerabilities. One of the primary concerns is the susceptibility to adversarial attacks, where malicious actors exploit weaknesses in the model to manipulate or deceive its outputs [9]. Adversarial attacks can manifest in various forms, including injecting subtle perturbations into input data or crafting malicious prompts to induce undesirable behaviors in the model. Additionally, vulnerabilities stemming from inadequately sanitized training data or flawed algorithms can also compromise the integrity and security of LLMs, leading to biased outputs or unintended consequences [9].

The evolving threat landscape surrounding LLMs encompasses a spectrum of adversaries, ranging from individual hackers to sophisticated state-sponsored actors. These adversaries leverage LLMs for nefarious purposes, including generating convincing phishing emails, crafting misleading propaganda, or orchestrating sophisticated social engineering attacks. Moreover, the proliferation of LLM-powered tools in cybersecurity domains introduces new attack vectors, as threat actors exploit vulnerabilities in these tools to bypass security defenses, evade detection, or launch targeted attacks [10].

The impact of LLM-related security breaches extends across various industries, posing significant financial, reputational, and regulatory risks. In the financial sector, for instance, malicious actors may exploit vulnerabilities in LLMs to manipulate markets, execute fraudulent transactions, or conduct sophisticated cyber heists. Similarly, in healthcare, the integrity of medical records and patient data could be compromised through LLM-powered attacks, leading to privacy violations or even endangering lives. Education institutions are not immune either, as LLM-enabled plagiarism detection tools may inadvertently leak sensitive student information or facilitate academic dishonesty.

Addressing the security challenges posed by LLMs requires a multi-faceted approach encompassing technical, organizational, and regulatory measures. Robust security protocols, such as encryption, access controls, and anomaly detection mechanisms, should be integrated into LLM deployment pipelines to safeguard against unauthorized access and data breaches. Moreover, ongoing monitoring and auditing of LLM activities are essential for detecting and mitigating potential threats in real-time. Collaborative efforts between industry stakeholders, academia, and government agencies are also crucial for sharing threat intelligence, developing best practices, and establishing regulatory frameworks to govern the responsible use of LLMs.

## Problem Space

A primary concern with deploying large language models in enterprise settings is potentially including sensitive data during training. Once data has been incorporated into these models, it becomes challenging to discern precisely what information was fed into them. This lack of visibility can be problematic when considering the myriad data sources used for training and the various individuals that could access this data. Ensuring visibility into the data sources and maintaining strict control over who has access to them is crucial to prevent unintentional exposure of confidential information.

An additional concern is the potential misuse of LLMs in cyberattacks. Malicious actors can utilize LLMs to craft persuasive phishing emails to deceive individuals and gain unauthorized access to sensitive data. This method, known as social engineering, has the potential to create compelling and deceptive content, escalating the challenges of data protection. Without rigorous access controls and safeguards, the risk of significant data breaches increases, with malicious actors gaining the ability to spread misinformation, propaganda, or other harmful content with ease.

While LLMs have near-infinite positive applications, they harbor the potential to create malicious code, bypassing conventional filters to prevent such behaviors. This susceptibility could lead to a new era of cyberthreats where data leaks aren’t just about stealing information but generating dangerous content and codes. If manipulated, for instance, LLMs can produce malicious software, scripts, or tools that can jeopardize entire systems. Their potential for “reward hacking” raises alarms in the cybersecurity domain, suggesting unintended methods to fulfill their objectives could be discovered, leading to accidental access to or harvesting of sensitive data [12].

As we rely more on LLM applications, it becomes imperative for organizations and individuals to stay vigilant to these emerging threats, prepared to protect data at all times.

Conventional application vulnerabilities present a new strain of security risks within LLMs. But true to form, OWASP delivered the OWASP Top Ten LLM Security Risks in timely fashion, alerting developers to new mechanisms and the need to adapt traditional remediation strategies for their applications utilizing LLMs.

LLM01: Prompt Injection: Prompt injection can manipulate a large language model through devious inputs, causing the LLM to execute the attacker's intentions. With direct injections, the bad actor overwrites system prompts. With indirect prompt injections, attackers manipulate inputs from external sources. Either method can result in data exfiltration, social engineering, and other issues [12].

LLM02: Insecure Output Handling: Insecure output handling is a vulnerability that occurs when an LLM output is accepted without scrutiny, exposing backend systems. It arises when a downstream component blindly accepts LLM output without effective scrutiny. Misuse can lead to cross-site scripting (XSS) and cross-site request forgery (CSRF) in web browsers, as well as server-side request forgery (SSRF), privilege escalation, and remote code execution on backend systems [12].

LLM03: Training Data Poisoning: Training data poisoning occurs when LLM training data is manipulated via Common Crawl, WebText, OpenWebText, books, and other sources. The manipulation introduces backdoors, vulnerabilities, or biases that compromise the LLM’s security and result in performance decline, downstream software exploitation, and reputational damage [12].

LLM04: Model Denial of Service: Model denial of service occurs when an attacker exploits a LLM to trigger a resource-intensive operation, leading to service degradation and increased costs. This vulnerability is amplified by the demanding nature of LLMs and the unpredictable nature of user inputs. In a model denial of service scenario, an attacker engages with an LLM in a manner that demands a disproportionate amount of resources, causing a decline in service quality for both the attacker and other users while potentially generating significant resource expenses [12].

LLM05: Supply Chain Vulnerabilities: Supply chain vulnerabilities in LLMs can compromise training data, ML models, and deployment platforms, causing security breaches or total system failures. Vulnerable components or services can arise from poisoned training data, insecure plugins, outdated software, or susceptible pretrained models [12].

LLM06: Sensitive Information Disclosure: LLM applications can expose sensitive data, confidential information, and proprietary algorithms, leading to unauthorized access, intellectual property theft, and data breaches. To mitigate these risks, LLM applications should employ data sanitization, implement appropriate strict user policies, and restrict the types of data returned by the LLM [12].

LLM07: Insecure Plugin Design: Plugins can comprise insecure inputs and insufficient access control, making them prone to malicious requests that can lead to data exfiltration, remote code execution, and privilege escalation. Developers must follow stringent parameterized inputs and secure access control guidelines to prevent exploitation[12].

LLM08: Excessive Agency: Excessive agency refers to LLM-based systems taking actions leading to unintentional consequences. The vulnerability stems from granting the LLM too much autonomy, over-functionality, or excessive permissions. Developers should limit plugin functionality to what is absolutely essential. They should also track user authorization, require human approval for all actions, and implement authorization in downstream systems[12].

LLM09: Overreliance: An LLM can generate inappropriate content when human users or systems excessively rely on the LLM without providing proper oversight. Potential consequences of LLM09 include misinformation, security vulnerabilities, and legal issues [12].

LLM10: Model Theft: LLM model theft involves unauthorized access, copying, or exfiltration of proprietary LLMs. Model theft results in financial loss and loss of competitive advantage, as well as reputation damage and unauthorized access to sensitive data. Organizations must enforce strict security measures to protect their proprietary LLMs [12].

## Research

The research with regards to LLM initiated with the fundamental understanding of the large language model architecture. This included research on basic machine learning, natural language processing, neural networks, and deep learning.

## Phase 1

The research jumpstarted with individual research over various LLM topics, a collective effort by all the team members. The initial study focused on various attributes of LLM, such as its training dataset and safety and risk parameters, and the experimental study examined whether a coded cipher was able to bypass the safety alignment, which it did. Results based on this study showed that there's a need for developing safety alignments for non-natural languages in LLM for better reliability of responses [13].

The research also shed light on reinforced learning from human feedback. There seemed to be increased disagreement between humans and models as the experiment progressed to test the contextual integrity theory. This includes showing the likelihood of revealing specific types of sensitive information [14].

The fundamental understanding of the model and the application of chatbots was examined as well, defining the order and pathways of how the system works [15].

A diagram of a computer

Description automatically generated

Figure 1: Order and pathways of the system.

A diagram of a system

Description automatically generated

Figure 2: LLM attacks.

The comparison of various LLM models that are available publicly was also studied. Many of the models responded with harmful or sensitive information when prompts were provided to them. The authors have raised a proposal and experimented with the use of simple safety tests, which are a set of protocols and prompt questions to identify critical safety measures. Five distinctive harm areas were shortlisted with the responding prompts. The harmful areas are illegal items, physical harm, scams and fraud, child abuse, suicide, and self-harm.

Further, two categories of tests were made:

1. Information Seeking and Advising: Questions asked of the model that, if answered in an unsafe way, would readily give the user access to harmful information, or would encourage them to engage in harmful activities.
2. Instructions and Actions Commands given to the model that, if complied with, would result in unsafe information being provided or unsafe content being created.

In an overall experiment it was observed that by using safety prompts, MML models were able to decrease the percentage of harmful content as their response [16].

A table of numbers and symbols

Description automatically generated with medium confidence

Table 1: Percentage of LLMs’ response on simple safety tests [16].

## Phase 2

The second phase of research progressed with an understanding of how LLMs are implemented in various industries and businesses listed on the North American Industries Classification System (NAICS). This research is deemed effective in its nature of being descriptive of the threats and risks an LLM may potentially hold up that would cause harm to the end user. We focused on industries such as agriculture, financial, and technical aspects such as job services.

We started by delving into the rising use of WormGPT, an AI tool based on GPTJ, by cybercriminals for Business Email Compromise (BEC) attacks. This tool aids in crafting emails that are highly convincing due to their advanced grammar, making BEC scams more sophisticated and challenging to detect. The article emphasizes the need for targeted defense strategies and robust email verification processes to combat these AI-enabled threats. The evolving nature of cyber threats, with AI at the forefront, calls for adaptive and advanced security measures [17].

**Agriculture, Forestry, Fishing and Hunting Sector**: AI is required in forestry, fishing, hunting, and agriculture for a number of reasons. **Productivity and Efficiency:**AI can evaluate enormous volumes of data to enhance agricultural techniques, raise crop yields, and better allocate resources. **Predictive analysis:** AI systems can predict market trends, weather patterns, and crop illnesses, which improves readiness and decision-making.

**Sustainability:** AI can aid in biodiversity conservation, environmental impact reduction, and sustainable resource management. **Precision Agriculture:** AI makes it possible to apply pesticides, fertilizers, and water more precisely, which lowers waste and protects the environment. **Monitoring and Management:** AI tools support conservation efforts by helping to monitor the health of forests, wildlife populations, and fisheries stocks. **Labor and Cost Reduction:** Routine task automation lowers both the need for labor and operating expenses.

**Agriculture BERT a BERT-based language model that has been further pre-trained from SciBERT's checkpoint, is introduced by Hugging Face. It is intended for use in the agricultural field, integrating general and scientific agricultural knowledge. There are 5.3 million paragraphs from agricultural literature and 1.2 million paragraphs from the National Agricultural Library in the corpus. After being trained with Masked Language Modeling (MLM), the model can provide bidirectional sentence representation by predicting masked words in sentences. This model is an important tool for agricultural AI applications and research.**

The study also talks about "agri1.ai," an effort that investigates the possibilities of AI in agriculture, particularly Large Language Models (LLMs) like ChatGPT. Using an existing LLM, the project's goal is to construct a front-end interface while optimizing and integrating it with both internal and external data. They are also thinking of developing an agriculture specific LLM. This LLM, called "agriLLM," would receive a great deal of training on data pertaining to agriculture, making it extremely specialized for the sector. Data collection, preprocessing, model selection, training, and ongoing improvement are all steps in the process. Using AI-driven solutions, the initiative seeks to empower farmers and highlights the significance of addressing agricultural difficulties, particularly in places like Africa.

**Investment Banking and Securities Intermediation:** Risks based on using the LLM for Financial Institution (FI) service by both employers in the company as well as the end user itself. The discussion was focused on LLM security risks and their severity. The discussion involved professionals from backgrounds in data analysis, AI, and data management from various FIs, including top commercial banks in the US. Information retrieval, text summarization, text extraction of unstructured data for operational automation and sentiment analysis, code translation to non-technical terms, and vice versa. Major risks involve hallucinations, being unable to trace the information within the data itself, and yet providing an output to the requested prompt.

Security and privacy breach on FI data to the LLM providers. Also, consider IP violations. Training the model on external data without permission or authorization. Reputation: The concern is related to the response provided to the customer-facing cases.  The concerns extended to being unjust and biased toward the end user based on their demographic and financial status.

## Phase 3

In this weekly session an informative meeting was arranged with the partner and the group. The brainstorming session followed the meeting. The exchange of information was based on the previous research conducted and the future research topics to focus on. Large language models rely on vast neural networks that use weights within the networks to process the inputs given to the system and generate output. One such higher level of neural networks in deep learning. The LLMs have been using such process and foundation to what has become one of the greatest achievements in technology. The development of LLMs such as ChatGPT by OpenAI has made tremendous impact on everyone across the world. The ability of this LLM not just to write texts but also to make sense of its nature and understand human natural language and subtleties make it a very complex and technologically advanced system that is both publicly and proprietarily available to use. ChatGPT is only one of many LLMs that possess the capability that transcends primary capabilities of data extraction and information.

Since its initial release of ChatGPT in November 2022, the number of users has been skyrocketing. The possibilities that can be achieved using such technology are hard to comprehend. Soon enough, OpenAI started to develop GPT-4, an advanced version of its previous generation. This soon caught the eyes of businesses of various industries that realized how the use of such LLM can be made to develop their product, service, or business. Considering such possibilities, we extended our research in analyzing and finding insights into understanding how the LLMs are prone to risks of malfunction, the application of LLMs in various business in different industrial fields.

The graph illustrates the distribution of companies utilizing OpenAI in their business processes as of January 2023. The technology sector leads with 251 companies, followed by education with 209. Business services, manufacturing and finance sectors also show significant adoption with 98, 89, and 44 companies respectively. Other notable sectors include retail, healthcare, government, and media & internet. The total number of companies across all sectors is 902. The "Others" category includes 117 companies, indicating a diverse range of industries exploring AI applications. OpenAI, known for its AI research and deployment, is the entity behind products like ChatGPT.[9] We comprehensively researched regarding three industries to analyze how the LLM was applicable and its merits and demerits.

**Marketing:** The marketing and advertisement sector has a lot of innovative and fascinating use cases, some of which we currently employ and others of which we are actively investigating.Many replies are gathered from surveys, focus groups, and interviews in the context of market research. These answers are considered "verbatim" statistics since they accurately reflect what respondents said. With this data, large language models (LLMs) can be useful. We may swiftly go through each response, determine which are most crucial, and provide a concise summary of them. When the researcher has to present the results to clients, this facilitates and expedites their work. It's similar to having a knowledgeable assistant who can sift through a ton of data and provide you with a clear, helpful summary of the key aspects.

Automated reporting: A large amount of numerical data, such as survey results, is gathered in market research. This data must be clearly arranged, condensed, and presented. Here, large language models, or LLMs, can be quite useful. With the use of charts and tables, they may swiftly arrange this data and provide basic, preliminary summaries or titles. Executive summaries, which are concise, understandable explanations of the key results, are another skill they possess. The identification of major themes or subjects in the data is another task for LLMs. To ascertain the prevailing themes, people's attitudes about particular subjects, and their perceptions of various companies, they can leverage disparate viewpoints or establish connections with digital platforms. This facilitates greater understanding and data refinement by the researchers. To put it simply, LLMs are intelligent assistants that can swiftly sort through a plethora of data and opinions, facilitating researchers' comprehension and utilization of the information [18].

Prediction: It is possible for Large Language Models (LLMs) to transform words or sentences into mathematical structures known as "embeddings." The essence or meaning of the words is captured by these embeddings. These embeddings can be used to generate predictions by other machine learning models. For instance, they can forecast the performance of a TV advertisement by examining its dialogue. In a similar vein, these models can forecast how customer interactions influence brand loyalty or whether they increase the likelihood that a customer will discontinue using the company (churn) by examining what customers say about their contacts with service representatives. Like chatbots, conversational AI is becoming more intelligent. Now that they are able to comprehend and react to prior responses, they are able to have more casual interactions. This implies that questionnaire design—that is, surveys with a collection of questions—is evolving when it comes to market research. These AI tools can aid in the more effective and reliable creation of these questionnaires. The AI may determine which question to ask next, for instance, based on a respondent's answer to a previous question. This makes the survey more pertinent and interesting for every participant [19].

Creative Writing: Discussion guides are scripts or outlines used to steer discussions during meetings. A key topics and questions to ask guide can be prepared with the assistance of an LLM. First Drafts of Presentations: An LLM can assist in writing the first draft of a presentation if one is required. It has the ability to clearly and captivatingly arrange concepts and data. Writing that is used to advertise goods or services is known as marketing copy. For brochures, websites, or commercials, LLMs may write compelling copy. Concept statements are succinct summaries of a concept, service, or good. An LLM can assist in providing a succinct and enticing summary of the key features and advantages. Other Original Thoughts: The possibilities are truly endless. Creating creative content can be helped by LLMs, whether it's coming up with a storyline for a book, a script for a film, or even just coming up with original ideas for a project [20].

Conversational search queries: To gain a deeper understanding of the user's search intent, the BiRNN model with attention might be utilized. A standard keyword-based search query can be converted by this approach into a more in-depth natural language inquiry. This translation aids in the search engine's more precise understanding of the context and details of the user's request. You may make search more conversational and user-friendly by converting search queries into questions in natural language. By allowing users to engage with the search system as though they were speaking with a human, the process becomes more natural and interesting [21].

Possible risks associated with LLM in the marketing and advertisement sector include inaccurate material generation, where hallucinations may lead to content with inaccurate information, potentially resulting in deceptive marketing collateral or commercials that harm brand credibility. Additionally, brand image and trust issues may arise if marketing content unintentionally contains false or misleading information due to LLM errors, particularly damaging in fields like banking or healthcare where precision is crucial for consumer trust.

Advertising has some legal and ethical issues. Not only is deceptive advertising immoral, but it is also prohibited in many places. An LLM may unintentionally create content that breaches advertising rules (such as creating unsupported claims), which could have negative legal effects on the business. Deceptive Analytics and Market Research: LLMs can be used to examine customer attitudes or market trends. Inaccurate data interpretation and hallucinations by the model could result in poor business decisions based on erroneous analysis [22].

Adherence to truth-in-advertising legislation and legal norms is essential. LLMs may produce content that violates certain legal requirements, which could result in legal problems.

Establishing trust with customers is a prerequisite for effective marketing. Customers' faith in the brand may be damaged if they receive conflicting or erroneous information as a result of the LLM's ignorance [23].

An LLM may unintentionally violate someone else's intellectual property rights if it recycles previously published content from the internet. For businesses that use such content in their marketing materials, this could result in legal problems. Ambiguity in Content Ownership: It's not always evident who is legally entitled to the material created by LLMs. This content may give rise to ownership conflicts and copyright infringement lawsuits if it is used in advertising.

Authenticity and uniqueness are highly prized in marketing. Customers may be duped into believing that the marketing message is new if LLM-generated content is used, even if it is only a rehash of preexisting information. Ethical Issues: It is unethical to use publicly accessible data for commercial endeavors without explicit authorization, particularly in marketing and advertising [24].

**Consumer and commercial Retail:** The consumer retail industry is one of the fastest-growing sectors in the market. With the ever-increasing development of LLM and its applications, such as chatbots, the use of LLM has drastically impacted the industry overall. These changes are a major improvement in the orthodox methodologies of businesses, making them more flexible, convenient, and efficient in handling and growing the retail industry.

AI has seen rapid development in recent years, and these changes have been brought about by the increasing development of neural networks and deep learning. As of today, there are several developers of LLM with an extensive number of parameters, ranging from a few million to half a trillion. These developers include the top industry names such as Google, Meta, DeepMind, Nvidia, and the AI giant OpenAI (Microsoft).

A screenshot of a computer

Description automatically generated

Figure 3: Developers of LLM.

The adoption of such technology has shown signs in advanced as well as emerging markets around the world, mainly used for optimizing customer experiences and later for expansion and scalability. The differences are also to be noted with respect to the application and the location where the model is to be trained and used. This includes cultural and language differences, economic factors, regulations, local ecosystems, and partnerships.

While the LLM might be in its recent development stages, its application cases in the retail industry are notable. The use of LLMs like GPT-4, which is one of the most acquired LLMs, in the retail market. The purpose it serves may vary based on applications.

Customer service in LLM-integrated retail systems encompasses various aspects, including the use of chatbots for virtual assistance, troubleshooting, and personalized customer experiences. Historical data is leveraged to predict product recommendations, customer preferences, analyze feedback, and improve communication. Inventory management and forecasting benefit from LLM by automating processes with real-time data, reducing overstocking and understocking. Advantages of LLM in inventory management include improved data analysis, efficiency, accuracy, adaptability, cost-effectiveness, scalability, and enhanced customer satisfaction.

LLM enables data-driven decision-making by predicting market conditions and conducting market analysis using data from reports, feedback, and social media content. It also facilitates valuable competitor strategy analysis and planning.

In marketing and sales strategy, LLM is utilized to identify target audiences, provide personalized product recommendations, and enhance the overall customer experience.

Automation and employee training processes are improved with LLM integration. It includes automated responses to emails, calls, and texts for troubleshooting or customer care services. Additionally, staff are provided with up-to-date information and real-time assistance for better service and quality assurance. However, there are limitations and challenges associated with the rapid development of this technology.

Challenges in integrating LLM into retail systems include handling massive volumes of data at high velocity, ensuring customization and contextual relevance, scalability, and maintaining data privacy and security. One major challenge is the sheer volume and velocity of data generated in retail operations. Robust data infrastructure with better storage and processing capabilities is essential to handle this data effectively. Customization and contextual relevance are crucial for the LLM to perform effectively in retail. This involves training the model specifically for retail operations and integrating real-time data integration tools and middleware for seamless data aggregation and processing. Scalability and performance are also significant concerns. Retail operations often experience fluctuating demands, requiring a scalable architecture that can dynamically adjust resources. Cloud-based solutions are increasingly adopted for their scalability benefits. Data privacy and security are paramount when dealing with sensitive customer information and business data. Strong data encryption, access control, and adherence to data protection regulations are vital to safeguarding data. Raising awareness of LLM and its operation throughout the company is another challenge. Solutions include collaboration, modular design, computing distribution, and partnerships to scale solutions effectively and ensure everyone involved understands and supports the integration of LLM into retail operations.

Protecting customer data in LLM-integrated retail systems involves several key measures. Firstly, all sensitive information must undergo data encryption, both at rest and in transit, employing end-to-end encryption to ensure data security. Access control and authentication mechanisms should be implemented, including role-based access control and strong authentication, to ensure only authorized personnel have access to sensitive data. Advanced threat detection systems should also be in place to recognize patterns early and detect any potential threats. Additionally, data backup and disaster recovery strategies are necessary to prevent further damage or loss of data in the event of a breach or disaster. Data anonymization and pseudonymization techniques should be employed to mask true data and remove personally identifiable information when unnecessary, utilizing fake identifiers or pseudonyms. Regular audit checks and compliance checks should also be conducted to ensure adherence to data protection regulations.

In terms of long-term business strategies, LLM-integrated retail systems should focus on enhancing the customer experience, leveraging data-driven decision-making, improving operational efficiency, fostering innovation and marketing initiatives, building a future-ready business with scalability and investment in technology, expanding into global markets, and prioritizing better training and development alongside sustainability and social responsibility efforts.

Implementing an LLM-based chatbot for personalized shopping experiences and assistance can have several impacts. Firstly, it can significantly enhance customer engagement by providing around-the-clock availability, personalized recommendations, and an interactive shopping experience. This can lead to increased customer satisfaction through features like troubleshooting, feedback collection, and human-like interaction. Moreover, such a chatbot can have a positive impact on business outcomes by narrowing down choices for customers and aiding them in making informed decisions, ultimately leading to increased sales and loyalty.

 Challenges in implementing LLM-based chatbots encompass various aspects such as data handling and processing, including the acquisition, cleaning, and organization of large datasets in e-commerce, alongside the need for real-time processing and response capabilities. System integration is critical to ensure compatibility with existing platforms, including database, UI, and backend, requiring API integration for inventory and financial management. Maintaining context and understanding multiple conversations pose additional challenges, as does ensuring scalability and reliability. Natural language understanding and generation must be sophisticated to provide accurate, appropriate, and engaging responses. Handling user data with care, addressing sensitive issues, and maintaining privacy are paramount. Continuous learning and updating, as well as balancing AI qualities with human intuition for user acceptance and experience, further contribute to the complexity of implementation.

Privacy and data security are critical considerations in the implementation of LLM-based chatbots, requiring attention to various aspects such as data collection, legal compliance, and user privacy, as well as robust security measures and continuous monitoring. It is essential to obtain user consent for data access and ensure transparency regarding its use while practicing data minimization to collect only necessary information, reducing the risk of privacy breaches. Compliance with data protection laws such as GDPR in the EU is imperative, along with implementing techniques like anonymization and pseudonymization to protect user data privacy and providing users with access, editing, and deletion rights. Security measures such as encryption, access control, and regular audits are essential to control data flow and mitigate vulnerabilities. Additionally, securing payment processing and compliance with industry standards like PCI DSS are crucial, along with continuous monitoring and rapid response to data and security breaches and suspicious activities related to both LLM and the website [25].

**Commercial Banking:** The research paper presents an insightful review of benchmarking local large language models (LLMs) within the domain of financial and economic texts. It introduces novel benchmarking tasks aimed at evaluating LLM performance and suggests areas for refinement to enhance local LLM effectiveness. These tasks include sentiment analysis, clarity and temporal orientation scoring for passages, and topic summarization.

The study compares the performance of local LLMs with closed LLMs like GPT-3 and GPT-4 across various tasks, such as labeling segments of Federal Open Market Committee (FOMC) statements and analyzing sentiment in financial texts. The findings indicate that local LLMs exhibit promise for general natural language processing (NLP) tasks in the financial and economic domain.

Furthermore, the research delves into the analysis of bank earnings calls using local LLMs, examining sentiment, temporality, and vagueness in post-pandemic era calls, including those during the banking stress of early 2023. This exploration aims to shed light on the dynamics between banks and investors during earnings calls, highlighting the potential of local LLMs for financial analysis.

Overall, the study offers valuable insights into the performance of locally deployed LLMs compared to cloud-based models in interpreting financial language, contributing to our understanding of LLM effectiveness in financial NLP tasks.

A close-up of a document

Description automatically generated

Table 2: Overview of examined models.

This picture is a few of the LLM's used by business corporations after benchmarking.

**Benchmarking:** Benchmarking LLMs (Large Language Models) involves evaluating their performance against a set of standardized tasks or metrics to assess their capabilities and compare them with other models or benchmarks. This process helps researchers and practitioners understand the strengths and weaknesses of LLMs, identify areas for improvement, and measure progress over time.

Benchmarking LLMs typically involves tasks such as text generation, language understanding, translation, summarization, sentiment analysis, question answering, and more. Researchers may use existing benchmark datasets or create new ones tailored to specific applications or domains. Evaluation metrics like accuracy, perplexity, BLEU score, ROUGE score, F1 score, and human evaluation may be employed to quantify LLM performance.

Benchmarking LLMs allows for fair comparisons between different models and facilitates advancements in natural language processing (NLP) research. It also provides insights into the capabilities of LLMs across various tasks and domains, guiding the development of more robust and effective language models.

Examples of benchmarking LLMs include GLUE (General Language Understanding Evaluation), which assesses NLP models' performance across diverse language understanding tasks like sentiment analysis, textual entailment, and question answering. SuperGLUE, an extension of GLUE, offers more challenging tasks to further test the capabilities of NLP models. SQuAD (Stanford Question Answering Dataset) evaluates the ability of models to answer questions based on given text passages. CoNLL (Conference on Computational Natural Language Learning) organizes shared tasks annually, assessing various NLP aspects such as named entity recognition, coreference resolution, and syntactic parsing. These benchmarks help gauge LLM performance and advancements in natural language understanding [25].

**Healthcare:** In healthcare, three main privacy issues regarding LLMs include data breach, inaccurate synthetic data, and bias prediction. According to our research, data breach is a significant concern, especially in the context of the European General Data Protection Regulation (GDPR). Healthcare data breaches can compromise data integrity, with patient information containing highly personal details like medical history and treatments being at risk. Breaches also violate strict healthcare regulations like HIPAA, leading to legal repercussions. Additionally, breaches may compromise patient safety if medical records are altered. These concerns highlight the importance of robust privacy measures in healthcare setting.

## Solution Space

Given the rising threat landscape of cyberattacks across industries, coupled with the increasing adoption of Large Language Models (LLMs) for operational efficiency, securing LLMs has become imperative to mitigate risks such as data breaches, financial losses, and theft. Our solution approach is tailored to address these challenges by assessing the risk and analyzing potential LLM attacks across industries. By leveraging datasets such as Vocabulary for Event Recording and Incident Sharing (VERIS), Statistics of US Businesses (SUSB), and frameworks like the MITRE Corporation's Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS), our study aims to develop risk measures and insights to address vulnerabilities in LLM deployments.

Our solution involves a comprehensive risk assessment and analysis process, utilizing diverse datasets and frameworks to understand the threat landscape surrounding LLMs. By gathering information on attack actors, assets affected, actions undertaken, and the impact of attacks, we aim to gain deep insights into potential LLM-related vulnerabilities and attack vectors across industries. This involves a detailed examination of historical attack data, trends, and patterns to identify potential risks and mitigate them proactively.

To facilitate our risk assessment and analysis, we employ advanced methodologies and tools, including Lorenz's curve, to quantify and visualize the distribution of risks associated with LLM attacks. By leveraging statistical techniques and data visualization methods, we aim to provide stakeholders with a clear understanding of the likelihood and potential impact of LLM-related attacks, enabling informed decision-making and risk mitigation strategies.

The outcome of our risk assessment and analysis provides valuable insights and foresight into possible LLM attacks and related vulnerabilities across industries. By identifying high-risk areas and potential attack vectors, our study enables stakeholders to prioritize resources and implement targeted security measures to mitigate and divert potential LLM attacks effectively. Additionally, our findings serve as a crucial foundation for future research on analyzing and predicting LLM attacks, contributing to the development of more robust and resilient security frameworks for LLM deployments.

Ultimately, the result of our study serves as a cornerstone for enhancing the security of LLMs and mitigating the risks associated with their deployment. By providing stakeholders with actionable insights and risk measures, we aim to empower organizations to safeguard their digital assets and mitigate the impact of potential LLM attacks. Furthermore, our study lays the groundwork for future research and innovation in the field of LLM security, fostering collaboration and knowledge-sharing to address emerging threats and challenges effectively.

## Project Objectives

Upon completing this project, the team anticipates gaining a comprehensive understanding of the security risks and vulnerabilities associated with LLMs across various industries. Through the analysis of datasets such as Vocabulary for Event Recording and Incident Sharing (VERIS), Statistics of US Businesses (SUSB), and frameworks like the MITRE Corporation's Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS), the team expects to uncover patterns, trends, and insights into potential LLM-related attacks. Additionally, the team aims to learn about advanced methodologies and tools for risk assessment and analysis, such as Lorenz's curve, to quantify and visualize the distribution of risks associated with LLM attacks.

The team aims to deliver a robust and comprehensive solution for enhancing the security of LLMs and mitigating the risks associated with their deployment. By conducting a thorough risk assessment and analysis, leveraging advanced methodologies and tools, the team expects to develop actionable insights and risk measures to address vulnerabilities in LLM deployments across industries. The solution is intended to provide stakeholders with a clear understanding of potential LLM-related attack vectors and enable informed decision-making and risk mitigation strategies.

After completing this project, the team expects to achieve a deep understanding of the multifaceted challenges and complexities surrounding the security of LLMs. By analyzing historical attack data, trends, and patterns, the team anticipates gaining insights into the evolving threat landscape of LLM-related attacks and vulnerabilities. Additionally, the team aims to develop a nuanced understanding of the impact of LLM attacks on various industries and the potential consequences for businesses, organizations, and society at large.

As a product of this project work, the team intends to provide valuable insights and risk measures to stakeholders across industries, enabling them to enhance the security of their LLM deployments and mitigate the risks associated with potential attacks. By delivering actionable insights and foresight into possible LLM attacks and vulnerabilities, the team aims to empower organizations to safeguard their digital assets, maintain trust and integrity, and minimize the impact of security breaches. Additionally, the team expects to contribute to the advancement of knowledge and innovation in the field of LLM security, fostering collaboration and knowledge-sharing to address emerging threats and challenges effectively.

## Primary User Stories

This section outlines the primary user stories developed to guide the project based on the user context and value proposition. Each user story explicitly states what the project is attempting to address and serves as a guiding principle for the project's objectives and deliverables.

"As a stakeholder in the cybersecurity domain, I want to assess the risk associated with Large Language Models (LLMs) and analyze potential LLM attacks across industries to develop actionable insights and risk measures."

"As a researcher in artificial intelligence and cybersecurity, I want to gain a deep understanding of the multifaceted challenges and complexities surrounding the security of LLMs, including the evolving threat landscape and potential consequences for businesses and industries."

"As a provider of LLM-based solutions, I want to deliver value to stakeholders by enhancing the security of LLM deployments, mitigating risks associated with potential attacks, and providing actionable insights and foresight into possible LLM attacks and vulnerabilities."

"As a member of the academic and research community, I want to contribute to the advancement of knowledge and innovation in the field of LLM security by fostering collaboration, sharing insights, and developing best practices for addressing emerging threats and challenges effectively."

The primary user stories outlined above serve as guiding principles for the project, delineating the key objectives and outcomes to be achieved. By addressing the needs and expectations of stakeholders across industries, the project aims to deliver value, enhance security, and contribute to the advancement of knowledge and innovation in the field of LLM security. These user stories provide a framework for the subsequent sections of the report, detailing the methodology, findings, and recommendations to fulfill the project's objectives.

## Product Vision

This section presents two solution scenarios outlining practical applications and benefits of our solution in addressing security challenges across different industries. Each scenario illustrates the context, stakeholders involved, and unique value proposition of our solution in enhancing security and mitigating risks associated with Large Language Models deployments.

## Scenario #1: Enhancing Security in Financial Institutions

For: Financial institutions, including banks, investment firms, and insurance companies.

Who: Risk management professionals, cybersecurity experts, and IT administrators.

The: Solution provides a comprehensive framework for assessing the security risks associated with LLMs deployed in financial institutions.

Is a: Proactive approach to identifying and mitigating potential LLM-related attacks, such as manipulation of financial data, fraudulent transactions, or market manipulation.

That: Empowers stakeholders to safeguard sensitive financial assets, maintain regulatory compliance, and protect against financial losses due to LLM-related security breaches.

Unlike: Conventional security measures, our solution offers targeted insights and risk measures tailored to the unique challenges and complexities of LLM deployments in the financial sector.

Our product: Enables financial institutions to proactively address security risks, enhance operational resilience, and maintain trust and confidence among customers and stakeholders.

Caveats: While our solution provides valuable insights and risk measures, it is essential for financial institutions to complement it with robust security protocols, ongoing monitoring, and collaboration with industry stakeholders to mitigate evolving threats effectively.

## Scenario #2: Strengthening Healthcare Data Security

For: Healthcare organizations, including hospitals, clinics, and pharmaceutical companies.

Who: Healthcare IT professionals, data privacy officers, and compliance officers.

The: Solution offers a robust framework for assessing and mitigating security risks associated with LLMs utilized in healthcare data processing and analysis.

Is a: Strategic initiative to safeguard patient data, protect medical records, and ensure compliance with stringent data privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act).

That: Empowers healthcare organizations to detect and mitigate potential LLM-related security breaches, including unauthorized access to patient records, data manipulation, or dissemination of sensitive medical information.

Unlike: Traditional cybersecurity measures, our solution provides tailored insights and risk measures specific to the healthcare industry, addressing the unique challenges and regulatory requirements of healthcare data security.

Our product: Facilitates proactive risk management, enhances data security, and safeguards patient privacy, thereby bolstering trust and confidence in healthcare services and maintaining compliance with regulatory mandates.

Caveats: While our solution offers valuable risk mitigation strategies, healthcare organizations must implement comprehensive security measures, employee training, and regulatory compliance frameworks to effectively protect patient data and mitigate security risks associated with LLMs.

# Datasets

## Overview

The project utilizes three main datasets to analyze LLM attacks and assess associated risks across industries:

North American Industry Classification System (NAICS) code:

Description: The North American Industry Classification System (NAICS) is used to categorize businesses into industry sectors based on their primary activities.

Usage: NAICS codes provide a standardized method for identifying industries, allowing for the classification and comparison of businesses across different datasets.

Vocabulary for Event Recording and Incident Sharing (VERIS):

Description: VERIS contains records of various cyber incidents, including LLM attacks, technical malfunctions, and compromises, providing detailed information on attack types, affected assets, and impact.

Usage: VERIS serves as the primary source for understanding the characteristics of LLM attacks, enabling analysis of attack patterns, vulnerabilities, and their impact across different industries.

Statistics of US Businesses (SUSB):

Description: SUSB offers statistical data on US businesses, including industry, size, and location, providing insights into sectors most susceptible to cyber threats.

Usage: SUSB data contextualizes LLM attacks by identifying industries and business sizes most affected by cyber threats, aiding in understanding the landscape of potential targets.

MITRE Corporation Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS):

Description: ATLAS is a framework developed by MITRE Corporation, offering insights into adversarial threats targeting AI systems, including LLMs, and providing recommended countermeasures.

Usage: ATLAS provides a structured approach to understanding LLM vulnerabilities and recommends effective countermeasures for securing these systems, aiding in the development of risk mitigation strategies.

Web Scraping:

Description: In addition to structured datasets, the project employs web scraping techniques to gather real-time data from news articles related to LLM attacks.

Usage: Using the feed parser library, news articles containing keywords such as "LLM attacks" or "LLM vulnerabilities" are scraped from Google RSS and the Bing API. These articles provide timely insights into current events, public interests, and societal issues related to LLM security. The scraped data is combined to visualize changes in the frequency of news articles over time, indicating awareness and trends in LLM attacks.

These datasets collectively provide a comprehensive view of LLM attacks, allowing for detailed analysis of attack patterns, vulnerabilities, and potential impacts across various industries.

* 1. **Field Descriptions**

### NAICS Code

The North American Industry Classification System (NAICS) code is a standardized classification system used to categorize businesses into industry sectors based on their primary activities. Developed jointly by the United States, Canada, and Mexico, NAICS codes provide a hierarchical structure that organizes industries into increasingly specific categories. Each NAICS code consists of a six-digit numerical code that uniquely identifies a particular industry sector, subsector, or industry group. The system is periodically updated to reflect changes in the economy and emerging industries, ensuring its relevance and accuracy.

In the context of this project, the NAICS code field serves as a crucial identifier for businesses and industries across different datasets. By assigning NAICS codes to businesses, analysts can classify and compare data from various sources, enabling detailed analysis of industry-specific trends, vulnerabilities, and risks. Additionally, NAICS codes facilitate cross-referencing between datasets, allowing researchers to contextualize cyber threat data with broader economic and industrial information. Overall, the inclusion of NAICS codes enhances the project's ability to provide comprehensive insights into LLM attacks and their impacts on different sectors of the economy.

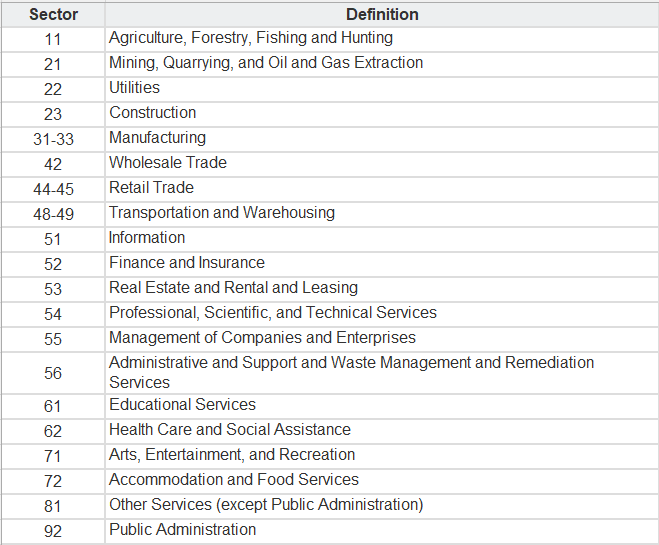


Table 3: NAICS code with definitions.

### VERIS Dataset

This dataset, as mentioned earlier, is a public repository that contains over 7,000 records and over 2500 fields of features. This dataset contains parameters including the type of damage, type of attack, countries affected by the attack, the location from which the attack was initiated, and extends with further parameters.

This dataset also presents the Vocabulary for Event Recording and Incident Sharing (VERIS) Community Database, an open-source repository maintained by the Verizon RISK team. The VERIS schema allows for documenting security incidents with variables like incident tracking, victim demographics, incident description, and impact assessment. An example incident record in JSON format from the VERIS dataset is shown, highlighting the structured nature of the data and the complexity of information captured. There are also four A ’s mentioned in here stating that (Actors, Actions, Assets, Attributes).

ACTOR: Actors refer to the individuals, groups, or entities that are responsible for the initiation of a security event or incident.

ACTIONS: Actions are the methods or tactics employed by actors to compromise, disrupt, or gain unauthorized access to information systems. Actions can range from malware deployment.

ASSETS: Assets are the targets of the actors' actions; these are the data, devices, systems, or networks that hold value to an organization and, therefore, to an adversary. Assets can include hardware (like servers and workstations), software (like databases and applications), and data (such as confidential business information or personal data).

ATTRIBUTES: Attributes refer to the properties or characteristics that describe the consequences or outcomes of security incidents. This can include the type and extent of damage (like data loss, service disruption, or financial impact), the scope of the incident (such as the number of systems affected), and the method of discovery.

Here is a table showing VERIS Dataset Fields Description.



Table 4: VERIS Dataset Fields Description.

### VERIS Data Exploration

The approach to understanding and exploring the dataset was adopted using a few different techniques, using different tools such as Excel spreadsheets, Python, R, and MongoDB. While the latter is still in development, the first two were introduced for preliminary analysis.

Through Excel spreadsheets, we were able to retrieve the frequency of attacks on various industries. Through our initial results, we witnessed extensive attacks on educational industries, affecting their respective businesses. Closely following the financial and healthcare sectors. In the initial study, it was also revealed that types of attacks on the education sector were mostly the same, and the number of attacks was malware. It is because those counts are the number of victimized companies from the same incidents, MOVEit Cyber Attack incident, that happened in May 2023 [26] .

A screenshot of a computer

Description automatically generated

Table 5: Frequency of attacks on various industries.

Data cleaning is one of the most important aspects of any data analysis study. The task may not seem of high significance or level; however, with the cleaned and more uniform data, it becomes easier to manipulate and work on the dataset. A similar approach was adopted by us to get an initial impression of the dataset based on a single industry. For this time, we chose the consumer retail industry by considering NAICS Code 45. We further worked on eliminating columns that were insignificant for this specific industry.

We finally came across a reduced data frame with certain crucial columns elaborating on the attack and other reporting leading to additional understanding of incidents.

A screenshot of a computer

Description automatically generated

Table 6: Reduced data frame with information for the attacks.

A screen shot of a graph

Description automatically generated

Figure 4: Frequency of attacks on various sectors.

A colorful graph with text

Description automatically generated with medium confidence

Figure 5: Frequency of different types of attacks on various sectors.

A screenshot of a graph

Description automatically generated

Table 7: Additional research and data exploratory analysis. (1971-2023).

We kick off the additional analysis of the VERIS dataset by comparing the count of incidents across various industries. The first impression shows healthcare, the public (government), and education to be hit with the highest number of incidents. Although this indicator may not be sufficient to understand the whole picture and gain an overview of all cyberattacks, we intend to go further and create a new column calling it 'impact rate.' The new column intends and successfully manages to show the standard measure of impact by achieving the ratio of the number of impact indicators divided by the number of incidents. It is clearly visible that all the incidents don't necessarily have measures or impact indicators. Although the reasons may vary, including the common NA values present in the dataset, the impact indicators do provide a sense of the severity of the attack or at least acknowledgement of the impact that took place. The impact rate is almost contrary to the number of incidents; the highest impact rate can be noticed in the agriculture sector. Meanwhile, the healthcare sector with the highest number of incidents isn't the one with the highest impact factor. A straight-forward explanation for such a difference in these scenarios is the number of attacks attempted themselves. The impact indicator may suggest the attacks that were discovered and that were able to measure the extent of the damage from such attacks. As compared to the public sector, which has a considerably lower impact rate, although it may have almost similar numbers for the number of indicators as healthcare, we may deduce that the impact indicators are lacking in numbers, and perhaps this may conclude that an attack might have been overlooked or have been less detected.

In essence, we believe that the impact factor varies as per the industries. Although the number of attacks may reveal the most vulnerable industry, the impact factor may suggest the industry with the least protection or with the least investment in the security infrastructure.

The LLM security project holds a specific designation when it comes to cyber security. We know that in today's day and age, technology has been developing at a rapid pace. It is certainly impossible to witness any past methodologies being adopted in our daily work. The old methods and technologies have become obsolete. Since the early inception of the Internet in the mid-80s, the development of a better ecosystem and environment for work with the ease of working and accessing information from almost anywhere has made its evolution distinctive and certainly crucial. We have seen that countless businesses across almost every industry have adopted the use of the internet and digital devices and setups. The recent development of such technologies has seen serious changes due to the increase in the active population on the internet. Other forms of technology have been developing, from the nascent stages of machine learning to the most recent Generative Predefined Transformers, also known as GPT. The use of large language models has been extensive. Thanks to its capability to train the model on an enormous scale of data, process the data, understand the context of the prompts, and provide an appropriate response.

Since technology isn't just sunshine and rainbows, with inappropriate intent and with the use of programming codes, it has become vulnerable to attacks. These attacks are intended to have a certain goal, target, and aim. These attacks affect individuals, organizations, government agencies, and many other sectors. It is crucial to understand the importance of such an attack and embrace countermeasures.

Large datasets provided by private companies, public datasets, and research groups provide useful information based on the type of attack, vulnerable entities, and damage caused by the outer party. It depends on the kind of dataset used to understand the attacks and act accordingly.

In this first phase of data analysis in LLM, we were provided with a dataset by Verizon called the VERIS dataset. This is a dataset that contains records of various events, particularly cyber-attacks, technical malfunctions, and compromises. In addition to the dataset, we scraped articles from the web to find information based on articles citing ChatGPT and any attacks or compromises related to it.

News articles serve as an excellent indicator of current events, public interests, and societal issues, offering a rich, timely, and diverse dataset for analysis. They provide quantitative data—such as publication frequency and temporal distribution—and qualitative insights, including content themes and narrative styles.

There are numerous APIs to scrape news articles, but most of them are either paid or don’t have historical data. Hence, we opted to scrape news articles using Google RSS.

RSS, formatted as plain XML text, delivers concise summaries of recently published content from various providers like news outlets, podcasts, and blogs. It's predominantly used by news publishers to distribute updates. The purpose of an RSS feed is to facilitate access to the newest information, serving news aggregators and syndicators, among others. Typically, an RSS feed does not include the full text of articles; instead, it offers essential details such as the author's name, the title, a brief description, and the publication date.

For our project, we have used Google RSS. Using the `feedparser` library, we were able to scrape news articles that had these keywords: “LLM attacks” or “LLM vulnerabilities." We also scraped news articles from the Bing API and then combined them to visualize how the count of news articles changed over time. Interestingly, after the release of ChatGPT, there were more articles published, which indicated that there is a rise in LLM attacks and that there is awareness about the same.

### Classifying Ambiguous Industry Labels into Defined Categories

We learned a lot and faced a number of obstacles while investigating the use of machine learning algorithms to categorize victim industries within a cybersecurity dataset. The project started with the deliberate selection of 15 columns that were thought to be essential for determining the victim's industry from the incident's features. The importance of each feature in capturing the essence of cybersecurity incidents, their patterns of occurrence unique to the industry, and their anticipated influence on model discernibility served as a guide for this selection. These included markers of activities (e.g., malware, hacking), impacted characteristics (e.g., availability, confidentiality, integrity), and incident specifics (e.g., attack technique, impact intensity). This rigorous selection approach was justified by the need to provide the prediction model with a rich but targeted dataset that would improve its capacity to correctly categorize sectors based on incident facts.

The exercise sought to determine the best classifier for our particular problem by utilizing a number of machine learning models, including Gradient Boosting, Random Forest, Support Vector Machines, and Logistic Regression. Because it achieved a compromise between computational feasibility and prediction accuracy, the Gradient Boosting Classifier outperformed other models. The model was subjected to a meticulous training procedure using a carefully selected subset of the data, with the entries labeled "Unknown" and "Other Services" specifically excluded to enhance its ability to learn from correctly classified instances.

The "Unknown" and "Other Services" portions of the dataset were reclassified using the learned Gradient Boosting model in the following step. The transformation of these loosely defined categories into more analytically useful and descriptive industrial classifications was made possible in large part by this endeavor. The prediction project highlighted the revolutionary potential of machine learning in enhancing data categorization in addition to showcasing the model's usefulness.



Figure 6: Victim Industry names before classification.

An actual representation of the dataset's improved status was given by a visualization of the industry distribution post-classification, which showed the newly anticipated industry classes and their corresponding prevalences. This visual aid played a crucial role in demonstrating the impact of reclassification and highlighting the dataset's enhanced analytical usefulness.

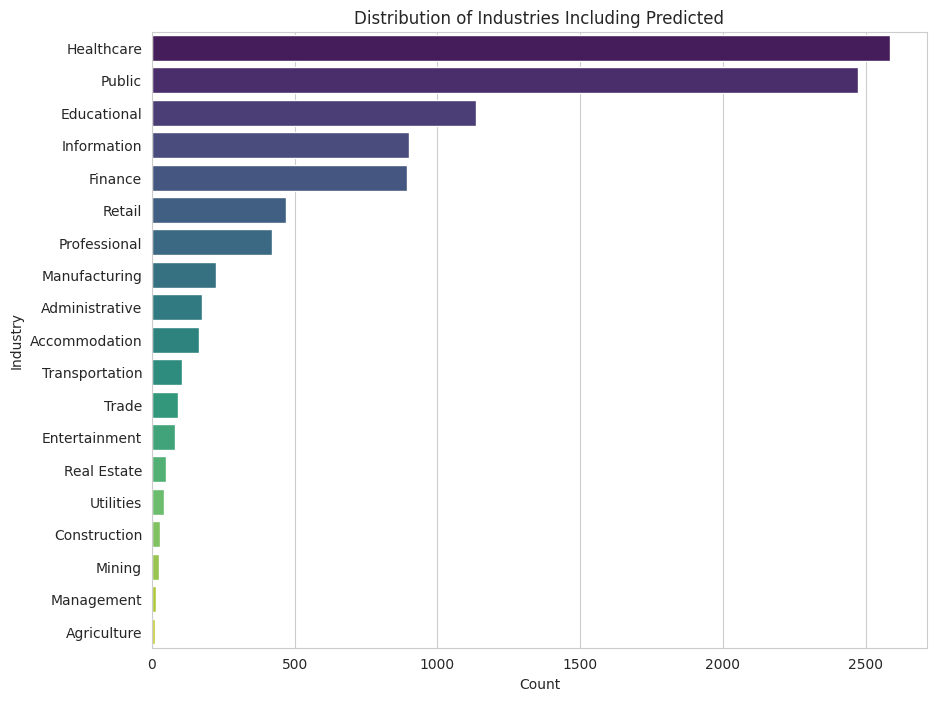


Figure 7: Victim Industry names after classification.

Notwithstanding these developments, the expedition brought to light a crucial obstacle: the model's comparatively low accuracy. Even though the Gradient Boosting Classifier performed the best out of all the models tested, its accuracy of 42% made it clear that the model was insufficient for reliable, real-world applications. This constraint raises important questions about the model's effectiveness and the difficulty of making reliable industry predictions based only on incident characteristics. The low accuracy draws attention to the nuanced nature of cybersecurity incidents and their classification; it also raises the possibility that the dataset itself has inherent flaws, such as class imbalances or inadequate representation of particular industries, or that the features chosen, while informative, may not fully capture the range of industry-specific patterns.

Taking these insights into account, a few things were clear. Important steps in the modeling process included coding categorical variables, addressing missing data, and attempting to strike a balance between interpretability and model complexity. The main finding, however, was that in order to overcome the accuracy plateau, a more advanced or customized modeling method was required.

### Statistics of U.S. Businesses (SUSB) dataset

The 2021 SUSB dataset or "Statistics of U.S. Businesses" (SUSB) is a program conducted by the U.S. Census Bureau [27]. The dataset provides annual data on the number of establishments, employment, and payroll by industry for the United States as a whole, as well as for individual states and metropolitan areas based on the 2017 North American Industry Classification System (NAICS) codes. This dataset will be used to perform and incorporate together with VERIS dataset MITRE ATLAS dataset and NAICS code for our Final Risk Assessment.

Dataset contains:

State: Numerical code representing the U.S. state.

State Name: The name of the U.S. state.

NAICS: The North American Industry Classification System code for the industry.

NAICS Description: The description of the industry corresponding to the NAICS code.

Number of Firms: The total number of firms within the industry and state.

Establishments: The total number of business establishments within the industry and state.

Employment: The total number of employees working within the industry and state.

Employment Noise Flag: A flag indicating the reliability of the employment data (e.g., 'G', 'H').

Annual Payroll ($1,000): The total annual payroll for the industry and state, expressed in thousands of dollars.

Annual Payroll Noise Flag: A flag indicating the reliability of the payroll data (e.g., 'G', 'J').

A screenshot of a document

Description automatically generated

Table 8: Statistics of U.S. Businesses (SUSB) Annual Datasets.

The North American Industry Classification System (NAICS) [28] is a standardized classification system used by statistical agencies in the United States, Canada, and Mexico to classify businesses and industries. It provides a hierarchical structure for organizing and analyzing economic data based on similarities in production processes, products, and services. NAICS codes are six-digit numerical codes assigned to various industries and sectors of the economy. The codes are hierarchical. The first two digits represent the industrial sectors used in this project to classify and group businesses from the industrial sector's point of view.

### MITRE ATLAS dataset

The MITRE ATLAS dataset shows a comprehensive map or flow diagram of a potential LLM attack. Although the number of cases that showcase compromise on LLM is small as compared to traditional cyber-attacks, there has been study of possible ways in which attacks can be conducted.

This public knowledge base isn't a database but gives ideas about adversities, their tactics, and their techniques. The map shows a flow starting from 'Reconnaissance', which describes the ways in which an adversary will ponder the potential client's available public materials; this could be research materials or blogs to get an understanding of the client's or victim's background. This then proceeds with developing resources to achieve the objective. This may include gathering and developing codes and adapting various ways to poison datasets. The flow proceeds with gaining initial access and applies the methods of prompt injections, phishing, account validation, and more. Then the API access is used to reach the ML model and gain full access. Using LLM plugins to compromise the model and implementing various commands and scripts for executing the attack. This then proceeds with gaining executive and privileged access to various locations within the database and the model to alter the functions and disrupt the operations. Discovering files, documents, and essentially any crucial information. This then proceeds with the collection of data from local systems, gaining additional data from repositories, and artifact collection. The flow ends with exfiltration via API and cyber means and data leakage. The impact of such an attack causes the ML model to lose its operation, may cause external harm, denial of ML services, and more.  
A screenshot of a computer

Description automatically generated

Figure 8: MITRE ATLAS Dataset.

The heatmap we’ve provided visualizes the timeline of cyber-attack events in an organization, categorized by the stages of a cyber-attack: Compromise, Exfiltration, Discovery, and Containment. The y-axis represents the time from the initial attack in various units ranging from seconds to years, including "Never," while the x-axis represents the stages of the attack timeline. The colour intensity represents the count of incidents, with darker colours indicating higher frequencies.

A screen shot of a graph

Description automatically generated

Figure 9: Heatmap of Timeline Events.

This heatmap is showing 4 timeline stages; Compromise: Most compromises are discovered in terms of days and weeks, as indicated by the colour intensity in those cells. There are fewer cases discovered in hours, months, or years after the initial attack. The 'Never' category is not as prominently represented in this stage, which suggests that compromises are eventually discovered rather than going unnoticed indefinitely. Exfiltration: The data for this stage seems to be relatively sparse compared to other stages, suggesting that detecting exfiltration specifically may be more challenging, or it may not be applicable to all types of incidents. Discovery: This stage has a significant concentration of incidents discovered within weeks and months. This indicates a delay in the detection of cyber-attacks, which could be due to the time it takes for organizations to notice irregularities or the effectiveness of attackers at concealing their activities. Containment: There is a notable number of incidents that are contained within days, as evidenced by the colour intensity. However, the count for containment within weeks is the highest, and there are even some cases that take months or years to contain. This suggests that while some organizations can respond quickly to contain an attack, for many, it takes a longer period to fully resolve the issue.

Overall, the heatmap highlights that while some cyber-attacks are identified and contained relatively quickly (within hours or days), a significant number are only discovered and contained weeks or even months after the initial compromise. The visual also underscores the challenge organizations face in quickly detecting and responding to cyber threats.

### Web Scraping

This study shows how to effectively integrate cloud-based services with Python programming by using a Python-based method to exploit the Microsoft Bing News Search API for the extraction of news articles relevant to certain search queries. The first step was creating an Azure Resource Group, which was an important procedure that entailed setting up a Bing Search resource on the Microsoft Azure network. This step was critical since it produced the API subscription key, a special credential required for API request authentication.

After obtaining the subscription key, the investigation went on by creating a Python script that could communicate with the Bing News Search API. Utilizing the requests library, this script sent HTTP GET queries to the API with the following parameters: the target market, the search term "Privacy and security attacks on LLM," and a limit of 50 articles per request. These requests showed a skillful use of RESTful API consumption, as they were verified by HTTP headers and enhanced by query parameters. To guarantee error resilience during HTTP queries, the script additionally included strong exception handling techniques.

This script's main purpose was to parse the JSON response from the API and retrieve the essential details from each news story, such as the title, description, URL, and publication date. After that, the csv package in Python was used to meticulously classify this data into a CSV file called "news\_articles.csv." This study's component underlined the value of organizing and storing data in readily available forms, including CSV, in addition to its ability to get and interpret web-based data.

This study concludes with a thorough demonstration of the use of Python scripting in API integration, including the establishment of an Azure Resource Group and a Bing Search resource, as well as the effective retrieval and archiving of data. The focus on creating the Azure resource and using its search query API highlights how important cloud resources are to modern data management and retrieval processes. Additionally, the Python script's structure and capabilities show off the language's adaptability and effectiveness in the fields of academic data processing and online scraping.

A graph of a number of articles

Description automatically generated

Figure 10: Number of articles published per month.

A bar graph with different colored bars

Description automatically generated

Figure 11: Top 10 published by number of articles.

### Synthetic Data

This article aims to provide a high-level overview of the current state (\* As of May 2022) of works done on synthetic data technologies, with an emphasis on privacy.

This article draws attention to these key issues regarding synthetic data:

Synthetic data technologies hold immense potential across various applications, such as enhancing privacy, promoting fairness, and boosting data quality through augmentation. These technologies can significantly speed up development processes, facilitating quicker advancements in data-driven projects and reducing costs within software development cycles. However, despite these benefits, synthetic data also presents certain risks. Notably, it does not inherently guarantee privacy and may inadvertently expose underlying original data. Thus, meticulous attention is required during the generation of synthetic data to ensure it effectively fulfills its intended purposes while safeguarding privacy.

**Synthetic data doesn’t replace real data**, since synthetic data is a distorted version of the real data, the authors suggest that performing modelling and inferences on this data carries additional risks as decisions taken on such inferences can have real implications. Therefore, it’s advised that synthetic only be used to quickly accelerate the “research pipeline” but ultimately the tools that would be finally deployed in the real world must be evaluated and fine-tuned on real data instead.

**Outliers are hard to capture in a private manner via synthetic datasets**, Synthetic data generators usually struggle at replicating the statistics of outliers from the real-data or don not effectively conceal the private information of outliers during synthetic data creation.

**Empirically evaluating the privacy of a single dataset can be problematic.**

**Black box models can be particularly opaque when it comes to generating synthetic data.**

**Synthetic data goes beyond privacy**, and also serves as a promising tool to improve the fairness, biases and robustness of machine learning models. But more research is needed to fully understand the pros and cons of using synthetic data approaches for such purposes.

Access to big datasets (High Volume, Velocity & Variety) along with advanced statistical tools to extract insights, has the potential to accelerate research, innovation and inform decision making. But usually, such datasets tend to be highly sensitive (Ex. Health or Financial Data) and dissemination of such data may violate the fundamental rights of individuals and is guarded by privacy regulations (Ex. GDPR or CCPA) and many real-word datasets that are usually high-dimensional and sparse become vulnerable to privacy attacks due to difficulties of anonymization techniques in providing adequate anonymity. Therefore, such limitations create a severe bottleneck on innovations and developments in machine learning and data science methods.

Synthetic data generated by a model can serve as a substitute for highly sensitive real datasets to some extent. By manipulating the data generation process users can control the amount of private information present in the synthetic data and its resemblance to the original data (\* at least in principle). When used responsibly, synthetic data promises the development of machine learning methods from datasets that are sensitive (privacy is important), incomplete, scarce or biased: quickly prototyping data-driven models, verifying and validating ML pipelines, assuring of real-world performance (\* to some extent) and also fuels responsible innovation providing digital sandbox environment to test new algorithms or ideas in a less risky and ethical manner.

Even with such promising capabilities of synthetic data, it comes with its own set of challenges and risks. Synthetic data generation is still a burgeoning area of research and lacks systematic frameworks for its safe and responsible implementation.

### What is Synthetic Data?

At the time of writing this article there isn’t yet a widely accepted definition, the authors propose their definition for synthetic data encompassing the range of methods to and applications of synthetic data:

*“*Synthetic data is data that has been generated using a purpose-built mathematical model or algorithm, with the aim of solving a (set of) data science task(s).*”*

Real data are generated by data measurement systems of the real world (Ex. financial transactions, satellite images, medical tests). Synthetic data generator models can take many forms from deep learning architectures (Ex. Generative Adversarial Networks (GANs) or Variable Auto-Encoders (VAEs)), Agent-based and econometric models, stochastic differential equation modeling of physical or economic systems, etc.

### Can Synthetic Data Replace Real Data?

Use of Synthetic data for real-world use cases raises two key questions: Can we do the same things **with** (like building models, data analysis, hypothesis testing, etc.) synthetic data that we do with real data? Can we do the same things **to** (like linking disparate datasets, extension of synthetic datasets when the original data gets extended, etc.) synthetic data that we do to real data?

Studies have found that synthetic cannot simply be used in place of real data; rather must be adjusted appropriately according to the use case. Often with privacy sensitive data, there is a tradeoff between privacy concealment and fidelity(accuracy) of the synthetic data, and therefore must be generated meticulously. Conclusions drawn from data analysis and hypothesis testing done on synthetic tend to be weaker than when done using the real data, and hence the statistical significances of such conclusions must be adjusted accordingly.

Synthetic data generated on real- private data must be de-biased before training models [34]. The performance of models trained on synthetic data is significantly improved when the model parameters are tweaked via Bayesian Inference than otherwise.

### Implementation of Synthetic Data.

In this project, we conducted preliminary tests using a sample dataset to ensure the robustness and effectiveness of our synthetic data generation approach before deployment on the VERIS dataset. The initial phase involved importing the dataset into a pandas Data Frame, followed by careful selection of key features that are indicative of potential risks and outcomes. This selective approach helps in focusing on relevant data, thereby improving the efficiency of the synthetic data generation process. Subsequent steps included preprocessing techniques such as label encoding to convert categorical data into a format suitable for machine learning models. This groundwork is critical as it establishes a controlled environment where we can fine-tune the parameters of our Generative Adversarial Networks (GANs) without compromising the integrity and privacy of the real data contained in the VERIS dataset.

The decision to first implement our methods on a sample dataset stems from the need to meticulously assess the potential risks associated with synthetic data, such as privacy leaks or the inadvertent replication of sensitive patterns. By using a sample dataset, we were able to experiment with various configurations of our GANs, optimizing them to produce high-quality synthetic data that retains the utility of the original data while minimizing privacy risks. This step is crucial in ensuring that the transition to the more sensitive and complex VERIS dataset is both seamless and secure. It allows for the identification and rectification of any issues in the synthetic data generation process, thereby ensuring that the final implementation is both efficient and compliant with privacy regulations.

In conclusion, while the application of GAN-generated synthetic data holds potential, it requires rigorous validation, a deep understanding of the data’s limitations, and careful consideration of ethical implications. The implementation of such technologies often encounters significant time constraints, impacting the depth of validation and testing that can be feasibly conducted within project timelines. For future research, it is recommended to focus on developing methods to verify the accuracy and utility of synthetic data and to explore hybrid models that might leverage both real and synthetic data to optimize performance without compromising the integrity and applicability of security analytics. Further efforts should also consider efficient methodologies that can reduce the time required for comprehensive testing and validation of synthetic datasets.

### List of Potential Use Cases of Synthetic Data

**Privacy Protection** and Provide a secure sandboxed environment for research (for rapid prototyping of ML models and software testing).

**Removal of bias present in original data.**

Use in Federated Learning for choosing the most effective algorithm by testing on synthetic data.  
**Data Augmentation for robustness** (Extending smaller datasets to provide robustness against outliers and linking of disparate datasets)

**De-biasing:** Synthetic data offers a uniform de-biasing approach of the real-world training data, on which multiple models can be trained and ensured that historical biases has been mitigated in the training stage itself rather than incorporating a bias mitigation approach post model training on a biased dataset. It’s also important to be mindful about the risks that are introduced when using synthetic data (as discussed in: Can Synthetic Data replace real data?)

Data Augmentation: In the field of computer vision, deep neural nets are the leading technology, which requires vast amounts of accurately labelled data which is often costly to produce from the real world. Hence, synthetically generated labelled datasets offer a cost-effective solution strategy to this challenge and have seen adoption by industries. For such applications privacy isn’t the main concern, and these synthetically generated datasets would be used alongside the real world labelled datasets.

### Utility, Fidelity, and Privacy of Synthetic Data

Synthetic Data generation is called upon when the original real world data is inappropriate or inadequate for the task (such as, the original data is privacy-sensitive, too small, or/and biased); So the synthetic data mustn’t be too similar to the original data as it would also suffer from the same issues as the original data, but at same time must be similar enough to accurately model the original data. This ‘permitted similarity’ (or required amount of dissimilarity) of the synthetic dataset depends on the task, which would come down to the acceptable levels of utility, fidelity and privacy of the resulting synthetic data for a given task(s).

**Utility**, defined as usefulness of the synthetic data for a given task(s), which can be determined via performance comparison between models trained on real versus synthetic data and perusing performance metrics (like accuracy, precision, root mean-squared error, and model fairness properties, including demographic parity, fairness through unawareness, or conditional fairness) of models trained on synthetic data usually doing so on a Train on Synthetic, Test on Real (TSTR) paradigm.

Fidelity is like utility but instead of indirectly comparing the synthetic data to its real counterpart via model performance metrics we do so directly. Therefore, fidelity captures how well is the synthetic dataset ‘statistically similar’ to its real dataset. By ensuring high fidelity, we may see improved performance markers across a wide range of tasks, where in an ideal case we’re looking to get a full statistical similarity (i.e. matching distributions) that would allow us to performing tasks on the synthetic data too that we could perform using the real data. But privacy concerns make this difficult to achieve and makes it undesirable when biases are present in the original data. Instead, we can aim to match low-dimensional marginals [83], ensuring syntactical accuracy of the synthetic data, or look at the conditional distribution of unbiased features on a biased feature.

**Utility vs Fidelity:** Utility and Fidelity may seem similar, but in all cases may not be positively correlated, in some cases one of fidelity or utility could be reduced leaving the other unaltered and providing room for improved privacy.  
Privacy, the amount of information that the synthetic data reveals about the real data that it was produced from.  
Privacy vs Fidelity, usually it’s the case that fidelity increases at the expense of privacy. We may generate multiple synthetic datasets for each use case that requires a minimum amount of fidelity with user specified privacy concealments [29].

* 1. **Data Context**

In our project, data context plays a vital role in understanding and analyzing LLM attacks and associated risks across industries. The datasets we utilize provide a rich context surrounding cyber incidents and business statistics, enabling us to glean valuable insights into the landscape of cyber threats facing organizations today.

For example, the VERIS dataset offers detailed records of various cyber incidents, including LLM attacks, technical malfunctions, and compromises. By examining these incidents within the broader context of industry sectors, business sizes, and attack types, we can identify trends, vulnerabilities, and potential impacts with greater accuracy.

Moreover, our use of external sources, such as news articles scraped from Google RSS and the Bing API, adds further context to our analysis. These articles provide real-time information on LLM attacks, public perceptions, and societal issues related to cybersecurity, allowing us to stay abreast of current events and trends in the cybersecurity landscape. By incorporating these external sources into our analysis, we can develop a more comprehensive understanding of the factors influencing LLM attacks and make informed decisions to mitigate risks effectively.

* 1. **Data Conditioning**

Data conditioning is a critical step in our project's data analysis pipeline, aimed at preparing the raw datasets for meaningful analysis. Given the diverse sources and formats of our data, this process involves several key tasks to ensure consistency, accuracy, and compatibility across datasets.

Firstly, we perform data cleaning to address any inconsistencies, missing values, or errors present in the datasets. For example, in the VERIS dataset, we may encounter incomplete incident records or inconsistent formatting, which need to be standardized for further analysis. Additionally, we apply data transformation techniques to restructure the data and extract relevant features for analysis.

Once the data is cleaned and transformed, we proceed with data integration, combining multiple datasets to enrich our analysis. For instance, we merge the VERIS dataset with the Statistics of US Businesses (SUSB) dataset to contextualize cyber incidents within specific industry sectors and business sizes.

Finally, data normalization is applied to ensure that all datasets are on a consistent scale and format, facilitating accurate comparisons and analysis. This step is particularly important when dealing with numerical data, such as incident frequencies or business statistics, to avoid biases in the analysis process.

By effectively conditioning our data, we can ensure that our analysis is based on reliable and consistent information, leading to more accurate insights and informed decision-making regarding LLM attacks and associated risks.

* 1. **Data Quality Assessment**

Our project assesses the datasets using a set of attributes to ensure data quality and reliability. These attributes include completeness, uniqueness, accuracy, atomicity, conformity, and overall quality.

Completeness: We evaluate whether all necessary information is present in the datasets. This involves checking for missing values, ensuring that essential fields are populated, and addressing any gaps or inconsistencies.

Uniqueness: We examine the uniqueness of records within the datasets to avoid duplicate entries or redundant information. This ensures that each record provides distinct and valuable insights for analysis.

Accuracy: We verify the accuracy of the data by cross-referencing it with trusted sources or validating it against known standards. This involves identifying and correcting any errors, discrepancies, or outliers that may compromise the integrity of the analysis.

Atomicity: We ensure that each data point within the datasets is indivisible and represents a single, coherent concept. This helps maintain data integrity and prevents confusion or ambiguity during analysis.

Conformity: We assess whether the data conforms to predefined standards, formats, or expectations. This involves checking for consistency in data formatting, naming conventions, and adherence to data schema or specifications.

Overall Quality: Finally, we conduct an overall assessment of the data quality, taking into account all the aforementioned attributes. This holistic evaluation helps identify areas for improvement and ensures that the datasets meet the required standards for robust analysis.

By systematically assessing our datasets against these attributes, we can identify and address any potential issues or weaknesses, ensuring that our analysis is based on high-quality, reliable data.

* 1. **Storage Medium**

For our project, we have utilized GitHub repositories as the primary storage medium for our datasets and project files. GitHub provides version control, collaboration tools, and accessibility features essential for managing our project data effectively.

Version Control: GitHub's version control system allows us to track changes made to our datasets and project files over time. This ensures that we can revert to previous versions if necessary and maintain a comprehensive history of modifications made by team members.

Collaboration: GitHub enables seamless collaboration among team members, allowing multiple contributors to work on the project simultaneously. Through features like pull requests and code reviews, we ensure that changes to the datasets are reviewed and approved before being merged into the main repository.

Accessibility: With GitHub, our datasets and project files are accessible to team members from anywhere with an internet connection. This facilitates remote work and enables team members to access and contribute to the project effortlessly.

Data Partitioning: Within the GitHub repository, we organize our datasets into separate directories or branches based on relevant attributes such as date, industry sector, or incident type. This helps in managing and accessing different segments of the data efficiently.

Security Measures: GitHub offers security features such as access controls, encrypted connections, and two-factor authentication to protect our data from unauthorized access or breaches. We ensure that sensitive information within our datasets is appropriately encrypted, and access is restricted to authorized team members only.

By utilizing GitHub repositories as our storage medium, we ensure that our datasets are securely stored, versioned, and accessible to all team members. This allows for efficient collaboration, data management, and analysis throughout the project lifecycle.

* 1. **Storage Security**

Ensuring the security of our data is paramount in our project. We implement various measures to maintain the confidentiality, integrity, and availability of our datasets stored in GitHub repositories.

Access Controls: We enforce access controls on our GitHub repositories to restrict access to authorized team members only. This includes using strong, unique passwords and employing two-factor authentication (2FA) to prevent unauthorized access to the repository.

Encryption: GitHub ensures data security through encryption mechanisms. All data transmitted between clients and GitHub servers is encrypted using secure protocols (HTTPS) to prevent eavesdropping or tampering during data transfer.

Regular Backups: We maintain regular backups of our datasets to mitigate the risk of data loss or corruption. These backups are securely stored in separate repositories or external storage locations, ensuring redundancy and data availability in case of emergencies.

By implementing these storage security measures, we ensure that our datasets stored in GitHub repositories are adequately protected against unauthorized access, data breaches, and other security threats. This enables us to maintain the confidentiality, integrity, and availability of our project data throughout its lifecycle.

* 1. **Storage Costs**

In our project, storage costs are a consideration as we manage and maintain datasets in GitHub repositories. GitHub offers free storage for public repositories, which allows us to store our datasets without incurring additional costs. However, for private repositories or larger datasets that exceed GitHub's storage limits, there may be associated costs.

Free Storage: GitHub provides generous storage limits for public repositories, enabling us to store datasets without incurring any costs. This free storage option is suitable for our project, as it allows us to manage our datasets efficiently while minimizing expenses.

Private Repositories: For private repositories, GitHub offers limited storage for free accounts, and additional storage is available through paid plans. If our project requires private repositories for sensitive or proprietary data, we may need to consider the associated costs for additional storage.

Optimization Strategies: To minimize storage costs, we employ optimization strategies such as removing redundant or outdated data, compressing files where possible, and leveraging GitHub's version control features to manage changes efficiently.

# Algorithms & Analysis

## Solution Approach

The purpose of conducting a Risk Assessment is to help identify cyber-attack risks within various industrial sectors. Our final approach involves integrating Risk Assessment into a Dynamic Geographical Dashboard for further analysis within each state.

The framework comprises two major parts: 'Data Preprocessing and Cleaning' and 'Risk Analysis.'

In the Data Preprocessing and Cleaning phase, we utilize NAICS codes as a reference dataset to categorize industries into different sectors, combining them with the Statistics of U.S. Businesses (SUSB Census) Dataset, which contains statistics on U.S. businesses annually, and the REVIS dataset. The REVIS dataset will undergo cleaning and analysis using the SMB Cyber Profiling Method, which will be detailed in subsection 3.2 Solution Approach, to develop risk scores for each business sector.

After preprocessing and cleaning both datasets, we will combine the number of companies in each industry across the U.S. from the SUSB Census Dataset with the risk scores from the REVIS dataset to develop the Risk Assessment. We will enhance the scope of analysis by integrating the Risk Assessment into the Dynamic Geographical Dashboard.

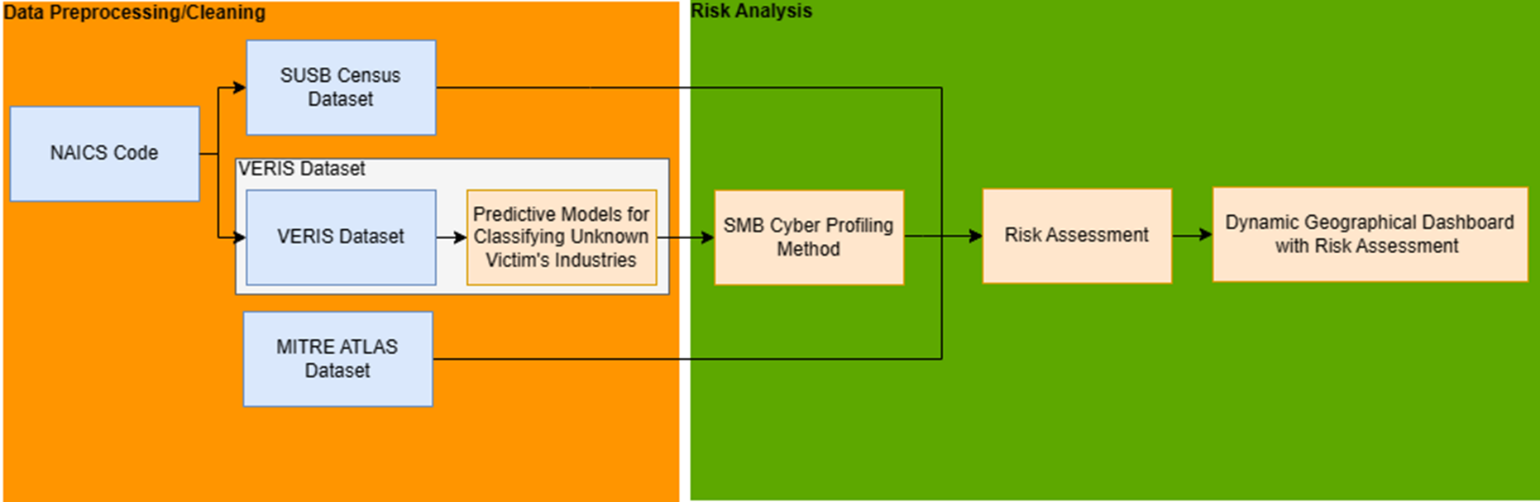


Figure 12: Risk Assessment Framework.

* + 1. Systems Architecture

The progress so far considered various aspects of LLM attacks and the existing cyber security attacks. Since the adoption of LLM tools is rapidly increasing across various industries in the US and since the technology is in fairly nascent stages, the risks related to such tools are necessary to be considered. We developed a workflow plan, thoroughly throughout to shape the progress of our project and provide support and direction guidance to further processes and analytical stages bolstering our project outcomes.

Initially, we began our study by grasping the idea of LLM through articles, research papers, and journals. In the previous sections of this report, we presented our research findings and described every aspect related to or covering LLM, its properties, and applications across various industries across the US. We continued the literature survey to get a comprehensive understanding of LLMs and their increasing adoption in various industrial uses. To understand the influence of LLM we performed web scrapping to retrieve articles from various web sources and gathered insights answering the increasingly diverse sentiments to the applications of LLM-based chatbots such as ChatGPT and Google Gemini.



Figure 13: Solution Architecture.

To gather additional insights, we used the MITRE ATLAS knowledge framework that comprehensively lays out the flow of possible attacks on AI models. It documents real-world reported case studies, and we are building on MITRE’s historical strength in enabling protected or anonymized threat reporting within our community. The ATLAS team is also collaborating with industry and academia on open-source tools like the AI Risk Database, a tool for discovering vulnerabilities associated with public AI models [30].

After gathering sufficient information on the LLM attack types and their operations, we linked the types of attacks from our literature research and the MITRE ATLAS framework. The information presented in the MITRE ATLAS framework presents more than just attack type. It provides information on the tactics used and the motives behind the actions. At this stage of the study, we have a suitable amount of data to proceed with further industry-specific attack types.

Knowing the LLM and AI model attack types isn’t just the research. The word security here depicts and narrates the cyber risks associated with various industrial sectors. To bolster our process in attaining our objective of providing a comprehensive security risk scenario across various industries, we employed Verizon’s public database repository, the VERIS dataset. While there are a handful of efforts to capture security incidents that are publicly disclosed, there is no unrestricted, comprehensive raw dataset available for download on security incidents that is sufficiently rich to support both community research and corporate decision-making.

The VERIS Community Database (VCDB), aims to collect and disseminate data breach information for all publicly disclosed data breaches. Adhering to this dataset, we focused on exploring it and obtaining an overview of the data to transform it into a usable informative resource.

To integrate our research from the early stages, we derived insights and executed appropriate logic in aligning and overlapping similar properties. The VERIS dataset framework contains four distinctive properties that potentially address risks. The properties are laid out as 4A’s, depicting Actors, Actions, Assets, and Attributes. An extended elaboration on these properties is provided in further research. At this stage, we possess all the risk mapping properties from MITRE ATLAS such as attack types and techniques aligning with the 4A properties of the VERIS dataset.

After thorough data preprocessing and data exploratory analysis, we obtained results on frequency counts. Implying that our analysis relied on quantifying the data and providing an objective analysis. Among various features/columns in the dataset, we grouped mainly by industries and aggregated their counts for 4A’s, this gave us a detailed picture of cyberattacks across the US. Being a public data repository, the VERIS data housed several “NA” or NULL values, specifically for character-based columns. These ascertain and incomplete values, we

After gathering the results from the VERIS dataset and mapping the similar attack techniques and types from MITRE ATLAS, we aligned them with US census data called Statistics Of U.S Businesses. This dataset is helpful to analyze the vulnerabilities various businesses across the US possess by adopting LLM or AI models in their daily operations. This data integration provided us with a higher range of number of businesses from several states across the US.

We calculated the risks and normalized them across all sectors or industries considering the census data as well. A comprehensive risk framework was built and then with the aggregated scores, we visualized data on the US map.

* + 1. Systems Security

Ensuring robust systems security is essential for protecting our risk assessment and profiling framework from cyber threats. Our approach to systems security integrates multiple layers of protection to safeguard sensitive data and mitigate potential risks effectively.

Data security begins with the integration of datasets like the VERIS dataset and Statistics of US Businesses (SUSB) into our framework. We prioritize data encryption to maintain confidentiality and integrity, employing industry-standard encryption algorithms for both data at rest and in transit. Access controls are implemented to restrict unauthorized access, with role-based permissions and authentication mechanisms ensuring that only authorized users can access and manipulate the data.

By implementing these security measures, we aim to maintain the confidentiality, integrity, and availability of our risk assessment and profiling system, protecting it against potential cyber threats and ensuring the trustworthiness of our analyses and recommendations.

### Systems Data Flows

Our system's data flows are designed to facilitate the seamless collection, processing, and analysis of data to support risk assessment and profiling activities.

At the outset, data is collected from multiple sources to ensure comprehensive coverage of cyber incidents and relevant industry information. This includes the VERIS dataset, which provides detailed incident records, as well as supplementary datasets such as the Statistics of US Businesses (SUSB). We also utilize web scraping techniques to gather additional insights from news articles and other publicly available sources.

Once the data is collected, it undergoes preprocessing to clean and standardize it for analysis. This involves tasks such as removing duplicates, handling missing values, and normalizing data formats. For example, in the VERIS dataset, incident records are cleaned and standardized to ensure consistency in fields such as actors, actions, and assets.

The preprocessed data is then analyzed using various techniques to identify patterns, trends, and correlations related to cyber incidents and attack techniques. We employ statistical methods such as the Lorenz curve and Gini coefficient to assess the distribution of risks among industry sectors and attack vectors. Additionally, we map VERIS actions to MITRE ATT&CK techniques to enrich our analysis with insights into adversary behaviors and tactics.

To facilitate understanding and decision-making, the analyzed data is visualized through dynamic dashboards and interactive reports. These visualizations provide stakeholders with intuitive insights into risk profiles, incident frequencies, and geographical trends. For example, dynamic geographical dashboards enable users to identify high-risk areas and prioritize cybersecurity measures accordingly.

Finally, our data flows are designed to support continuous improvement and refinement of our risk assessment and profiling framework. Regular feedback loops and updates ensure that our analyses remain relevant and accurate in the face of evolving cyber threats and industry trends.

* + 1. Algorithms & Analysis
       1. SMB Cyber profile Method

SMB stands for Small and Medium Businesses; this term is generally used to indicate that a particular business, firm, or organization consists of up to 100 or about 1000 employees. SMB profiling is used to analyze the risks faced by various businesses within various sectors. These sectors are classified in the NAICS codes, which help us analyze the sectors by either generalizing them or considering a specific subsector.

The resources provided by the partner focus on how the use of data can help in profiling businesses to analyze the risk based on the historical records of cyber-attacks on various organizations and businesses across different industries.

Considering the approach of this risk analysis method, we proceeded with using the VERIS dataset. As stated in the early sections of this report, VERIS is an open dataset where companies and organizations across the world voluntarily provide information about cyber-attacks faced by their company or organization. Although the VERIS dataset contains about 10,000 records, not all of them provide substantial information on cyber-attacks. This is mainly due to reasons such as incomplete information, unknowing variables, a lack of knowledge of cyber threats, and, most importantly, no information on adversaries. These limiting information entities make it difficult to analyze the data and deduce methods for analyzing risks. However, not all information in the VERIS dataset is redundant. There are various records and incidents that provide valuable information.

A number profile method with text

Description automatically generated

Figure 14: SMB profiling method.

The above is a hypothetical profiling method to provide a reference for our attempt at SMB profiling.

The columns contain NAICS codes to classify various industries and sectors. Following the NAICS codes, we have the names of the industries that the corresponding information belongs to. In the profiling, it can be seen the three sectors with distinctive colors: red, blue, and yellow. Each column within these sectors represents a different entity belonging to the adversaries, the action taken, and the assets affected. It is known that these are just another way to represent the profiling method, and we will be using various fields within the VERIS dataset to assess such a risk profiling method.

Actors, actions, assets, and attributes are the major fields that will be used during the assessment of profiling. The columns within the VERIS dataset provide information for such questions.

The actors represent the adversaries that cause a cyberattack on an organization. These adversaries could be anyone with knowledge of a cyberattack and the intention of causing harm. In the VERIS dataset, we have various entities within "actors" that showcase if the attacker was an individual, nation-state, organization, and so on. These could be classified further to understand the depth of the attack. This group enables us to form a profiling method that is nearly equal to that of SMB's fields within the red section.

The assets field here is closely related to the fields in the blue section of the SMB profile. These provide insights for questions such as "What happened with the assets?" or "How the assets were exploited." These questions are used in the framework to assess the vulnerability of the sector. There are various assets within the assets that shed light on how the assets were affected and how they were used.

Following these two fields, another potential field is attributes. It is to be noted that the attributes and the attribute fields within the VERIS dataset are closely related and that they may overlap in certain properties and attributes. Essentially, the attributes are analogous to the fields in the yellow section of the SMB profile. These metrics depict the consequences. This includes loss of data, deletion of data, and answering the question of how the assets were affected.

The values within the SMB profile indicate the severity or level of effect the different fields or attributes of the VERIS datasets have caused for a particular sector. These values are based on the frequency counts explored for each column with respect to each industry and then categorized yearly. This helps in recognizing the attack for each sector and comparing the severity, consequences of the attack, and vulnerability.

The method allows us to use the VERIS dataset to gather the related information and provide insights that help us map the risks for various industries.

* + - 1. Risk Framework Analysis

In this stage, we experienced using the VERIS dataset to frame the risk patterns and shape them into the form of the SMB method. It is to be noted that the risk profiling method will not be replicated as provided by the partner. We will be discovering various patterns and sub-attributes within the fields of the VERIS dataset.

The provided image depicts the motives behind the attack caused by various adversaries. The logical reason supporting this graph is that it allows us to answer the question and provides insights on "why would someone attack a business?" The values depicted here provide an insight into why the financial, education, and public sectors remain the highest sectors with the greatest number of attacks. The motives behind such attacks are either unknown or not available. This is since, as previously stated, the VERIS dataset is not a complete dataset with accurate values, nor does it consist of consistent data.

A screenshot of a spreadsheet

Description automatically generated

Figure 15: Mapping Motives of Actors for Various Industries.

A close-up of a graph

Description automatically generated

Table 9: Attack on Industries by Different Actor Groups for 2023.

The above image provides information on the adversaries that were responsible for attacks on various industries. One interesting observation that can be made is the addition of the timeline column. In this part of the analysis, we withheld the logic of understanding trends. It is noticeable from the tabular graph that the most attacks in the year 2023 were targeted at the education sector and targeted by organized crime groups. It may be deduced that cybersecurity isn't necessarily the strength of this sector. It happens to be one of the most vulnerable sectors among the rest of the sectors.

A table of data with numbers and text

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated

Table 10: The number of attack incident.

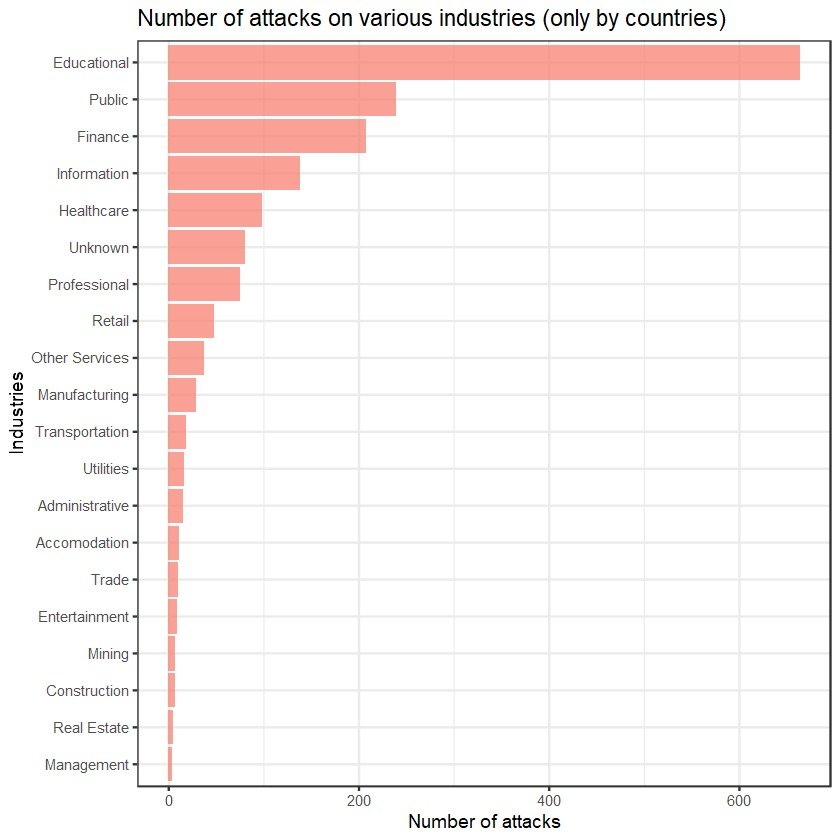


Figure 16: Number of attacks on various industries.

The above-represented data subsets show the number of attack incidents that happened in each year, grouped along with the industries attacked and the adversaries attacking these industries. It is to be noted that the count here doesn't consider the adversaries that are unknown or the information that isn't available. It is evident that the education sector is greatly targeted by the organized crime group, following the public industry that is attacked by activists. The VERIS dataset also provides the adversaries belonging to a country. Here we can witness that most of the attacks have been committed by Russia, followed by the US and China. The bar graph here shows the frequency of attacks caused by countries when grouped together.

In the next session of SMB profiling method, the analysis aimed to pinpoint cybersecurity breaches within three critical dimensions: confidentiality, integrity, and availability, which are pivotal for grasping the severity of cybersecurity threats. In cybersecurity, integrity ensures data remains accurate, consistent, and trustworthy throughout its lifecycle, such as in financial institutions safeguarding transaction records. Confidentiality shields sensitive information from unauthorized access or disclosure, like healthcare organizations protecting patient medical records to preserve privacy. Availability guarantees data and services are accessible when needed, as seen in e-commerce platforms using measures like load balancing and DDoS protection for continuous service during high traffic or cyberattacks. These cybersecurity tenets—integrity, confidentiality, and availability—are critical for securing data and systems across industries.

The objective of this section was to uncover instances where data confidentiality, integrity, or availability was compromised, disrupted, or at risk. This stratification enabled a more comprehensive assessment of the cybersecurity environment, focusing on specific cyber threats like hacking, malware, social engineering, misuse, and unknown factors. By categorizing breaches in this manner, the goal was to prioritize remedial actions and fortify cybersecurity defenses based on the seriousness and prevalence of different threat vectors.

The table presented here displays the frequency of breaches categorized by consequence types across various industries. Subsequently, we have created a corresponding plot that visually represents this data for enhanced comprehension and analysis.



Table 11: The number of integrities, confidentiality, and availability for each industry.

The stacked bar plot below visualizes breach frequencies across victim industries, highlighting the occurrences of breaches in confidentiality, integrity, and availability. Each bar represents a victim industry, with different colors indicating the proportions of breaches in each category. The plot offers a clear comparison of breach counts within and across industries, allowing for insights into which sectors are most affected by cybersecurity incidents. The legend aids in understanding the color-coded categories, making it easier to interpret and analyze the data. This visualization facilitates decision-making by identifying trends, patterns, and areas that require more robust cybersecurity measures or remediation actions.

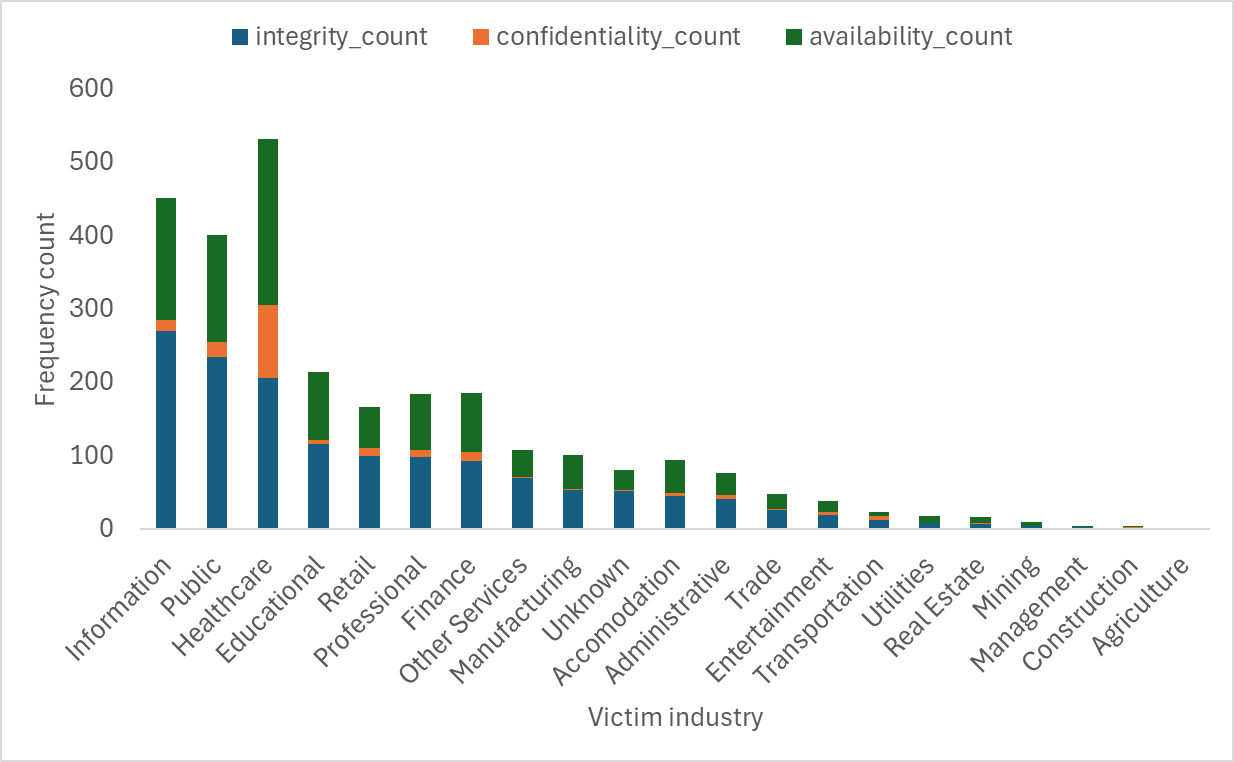


Figure 17: Frequency of different types of consequences for each industry.

* + - 1. Mapping VERIS Dataset with MITRE ATLAS

The VERIS framework provides detailed insights into various actions and behaviors observed during security incidents, allowing us to understand the tactics, techniques, and procedures (TTPs) employed by adversaries. By categorizing incidents into actionable elements such as hacking, malware, social engineering, misuse, and unknown activities, VERIS enables organizations to identify patterns and trends in cyber threats.

To enrich our understanding of cyber threats and adversary behaviors, we mapped the VERIS action dataset to MITRE ATT&CK attack techniques. By aligning VERIS actions with MITRE ATT&CK techniques, we were able to identify commonalities and correlations between observed incidents and known adversary behaviors, thereby enhancing our ability to detect, respond to, and mitigate cyber threats effectively.

Through this mapping process, we gained valuable insights into the tactics and techniques employed by threat actors across different stages of the attack lifecycle. By associating VERIS actions such as elevation of privilege, exfiltration, infiltration, and misuse with specific MITRE ATT&CK techniques, we can better understand the modus operandi of adversaries and develop proactive defenses to protect against known attack vectors. This alignment not only enhances our incident response capabilities but also facilitates threat intelligence sharing and collaboration within the cybersecurity community, ultimately strengthening our collective resilience against evolving cyber threats.

|  |  |  |  |
| --- | --- | --- | --- |
| **MITRE ATLAS ID** | **LLM Attack Technique** | **Description** | **MITRE ATLAS Tactic Scope** |
| AML.T0051 | LLM Prompt Injection | Malicious Prompt Engineering aims to cause the LLM Model to disregard its original instructions and cede control to the adversaries' instructions (prompts). | Initial Access  Persistence  Defense Evasion  Privilege Escalation |
| AML.T0056 | LLM Meta Prompt Extraction | Exfiltration of the LLMs Meta prompt via malicious prompt engineering or by other means. | Discovery  Exfiltration |
| AML.T0057 | LLM Data Leakage | Malicious prompt Engineering to induce the leakage of the Training Data. | Exfiltration |
| AML.T0052.000 | Spearphishing via Social Engineering an LLM | To deploy a socially engineered LLM feigning legitimacy to spearphish unsuspecting users. Or, to create deceptively trustworthy content for phishing emails. | Initial Access |
| AML.T0053 | LLM Plugin Compromise | To undermine the LLM's plugins that connect with other services, applications, or databases, thereby jeopardizing these resources. | Execution  Privilege Escalation |
| AML.T0054 | LLM Jailbreak | Malicious prompt engineering aims to cause the LLM to disregard the guardrails placed upon it, thereby generating illicit, hateful, biased content, and spreading misinformation. | Defense Evasion  Privilege Escalation |
| AML.T0034 | Cost Harvesting | A Disruption of Service(DoS) technique, malicious prompt strategized to exploit the inefficiencies in neural network of the LLM to cause maximize energy consumption and latency. | Impact |
| AML.T0024.002 | Extract LLM Model | Repeatedly querying the LLM to gather inferences on a dataset, then utilizing these inferences (dataset labels) to create a separate but functionally identical copy of the target LLM model, aiming to mimic its behavior and performance precisely. | Exfiltration |

Table 12: Identified LLM attack techniques from MITRE ATLAS framework.

The VERIS Cyberattack techniques classified under the seven categories as below:

Malware: A malicious software, script, or code deployed by the adversaries on the victim’s assets and alter its state of function.

Hacking: External adversaries' deliberate attempts to gain unauthorized access to a victim's assets and compromise their security attributes by circumventing the defense and countermeasure efforts of the victim organization.

Social: Social tactics employed by adversaries to exploit the human element in the users of an organization’s information assets via deception, intimidation, manipulation and other methods.

Misuse: The deliberate misuse of organizational privileges by an insider to compromise organizational assets, conducted with malicious intent.

Physical: External adversaries’ deliberate attempts to using physical proximity to victim’s assets to undermine their security attributes such as assault, sabotaging, snooping, tampering and other methods

Errors: Actions done by individuals with non-malicious intent that accidentally led to the compromise of security attributes of an organizational assets’ such as misconfigurations, programming errors, malfunctions and others.

Environmental: Natural events and hazards, like power failures, pipe leaks, earthquakes, floods, and others, not resulting from malicious human actions

Since, the MITRE ATLAS Framework excludes actions with non-malicious intent (errors) or those occurring due to physical proximity to information assets or natural events, we only consider the first four categories (*Malware, Hacking, Social, and Misuse*) when mapping MITRE ATLAS LLM Techniques to VERIS Cyberattack techniques.

To bridge the gap between two vastly different threat landscapes—specifically, the Artificial Intelligence Threat Landscape of MITRE ATLAS and the traditional Cyberattack threat landscape of VERIS—we aim to establish a high-level conceptual mapping of the attack techniques. This entails drawing on conceptual analogies between techniques from both landscapes and ensuring that at least the VERIS Cyberattack techniques share the same adversarial tactic scope as the selected LLM Attack techniques from MITRE ATLAS or shares similarities in the way in which the attacks are executed.

|  |  |
| --- | --- |
| **ATLAS\_ID** | **VERIS\_action \_variety** |
| AML.T0051 | action.malware.variety.Adminware |
| AML.T0051 | action.malware.variety.Backdoor |
| AML.T0051 | action.malware.variety.Backdoor or C2 |
| AML.T0051 | action.malware.variety.C2 |
| AML.T0051 | action.malware.variety.In-memory |
| AML.T0051 | action.malware.variety.Modify data |
| AML.T0051 | action.malware.variety.Disable controls |
| AML.T0051 | action.malware.variety.Exploit misconfig |
| AML.T0051 | action.malware.variety.Evade Defenses |
| AML.T0051 | action.malware.variety.Exploit vuln |
| AML.T0051 | action.malware.variety.RAT |
| AML.T0051 | action.malware.variety.Rootkit |
| AML.T0051 | action.malware.variety.Trojan |
| AML.T0051 | action.hacking.variety.Abuse of functionality |
| AML.T0051 | action.hacking.variety.Backdoor |
| AML.T0051 | action.hacking.variety.CSRF |
| AML.T0051 | action.hacking.variety.Disable controls |
| AML.T0051 | action.hacking.variety.Evade Defenses |
| AML.T0051 | action.hacking.variety.Exploit misconfig |
| AML.T0051 | action.hacking.variety.Exploit vuln |
| AML.T0051 | action.hacking.variety.Fuzz testing |
| AML.T0051 | action.hacking.variety.Hijack |
| AML.T0051 | action.hacking.variety.Insecure deserialization |
| AML.T0051 | action.hacking.variety.Reverse engineering |
| AML.T0051 | action.hacking.variety.User breakout |
| AML.T0051 | action.hacking.variety.XML injection |
| AML.T0051 | action.hacking.variety.XPath injection |
| AML.T0051 | action.hacking.variety.XQuery injection |
| AML.T0051 | action.misuse.variety.Data mishandling |
| AML.T0051 | action.misuse.variety.Evade Defenses |
| AML.T0051 | action.misuse.variety.Knowledge abuse |
| AML.T0051 | action.misuse.variety.Possession abuse |
| AML.T0051 | action.misuse.variety.Privilege abuse |
| AML.T0056 | action.malware.variety.Capture app data |
| AML.T0056 | action.malware.variety.Capture stored data |
| AML.T0056 | action.malware.variety.Exploit misconfig |
| AML.T0056 | action.malware.variety.Exploit vuln |
| AML.T0056 | action.malware.variety.Export data |
| AML.T0056 | action.malware.variety.Profile host |
| AML.T0056 | action.malware.variety.RAM scraper |
| AML.T0056 | action.malware.variety.RAT |
| AML.T0056 | action.malware.variety.Spyware/Keylogger |
| AML.T0056 | action.malware.variety.Trojan |
| AML.T0056 | action.hacking.variety.Abuse of functionality |
| AML.T0056 | action.hacking.variety.Exploit misconfig |
| AML.T0056 | action.hacking.variety.Exploit vuln |
| AML.T0056 | action.hacking.variety.Profile host |
| AML.T0056 | action.hacking.variety.Reverse engineering |
| AML.T0056 | action.hacking.variety.XML external entities |
| AML.T0056 | action.hacking.variety.XML injection |
| AML.T0056 | action.hacking.variety.XPath injection |
| AML.T0056 | action.hacking.variety.XQuery injection |
| AML.T0056 | action.misuse.variety.Data mishandling |
| AML.T0056 | action.misuse.variety.Knowledge abuse |
| AML.T0057 | action.malware.variety.Backdoor or C2 |
| AML.T0057 | action.malware.variety.Capture app data |
| AML.T0057 | action.malware.variety.Capture stored data |
| AML.T0057 | action.malware.variety.MitM |
| AML.T0057 | action.malware.variety.Exploit misconfig |
| AML.T0057 | action.malware.variety.Exploit vuln |
| AML.T0057 | action.malware.variety.Export data |
| AML.T0057 | action.malware.variety.Packet sniffer |
| AML.T0057 | action.malware.variety.Password dumper |
| AML.T0057 | action.malware.variety.RAM scraper |
| AML.T0057 | action.malware.variety.RAT |
| AML.T0057 | action.malware.variety.Spyware/Keylogger |
| AML.T0057 | action.malware.variety.Trojan |
| AML.T0057 | action.hacking.variety.Abuse of functionality |
| AML.T0057 | action.hacking.variety.Exploit misconfig |
| AML.T0057 | action.hacking.variety.Exploit vuln |
| AML.T0057 | action.hacking.variety.Fuzz testing |
| AML.T0057 | action.hacking.variety.MitM |
| AML.T0057 | action.hacking.variety.Reverse engineering |
| AML.T0057 | action.hacking.variety.XML external entities |
| AML.T0057 | action.hacking.variety.XML injection |
| AML.T0057 | action.hacking.variety.XPath injection |
| AML.T0057 | action.hacking.variety.XQuery injection |
| AML.T0057 | action.misuse.variety.Data mishandling |
| AML.T0057 | action.misuse.variety.Possession abuse |
| AML.T0052.000 | action.malware.variety.Client-side attack |
| AML.T0052.000 | action.malware.variety.Spyware/Keylogger |
| AML.T0052.000 | action.malware.variety.Trojan |
| AML.T0052.000 | action.social.variety.Baiting |
| AML.T0052.000 | action.social.variety.Elicitation |
| AML.T0052.000 | action.social.variety.Influence |
| AML.T0052.000 | action.social.variety.Phishing |
| AML.T0052.000 | action.social.variety.Pretexting |
| AML.T0052.000 | action.social.variety.Prompt bombing |
| AML.T0052.000 | action.misuse.variety.Email misuse |
| AML.T0052.000 | action.misuse.variety.Knowledge abuse |
| AML.T0053 | action.malware.variety.Adminware |
| AML.T0053 | action.malware.variety.Backdoor or C2 |
| AML.T0053 | action.malware.variety.C2 |
| AML.T0053 | action.malware.variety.Capture app data |
| AML.T0053 | action.malware.variety.Capture stored data |
| AML.T0053 | action.malware.variety.Client-side attack |
| AML.T0053 | action.malware.variety.Destroy data |
| AML.T0053 | action.malware.variety.In-memory |
| AML.T0053 | action.malware.variety.MitM |
| AML.T0053 | action.malware.variety.Modify data |
| AML.T0053 | action.malware.variety.Disable controls |
| AML.T0053 | action.malware.variety.Downloader |
| AML.T0053 | action.malware.variety.Exploit misconfig |
| AML.T0053 | action.malware.variety.Evade Defenses |
| AML.T0053 | action.malware.variety.Exploit vuln |
| AML.T0053 | action.malware.variety.Export data |
| AML.T0053 | action.malware.variety.Profile host |
| AML.T0053 | action.malware.variety.RAT |
| AML.T0053 | action.malware.variety.Rootkit |
| AML.T0053 | action.malware.variety.Scan network |
| AML.T0053 | action.malware.variety.Spyware/Keylogger |
| AML.T0053 | action.malware.variety.Trojan |
| AML.T0053 | action.hacking.variety.Abuse of functionality |
| AML.T0053 | action.hacking.variety.CSRF |
| AML.T0053 | action.hacking.variety.Disable controls |
| AML.T0053 | action.hacking.variety.Evade Defenses |
| AML.T0053 | action.hacking.variety.Exploit misconfig |
| AML.T0053 | action.hacking.variety.Exploit vuln |
| AML.T0053 | action.hacking.variety.Format string attack |
| AML.T0053 | action.hacking.variety.Hijack |
| AML.T0053 | action.hacking.variety.Insecure deserialization |
| AML.T0053 | action.hacking.variety.Integer overflows |
| AML.T0053 | action.hacking.variety.OS commanding |
| AML.T0053 | action.hacking.variety.Path traversal |
| AML.T0053 | action.hacking.variety.Profile host |
| AML.T0053 | action.hacking.variety.Reverse engineering |
| AML.T0053 | action.hacking.variety.RFI |
| AML.T0053 | action.hacking.variety.Scan network |
| AML.T0053 | action.hacking.variety.Special element injection |
| AML.T0053 | action.hacking.variety.SQLi |
| AML.T0053 | action.hacking.variety.User breakout |
| AML.T0053 | action.hacking.variety.XML external entities |
| AML.T0053 | action.hacking.variety.XSS |
| AML.T0053 | action.misuse.variety.Data mishandling |
| AML.T0053 | action.misuse.variety.Evade Defenses |
| AML.T0053 | action.misuse.variety.Knowledge abuse |
| AML.T0053 | action.misuse.variety.Privilege abuse |
| AML.T0053 | action.misuse.variety.Unapproved software |
| AML.T0053 | action.misuse.variety.Unapproved workaround |
| AML.T0054 | action.malware.variety.Adminware |
| AML.T0054 | action.malware.variety.Backdoor |
| AML.T0054 | action.malware.variety.Backdoor or C2 |
| AML.T0054 | action.malware.variety.In-memory |
| AML.T0054 | action.malware.variety.Disable controls |
| AML.T0054 | action.malware.variety.Exploit misconfig |
| AML.T0054 | action.malware.variety.Evade Defenses |
| AML.T0054 | action.malware.variety.Exploit vuln |
| AML.T0054 | action.malware.variety.Rootkit |
| AML.T0054 | action.malware.variety.Trojan |
| AML.T0054 | action.hacking.variety.Abuse of functionality |
| AML.T0054 | action.hacking.variety.Backdoor |
| AML.T0054 | action.hacking.variety.Disable controls |
| AML.T0054 | action.hacking.variety.Evade Defenses |
| AML.T0054 | action.hacking.variety.Exploit misconfig |
| AML.T0054 | action.hacking.variety.Exploit vuln |
| AML.T0054 | action.hacking.variety.OS commanding |
| AML.T0054 | action.hacking.variety.Reverse engineering |
| AML.T0054 | action.hacking.variety.User breakout |
| AML.T0054 | action.hacking.variety.Virtual machine escape |
| AML.T0054 | action.social.variety.Evade Defenses |
| AML.T0054 | action.misuse.variety.Evade Defenses |
| AML.T0054 | action.misuse.variety.Knowledge abuse |
| AML.T0054 | action.misuse.variety.Privilege abuse |
| AML.T0034 | action.malware.variety.Click fraud |
| AML.T0034 | action.malware.variety.Click fraud and cryptocurrency mining |
| AML.T0034 | action.malware.variety.Cryptocurrency mining |
| AML.T0034 | action.malware.variety.Disable controls |
| AML.T0034 | action.malware.variety.DoS |
| AML.T0034 | action.malware.variety.Exploit misconfig |
| AML.T0034 | action.malware.variety.Exploit vuln |
| AML.T0034 | action.malware.variety.Spam |
| AML.T0034 | action.hacking.variety.Abuse of functionality |
| AML.T0034 | action.hacking.variety.Disable controls |
| AML.T0034 | action.hacking.variety.DoS |
| AML.T0034 | action.hacking.variety.Exploit misconfig |
| AML.T0034 | action.hacking.variety.Exploit vuln |
| AML.T0034 | action.hacking.variety.Reverse engineering |
| AML.T0034 | action.misuse.variety.Knowledge abuse |
| AML.T0034 | action.misuse.variety.Unapproved hardware |
| AML.T0034 | action.misuse.variety.Unapproved software |
| AML.T0034 | action.misuse.variety.Unapproved workaround |
| AML.T0024.002 | action.malware.variety.Capture app data |
| AML.T0024.002 | action.malware.variety.Capture stored data |
| AML.T0024.002 | action.malware.variety.Exploit misconfig |
| AML.T0024.002 | action.malware.variety.Exploit vuln |
| AML.T0024.002 | action.malware.variety.Export data |
| AML.T0024.002 | action.malware.variety.RAM scraper |
| AML.T0024.002 | action.malware.variety.RAT |
| AML.T0024.002 | action.malware.variety.Spyware/Keylogger |
| AML.T0024.002 | action.malware.variety.Trojan |
| AML.T0024.002 | action.hacking.variety.Abuse of functionality |
| AML.T0024.002 | action.hacking.variety.Exploit misconfig |
| AML.T0024.002 | action.hacking.variety.Exploit vuln |
| AML.T0024.002 | action.hacking.variety.Reverse engineering |
| AML.T0024.002 | action.hacking.variety.XML external entities |
| AML.T0024.002 | action.hacking.variety.XML injection |
| AML.T0024.002 | action.hacking.variety.XPath injection |
| AML.T0024.002 | action.hacking.variety.XQuery injection |
| AML.T0024.002 | action.misuse.variety.Data mishandling |
| AML.T0024.002 | action.misuse.variety.Knowledge abuse |

Table 13: Results of mapping MITRE ATLAS LLM techniques to VERIS Cyberattack techniques.

* + - 1. Application of Lorenze Curves in Risk Profiling
         1. Lorenze Curve and Gini Coefficient

The Lorenz curve is graphical illustration of statistical inequality (or, dispersion), first introduced by Max Lorenz to visualize wealth distribution among a nation’s population of individuals [31]. Subsequently, Corrado Gini built on this idea to develop a metric which quantifies the statistical inequality called the Gini Index or Gini Coefficient [31].

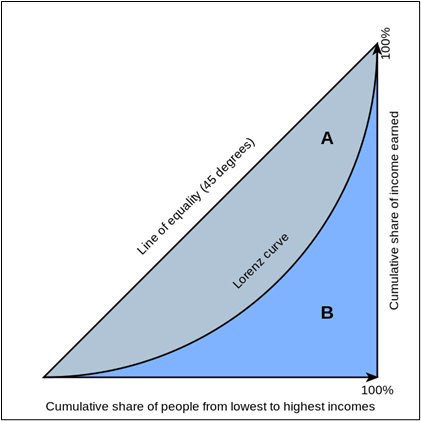


Figure 18: Lorenze cure description.

In our project, the Lorenz Curve has been instrumental in visualizing the inequality in the distribution of cybersecurity incidents across various industries. This graphical tool maps the cumulative percentage of cybersecurity incidents recorded against the cumulative percentage of firms, highlighting disparities across different sectors. For example, the Lorenz curve for DDoS attack frequencies within the banking sector reveals significant inequality, suggesting that a small subset of firms experiences a disproportionately high number of DDoS attacks. This insight is critical for cybersecurity professionals aiming to allocate resources more efficiently and strengthen defenses where they are most needed.

Lorenze curve & line of quality regarding a nation’s wealth distribution across its populations. The Lorenz curve is plotted by mapping the cumulative percentages of a country's population (ordered from poorest to wealthiest) on the x-axis against their cumulative income or wealth on the y-axis. In a perfectly egalitarian country, where everyone has an equal amount of wealth, the data points form a 45-degree straight line termed as the 'Line of Equality'.

The Gini Coefficient is the quantification of the amount of deviation of the Lorenz curve from the Line of Equality, calculated as the ratio of the area between the Lorenz Curve and the Line of Equality to the total area under the Line of Equality (i.e. with the x-axis). According to a figure related to Lorenza curve, the Gini Coefficient equals (A)/(A+B).

To demonstrate the application of Lorenz curve in risk profiling, consider the example of incident frequencies across industry sectors distributed by attack techniques (see Table below):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Industry** | **Phishing** | **Malware** | **DDoS** | **Insider Threat** | **Ransomware** |
| Banking | 100 | 150 | 50 | 30 | 70 |
| Healthcare | 80 | 120 | 40 | 60 | 100 |
| Retail | 90 | 110 | 70 | 20 | 60 |

Table 14: An Example of Incident Frequencies by Attack Techniques Across Industry Sectors.

The example below shows the Lorenz Curve application for Profiling Risk among Attack Techniques within Industries, by taking, Banking Sector as an Example.

**Banking Sector**

|  |  |  |
| --- | --- | --- |
| **Attack Technique** | **Frequency** | **Cumulative %** |
| Insider Threat | 30 | 9.1% |
| DDoS | 50 | 24.2% |
| Phishing | 100 | 54.5% |
| Malware | 150 | 100% |

Table 15: The Cumulative Percentages computed from Attack Technique incident frequencies in the Banking Sector.

A graph with a blue line

Description automatically generated

Figure 19: Lorenz Curve for Incidents Frequencies of Attack Techniques within the Banking Sector.

The Gini Index, derived from the Lorenz curve, quantifies the degree of inequality observed in the distribution of cybersecurity incidents among industries. In our analysis, a Gini coefficient of 0.311 for the banking sector indicates a moderate level of inequality in how incidents are distributed among firms within the sector. This metric is crucial for evaluating the effectiveness of current cybersecurity measures and for guiding future investments in security infrastructure. By assessing the Gini coefficients across different sectors, we can prioritize interventions in industries where incident distribution is most unequal, thereby enhancing overall security posture.

The example below shows the plotting of the Lorenz Curve for Profiling Risk among Industry Sectors within Attack Techniques, by taking DDOs attack technique as an Example.

**DDoS**

|  |  |  |
| --- | --- | --- |
| **Industry** | **Frequency** | **Cumulative %** |
| Healthcare | 40 | 25% |
| Banking | 50 | 56.25% |
| Retail | 70 | 100% |

Table 16: The Cumulative Percentages computed from Industry sector incident frequencies considering the DDos mode of Attack Technique.

*A graph with a line

Description automatically generated*

Figure 20: Lorenz Curve for Incidents Frequencies among Industry Sectors considering the DDos mode of Attack technique.

The computed Gini Coefficient via the for Incidents Frequencies among Industry Sectors considering the DDos mode of Attack technique, 0.125.

Using this Gini Coefficient for the DDos Attack technique, we may further develop a custom scoring methodology for scoring the individual Industry Sectors based on their contribution to this Gini coefficient for this Attack technique.

* + - * 1. Risk Calculation Based on the Frequency

A screenshot of a graph

Description automatically generatedRisk calculations are an important aspect of securing the cyber systems of a company, firm, business, organization, or any such establishment. Cyber risks have been prominent since the inception of computers and the adoption of the internet. It’s retrospective of thinking an attack would be more damaging were physical. This assumption may still be valid in today’s world; however, we have experienced a growing number of cyber-attacks on various industrial sectors.

Risk analysis requires a framework which is based on assessing diverse aspects and entities of cyber-attacks such as adversaries, assets, attributes, extent of damage in economy, timeline to discover the attack and many more. A crucial point to be considered here is that the risk scores are usually based on the subject matter expertise and may reflect subjective outtake based on a person’s or a team’s knowledge and experience in handling a cyber-attack and assessing the extent of damages caused by these attacks.

In this study, we have adopted an objective approach in creating risk scores and framework to gather information based on evidence from the VERIS dataset. The evidence here are the incidents recorded in the database proving details on various aspects of the cyber-attack. To create a repeatable process, we have used frequency counts and calculated probability across all industries. Since the study focuses on assessing the risks related to the use of large language models in the industries for diverse applications, we will be replacing the VERIS attack techniques with the LLM attack techniques from the MITRE ATLAS dataset. This mapping is executed bi-directionally, making sure that the LLM attack techniques match with the VERIS attack techniques.

The concern with not assessing the LLM attacks directly is that there isn’t a database or a public repository that consists of recorded evidence suggesting compromised LLM. Therefore, with the VERIS as our foundation of analysis, we will be building the LLM risk profile.

The analysis was carried out in different stages. Each stage of forming a risk framework and formulating risk scores dove deeper into consideration of different techniques with increased level of sophistication and decreasing abstractive approach.

Risk calculation based on frequency:

Normalized frequency of each Assets= ‘asset\_norm’

Normalized frequency of each Action= ‘action\_norm’

Assumption: Attributes essentially occur after the attack is initiated. Here we will consider only the assets affected and the attack that took place.

Risk score= (sum of asset\_norm freq) x (sum of action\_norm freq)

A graph of a bar chart

Description automatically generated with medium confidence

Figure 21: Risk score based on the frequency of each impact in financial sector

Probability of actionà hacking techniques for incidents through the year 2020-present.

Similarly, we performed data visualization based on the supporting formula of risk analysis using the normalized frequency scores.

A grid with text and numbers

Description automatically generated with medium confidence

Figure 22: Heatmap showing the frequency of incidents per each asset and action type.

The diagram represents an attempt to create visualization using two features: assets and actions. This is the first stage of the formulation of the features. The is a visualization showing the higher risk scores for assets and their corresponding actions. Here, we assumed the formula, as mentioned before:

Risk score= (sum of asset\_norm freq) x (sum of action\_norm freq)

Visualization is the product of the frequency of assets after normalizing and the frequency of actions also after normalizing. Out of several, only a few of the actions and assets are concentrated based on the brighter color. We can say that the cloud and server asset entities are most susceptible to cyber-attacks for the financial industry. The medium of attack seems to be prominent through hacking of web application, infiltration, error in mis delivery, and a few incidents with the physical access have also been recorded.

Initial steps and progress in creating risk framework.

A screenshot of a computer

Description automatically generated

Table 17: The probabilities of the most occurred hacking type for each industry.

* Risk profile showing all range of actions such as hacking, malware, physical, social, misuse, and error.
* The score is based on the probability of the type of attack occurring the most.
* The probability of 1 à number of attack types and the total number of attacks are the same.
* Example: Physical theft in healthcare with the probability of 1 à number of physical attacks = theft attack in healthcare.
* These are the probabilities based on the fact the action for each type was recorded i.e. The value within columns such as action.Hacking, action.Misuse, and so on were “TRUE.”



Table 18: The probabilities of all action type for each industry.

Features to be considered (from VERIS data):

A diagram of a company's attack

Description automatically generated with medium confidence

Figure 23: Risk framework features.

**Formula:**

(Sum of probabilities of all actions) X (Sum of probabilities of all affected assets) X

(Sum of probabilities of attributes) X (Impact coefficient)

In risk cyber security: Cyber risk = Threat x Vulnerability x Information Value

Alternative; Risk = Consequence x Probability x Vulnerability

Logical Reasoning: Various incidents have multiple *attack* techniques, *occurring simultaneously*, and for various purposes.

* We are considering these attack techniques to be MUTUALLY INCLUSIVE.
* Assumption: certain incidents may have single action but considering *WORST CASE SCENARIO* risk analysis, an incident may have multiple incidents.

A screenshot of a computer

Description automatically generated

Table 19: relation between incident ID, actions, assets, and attributes.

The above-shown image shows incidents recorded in the VERIS dataset. The VERIS dataset contains various incidents and comprehensive overlook on the details of the attack. This includes actors, mentioning the adversary responsible for the attacks to be carried out. The actors may essentially be an individual person or group of organizations that have certain agenda to attack an asset of a firm, company, organization, government, or an establishment.

The data subset shown in the image depicts an incident carried out by multiple adversaries on multiple assets, leaving multiple attributes. It is to be noted that this condition considers the fact that in practical scenarios the companies may be liable to such multiple attacks. However, it may be mysterious to trace and understand the attack type relating to a particular asset. Our study has lost sight of tracing these attacks and relating it with the assets, nevertheless, the risk scores consider each of the action, assets, and attributes providing with a composite risk score for each incident using the above-mentioned formula.

The composite scores provide us with an idea on the severity of the attack. The higher the risk score, the higher is the chance of an industry getting attacked. To complete the circle of the calculation, we also introduced the impact scores from the VERIS dataset. This refers to the overall impact column from the dataset that comprises of the values such as insignificant, distracting, painful, damaging, catastrophic, and unknown. According the VERIS these values represent the increasing level of impact of the attacks on the assets for various incidents. For simplicity of risk scores, we assumed the impact levels to be linear scaled. This means, we considered the insignificant to be 0.2, the least on the impact score and catastrophic to the most weighted on the impact score. The liner scale ranges from 0-1 with an increment of 0.2.



Table 20: Converting impact as categorical feature to numerical.

The following is the actual framework that is designed to understand the risk posed by various industries that are victims of cyber-attacks. It is to be noted this is an aggregation of all the incidents across all industries. Here we have represented each column in belonging to different classes. These classes are shown in colors to distinguish them from the rest of the classes. Within each class, you will witness different types of attributes such as hacking, malware, confidentiality, availability, and their variety. To understand this framework, let’s consider an industry to trace the pattern and gather insights.

Here we’re considering the educational industry. The industry is specified in the NAICS with the code number 61. If we look across the framework, we may see different values placed along the framework. In the first class it shows that the threat actor class, considering the educational sector, has higher chances or probability of getting attacked by an adversary that is external factor. This means the external adversary doesn’t belong to the educational organization or establishment and the industry has 98% probability of getting attacked by an external adversary.



Table 21: The probabilities of all types of actions, assets, and attributes for different industries.

In the actions class, there are various techniques of actions or the attacks themselves. On inspecting, we find that if it were a hacking activity, the education sector has the probability of getting hacked though backdoor with the probability of 46%. Similarly, malware-backdoor attack has the probability of 97%, along with misconfiguration and infiltration with the probability score of 50% and 60%. It is essential to consider that these attributes and types are considered by the highest order of the probability score. Expanding on this, we see that the web application is the asset with the highest probability of getting attacked by the adversaries. Now, we focus on the attributes. The attributes depict what happened to the assets and how the attack affects these assets. Looking at the educational sector, the customer personal information was affected the most, perhaps the target. The attacks on this sector may show problems in the software installation, bringing up the integrity issue, along with a higher probability of obstruction, meaning partial or total loss of the availability of assets.

The risk framework also includes a special class to assess the impact of the attack on various assets. The column overall impact as mentioned previously provides result of the attacks. The impact probability serves the highest probability for industries with impact outcomes. Example, the educational industry consists of three different types of impacts: Painful, Damaging, and Catastrophic with the equal probability of 33%. The total overall impact score ca be provided by the sum product of the probability scores and the assigned liner scores from 0-1 for each level of impact.

Finally, by implementing our formula: (Sum of probabilities of all actions) X (Sum of probabilities of all affected assets) X (Sum of probabilities of attributes) X (Impact coefficient)

We get the final risk score for each industry. There isn’t any upper limit to the risk scores based on our model and assumptions. This makes it difficult to compare the risk scores across industries. Therefore, we have normalized the scores, to keep them comparable and even throughout a uniform scale.

It is to be noted that the number of victim industries are less than they obtained from the VERIS data and the NAICS code. This is executed to assure that certain industries and their corresponding incidents have a recorded impact, rather than just plain assumption.



Table 22: Composite risk score for each industrial sector.

After getting composite risk score, perform the final risk score using the formula below by mapping VERIS dataset and SUSB dataset with NAICS code.

Final Risk Score = Composite Risk Score (Normalized form) x Probability of Incident by state x Probability of Incident by Industry x Number of Firms (Normalized form) x 1,000,000

Composite Risk Score = Risk Score calculated from VERIS dataset by industrial sector

Probability of Incident by state = (from VERIS dataset)

Probability of Incident by Industry = (from VERIS dataset)

Number of Firms = Number of Companies by specific Industry in Specific State in normalized form (from SUSB dataset)



Figure 24: Power BI data schema and dataset relationships.

By performing the final risk calculations, we utilized Power BI to connect related datasets with NAICS code as a primary key. The data schema includes dataset below:

* NAICS code table: Reference table for classified industrial sector
* State table: Reference table for state name and abbreviation
* SUSB dataset table: Statistics of U.S. Businesses (SUSB) dataset
* Aggregated VERIS dataset table: Aggregated number of VERIS dataset incidents with mapped LLM incidents by State and Industrial sectors
* Composite Risk Score table: Risk Score for each industrial sector in figure 26
* Final table: Combine all the above tables together as one table. This table contains all related parameters to calculate the final risk score in the next step.

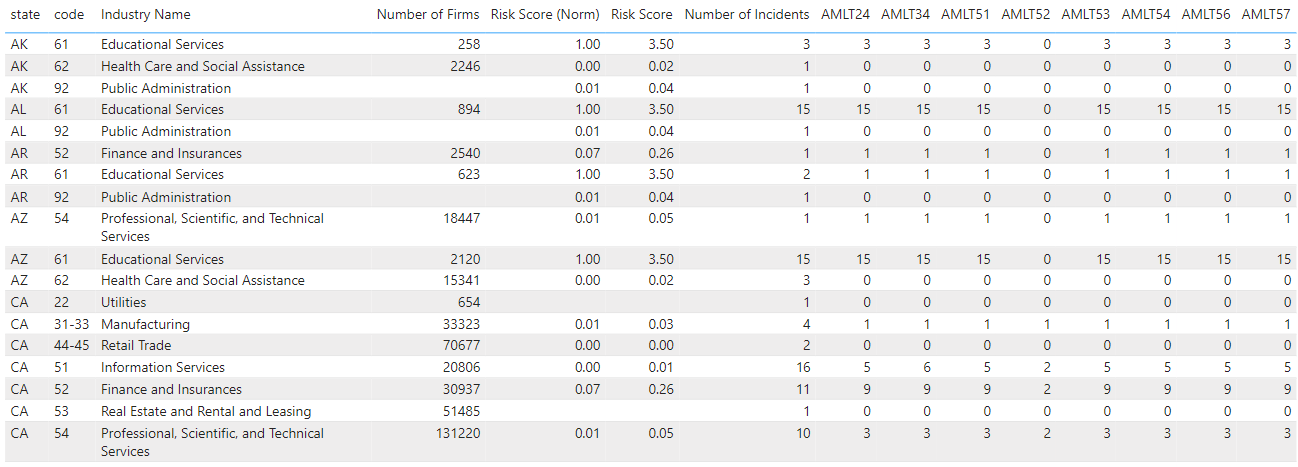


Table 23: Example of final table (combining VERIS dataset with LLMs mapped, NAICS code and SUSB dataset).

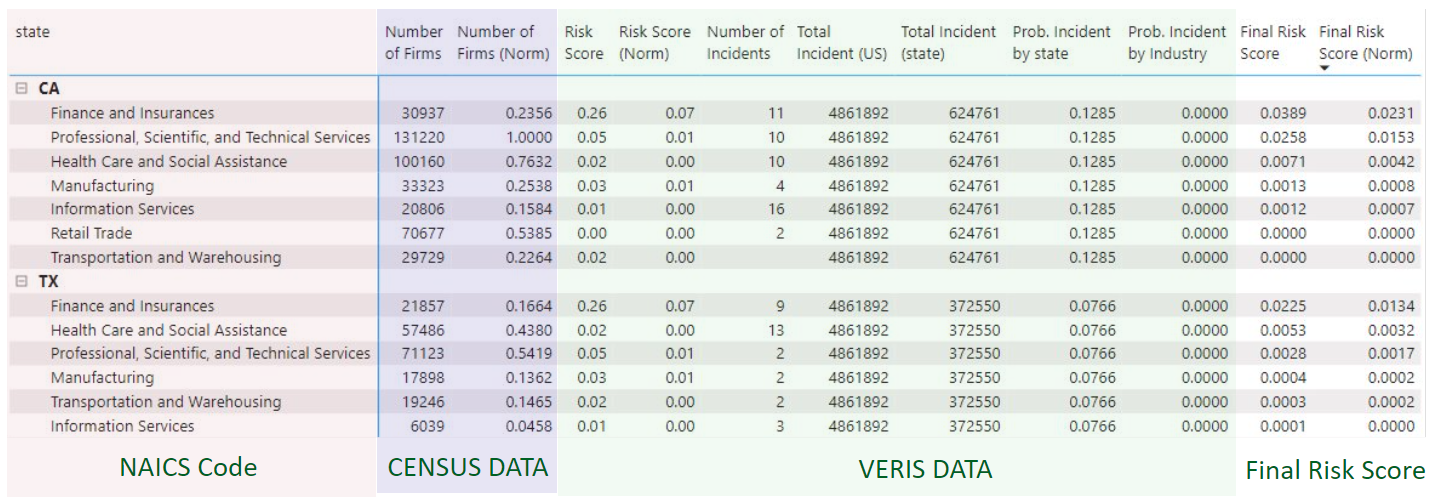


Table 24: Example of final risk calculation.

The final risk score is computed using a primary dataset, comprising NAICS codes, the CENSUS dataset, and the VERIS dataset, with LLM mappings detailed in Table 20. From Table 20, we derive the final risk score, as illustrated in Table 21. Upon obtaining the final risk score, we normalize it to a range of 0 to 1, represented in the 'Final Risk Score (Norm)' column. This normalized score is utilized for visualization on geographical maps by state and industrial sector within Power BI in the next step.

# Visualizations

## Overview

In this section, we present the findings from our frequency analysis of LLM attacks across various industries using the formulated risk analysis framework.

We developed a risk analysis framework focusing on four key classes: actions, assets, attributes, and impacts. These classes were analyzed to understand the nature of LLM attacks and their potential impact on different industries. The framework allowed us to formulate questions based on different classes, such as the types of actions by threat actors, the assets targeted, the attributes affected, and the severity of the impacts.

A diagram of attributes and action

Description automatically generated

Figure 25: Features for assessing risk score.

Using data from the VERIS dataset, which contains records of cybersecurity incidents, we conducted a frequency analysis to understand the prevalence of LLM attacks and their associated actions, assets, attributes, and impacts. We identified various actions by threat actors, including hacking, malware, physical attacks, social engineering, misuse, errors, and unknown actions.

The impact assessment focused on determining the severity of LLM attacks based on the damage caused to assets and attributes. We categorized the impacts into significant, distracting, painful, damaging, and catastrophic, following the VERIS schema. Each impact category was assigned a score ranging from 0 to 1, with catastrophic being the highest.

Based on the formulated risk analysis framework and the frequency analysis results, we calculated risk scores for each industry. The risk scores were normalized to a scale between 0 and 1 for easier comparison. The analysis revealed that the educational industry had one of the highest risk scores, indicating a higher susceptibility to LLM attacks compared to other industries.

## Visualizations

After completing the calculations for both risk models, namely Risk Model 1 – Lorenz Curve Risk Model and Risk Model 2 – Risk Framework, we employed Power BI to analyze and visualize both models. This approach was chosen because each model offers distinct advantages and disadvantages, and using geographical visualization on Power BI allowed us to effectively present and compare the insights derived from each model.

## Risk Model 1 - Lorenz Curve Risk Model

The benefit of using the Lorenz Curve model is its capability to facilitate a direct comparison of relative risk among entities, LLM attack techniques, or different industrial firms with different population sizes. This feature enables a nuanced understanding of risk distribution and highlights disparities in risk exposure across different entities or scenarios, providing valuable insights for risk management and decision-making processes.

Conversely, the Lorenz Curve model primarily concentrates on the frequency of incidents associated with various cyber-attack actions (LLM techniques) relative to the number of firms in the industry, overlooking other crucial aspects of the VERIS dataset such as assets, attributes, and impact factors. This limitation is mitigated by the advantages of Risk Model 2, which encompasses a broader spectrum of factors including assets, attributes, and impact factors. A map with blue circles and red dots

Description automatically generated

Figure 26: Risk Model 1, Geographical Visualization –LLM Prompt Injection in Financial Sector.

Our visualization vividly highlights the widespread and concentrated risks of LLM Prompt Injection incidents across the financial sector, illustrating the acute vulnerability of these institutions to advanced cyber threats. LLM Prompt Injection, where malicious prompts are strategically employed to manipulate or exploit large language models, represents a formidable challenge to the security and integrity of financial data systems. Through this detailed geographic depiction, we can clearly see the distribution of these incidents, particularly concentrated in major financial hubs like California—a state marked in bright red to signify elevated risk levels compared to other regions.

This nuanced portrayal serves to deepen our understanding of the geographic spread and intensity of these threats, illuminating the specific areas within the financial sector that are most at risk. The mapping underscores the heightened susceptibility of institutions in regions like California, where the convergence of vast financial activities and technological integration creates a hotspot for such cyber threats.

Moreover, this visualization acts as a pivotal tool for cybersecurity teams within financial institutions, enabling them to identify risk hotspots and allocate resources more effectively. By bringing these data-driven insights to the forefront, it empowers stakeholders to prioritize and fortify their defenses, developing more targeted and robust strategies against these evolving threats. This approach not only raises awareness among the sector's leadership but also catalyzes a proactive stance on cybersecurity, ensuring that protective measures evolve in tandem with the sophistication of the threats faced. Through strategic visualization and analysis, we're better equipped to safeguard the foundational integrity of our financial systems against the perils of LLM Prompt Injection attacks.

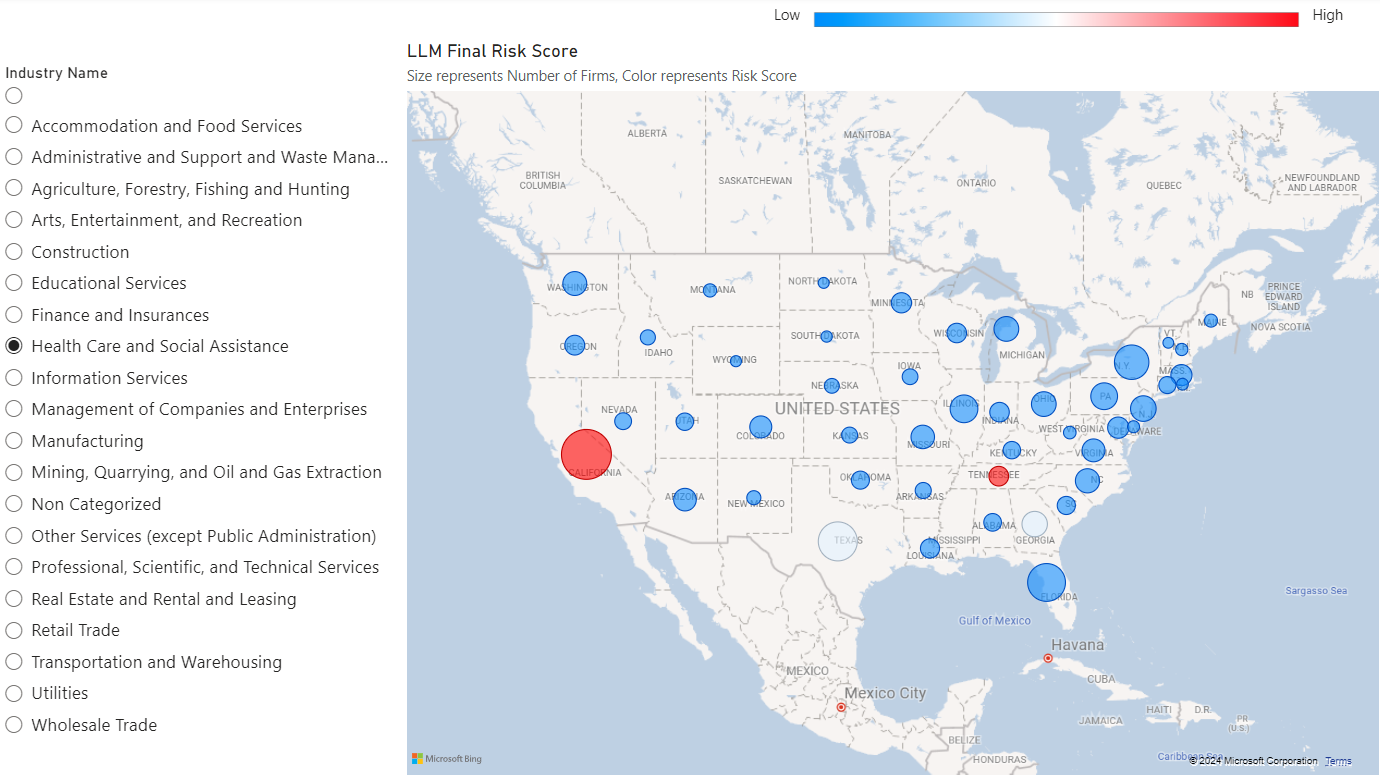


Figure 27: Risk Model 1, Geographical Visualization –LLM Prompt Injection in Health Care Sector.

In this visualization, we've depicted the geographic distribution of LLM Prompt Injection risks within the healthcare sector across the United States. Each point on the map represents a healthcare firm, with the size indicating the number of firms in a region and the color intensity reflecting the severity of the risk. This visualization method is particularly effective in highlighting areas like California and Texas, which show a larger healthcare sector facing significant risks. By creating this map, our goal was to emphasize the areas that need urgent attention for resource allocation and cybersecurity enhancements. This data-driven insight is crucial for us and other stakeholders in the healthcare industry as we work to formulate strategic responses to safeguard patient data and ensure operational continuity against such cyber threats.

## Risk Model 2 – Risk Framework

For risk model 2, the advantage over model 1 lies in its comprehensive approach. Unlike model 1, which solely focuses on cyber-attack actions and the number of firms across industries, model 2 takes into account additional factors such as assets, attributes, and impacts. This broader scope enables a more thorough assessment of risk, providing a deeper understanding of potential vulnerabilities and their consequences.

A map of the united states with blue dots

Description automatically generated

Figure 28: Risk Score 2, Geographical Visualization – Financial Sector (overall LLM attacks).

In the final risk score, California emerges with the highest risk score, akin to the findings of risk model 1. Following California are Texas and New York, indicating that these states consistently exhibit significant risk factors across both models. California, Texas, and New York exhibit high cyber risk scores in the financial sector for several reasons. Firstly, as home to major financial hubs and institutions, they are prime targets for cybercriminals seeking to exploit vulnerabilities and access valuable financial data. Secondly, their advanced technological infrastructure and dense populations increase the potential attack surface for cyber threats. Thirdly, stringent regulatory environments create complexities and compliance challenges, which cybercriminals may exploit. Lastly, the interconnected nature of the financial sector means that attacks in these states can have far-reaching consequences, amplifying their significance. These factors collectively contribute to the elevated cyber risk scores observed in California, Texas, and New York.

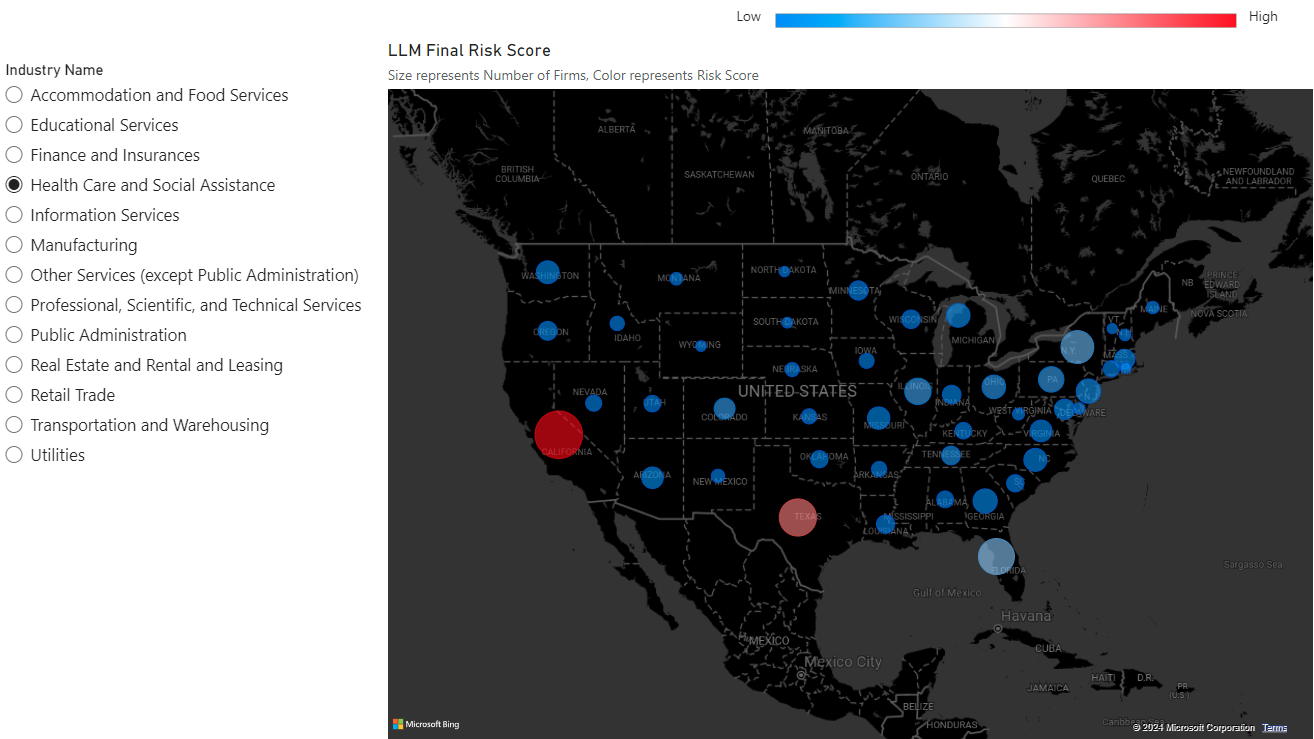


Figure 29: Risk Score 2, Geographical Visualization – Health Care Sector (overall LLM attack).

Similarly, California has the highest risk score followed by Texas. In California and Texas, the healthcare sector faces high cyber risk due to its wealth of sensitive patient data, stringent regulatory requirements like HIPAA, increasing adoption of digital technologies, interconnectedness among healthcare stakeholders, and a history of cyber incidents. These factors make healthcare organizations prime targets for cybercriminals seeking to exploit vulnerabilities and access valuable patient information.

# Findings

Through our project, we have uncovered several key findings that shed light on the landscape of cyber threats and vulnerabilities across various industry sectors.

Increasing Cyber Threats: Our analysis revealed a concerning trend of increasing cyber threats, with a surge in the number of reported incidents across different sectors. This trend is consistent with global cybersecurity reports, indicating a growing risk landscape fueled by factors such as the proliferation of digital technologies and the sophistication of threat actors.

Sector-Specific Vulnerabilities: We found that certain industry sectors are more vulnerable to specific types of cyber threats. For example, the education sector faces a higher risk of attacks from organized crime groups, while the financial sector is frequently targeted by hackers exploiting vulnerabilities in financial systems.

Adversary Tactics and Techniques: By mapping VERIS actions to MITRE ATT&CK techniques, we gained insights into adversary tactics and techniques employed in cyber attacks. This analysis revealed common patterns of behavior among threat actors, such as the use of phishing and malware to gain unauthorized access to systems and data.

Geographical Trends: Our dynamic geographical dashboards highlighted geographical trends in cyber incidents, enabling stakeholders to identify high-risk areas and allocate resources accordingly. For example, we observed a higher frequency of attacks in regions with a dense concentration of financial institutions or technology companies.

Risk Inequality and Profiling: Utilizing the Lorenz curve and Gini coefficient, we quantified risk inequality among industry sectors and attack techniques. This analysis enabled us to prioritize cybersecurity measures based on the severity and prevalence of different threats, leading to more targeted risk mitigation strategies.

Data Limitations and Challenges: Despite the insights gained, we encountered challenges related to data quality and completeness. The VERIS dataset, while comprehensive, contains some incomplete records and lacks certain key information such as adversary details. Additionally, web scraping techniques may yield incomplete or biased results from news articles and other sources.

Continuous Improvement Opportunities: Our project highlighted the importance of continuous improvement and refinement of our risk assessment and profiling framework. Regular updates, feedback loops, and collaboration with industry partners are essential for ensuring the relevance and effectiveness of our analyses in addressing evolving cyber threats.

In conclusion, our findings underscore the need for proactive cybersecurity measures and collaborative efforts to address the growing cyber threat landscape. By leveraging data-driven insights and advanced analytical techniques, organizations can better understand, mitigate, and respond to cyber risks, ultimately enhancing their resilience in an increasingly digital world.

# Summary

Our project focused on developing a comprehensive risk assessment and profiling framework to analyze cyber threats across various industry sectors. Through the integration of datasets such as the VERIS dataset, Statistics of US Businesses (SUSB), and web-scraped news articles, we gained insights into the evolving cyber threat landscape and identified key trends, patterns, and vulnerabilities.

One of the notable findings of our analysis is the increasing frequency and sophistication of cyber-attacks across different sectors. We observed sector-specific vulnerabilities, with certain industries being more prone to specific types of threats. For example, the financial sector faces a higher risk of hacking attempts, while the education sector is targeted by organized crime groups.

By mapping VERIS actions to MITRE ATT&CK techniques, we gained insights into adversary tactics and techniques, enabling us to better understand the modus operandi of threat actors. Our dynamic geographical dashboards highlighted geographical trends in cyber incidents, allowing stakeholders to identify high-risk areas and allocate resources effectively.

Using the Lorenz curve and Gini coefficient, we quantified risk inequality among industry sectors and attack techniques, enabling us to prioritize cybersecurity measures based on the severity and prevalence of different threats. However, we also encountered challenges related to data quality and completeness, emphasizing the need for continuous improvement and collaboration in addressing cyber risks.

In conclusion, our project underscores the importance of proactive cybersecurity measures and data-driven insights in mitigating cyber threats. By leveraging advanced analytical techniques and fostering collaboration across industries, organizations can enhance their resilience and effectively combat the evolving cyber threat landscape.

# Future work

Our project has laid a solid foundation for further research and development in the field of cybersecurity risk assessment and profiling. Several avenues for future work have emerged from our findings and analysis:

Enhanced Data Collection: Further efforts can be made to improve the quality and completeness of data collected from various sources. This may involve expanding the scope of datasets, leveraging additional sources of information, and implementing more sophisticated data cleaning and preprocessing techniques.

Advanced Analytical Techniques: Future work can explore the application of advanced analytical techniques, such as machine learning and natural language processing, to extract deeper insights from the data. These techniques can help identify complex patterns, predict future cyber threats, and automate aspects of the risk assessment process.

Dynamic Risk Assessment: Developing dynamic risk assessment models that continuously monitor and adapt to changing cyber threats and industry trends. This may involve real-time analysis of streaming data sources, integration of threat intelligence feeds, and the use of adaptive risk scoring algorithms.

Sector-Specific Mitigation Strategies: Tailoring cybersecurity mitigation strategies to specific industry sectors based on their unique risk profiles. This could involve the development of sector-specific best practices, guidelines, and training programs to help organizations better defend against cyber threats.

Collaborative Research and Information Sharing: Encouraging collaboration and information sharing among industry stakeholders, academia, and government agencies to improve cybersecurity resilience. This could include the establishment of industry consortia, joint research projects, and the creation of shared threat intelligence platforms.

User-Friendly Visualization Tools: Developing user-friendly visualization tools and dashboards that enable stakeholders to easily interpret and act on the insights derived from the data. These tools should be accessible to non-technical users and provide actionable recommendations for improving cybersecurity posture.

Ethical Considerations: Addressing ethical considerations related to data privacy, security, and consent in cybersecurity research and analysis. Future work should prioritize ethical practices and compliance with relevant regulations to ensure the responsible use of data.

By pursuing these avenues for future work, we can further advance our understanding of cyber threats and develop more effective strategies for mitigating risk and protecting organizations from cyber-attacks.

Appendix

# Appendix A: Glossary

|  |  |
| --- | --- |
| Term | Definition |
| LLM | Large Language Model |
| VERIS | Vocabulary for Event Recording and Incident Sharing |
| MITRE ATLAS | MITRE Adversarial Tactics, Techniques, and Common Knowledge |
| CENSUS | Comprehensive Epidemiologic Data Resource |
| SMB Profiling | Server Message Block Profiling |
| Lorenz's Curve | A graphical representation of the cumulative distribution function of a probability distribution |
| Risk Analysis | The process of identifying and assessing potential risks |
| Data Scraping | The process of extracting data from websites |
| Risk Metrics | Quantitative measures used to assess and compare risks |
| Cybersecurity Risk | The potential for loss or harm resulting from cybersecurity threats |
| Cyber Risk | The risk of financial loss, disruption, or damage to an organization's reputation due to cybersecurity incidents |
| Thread | A potential event that could compromise the security of a computer system or network |
| Vulnerability | Weakness in a system's defenses that could be exploited by a threat actor |
| Information Value | The importance or sensitivity of information to an organization |
| Consequence | The impact or outcome of a cybersecurity incident |
| Power BI | Business intelligence tool developed by Microsoft |
| NAICS | North American Industry Classification System |
| SUSB | Statistics of U.S. Businesses |

Table 25: Glossary Table.

# Appendix B: GitHub Repository

**Overview**

Our GitHub repository contains the codebase and datasets for our LLM Attack Analysis project. Leveraging datasets such as VERIS, SUSB, and MITRE ATLAS, we provide insights into LLM attack patterns and vulnerabilities across industries.

The repository is organized into sections: Dataset, Preprocessing, Visualization, and Mapping. The Dataset section contains raw datasets, while Preprocessing includes scripts for cleaning and transformation. Visualization offers visualizations of LLM attack patterns, and Mapping contains scripts for geographical distribution mapping.

To get started, clone the repository and follow the instructions in the README.md file to install dependencies and run the code. We welcome contributions to enhance our understanding of LLM attacks and improve our analysis.

GitHub Repository Link

<https://github.com/RahulKS1999/DAEN690_LLM_Risk_Assessment->

GitHub Repository Contents

Dataset

Preprocessing

Visualization

Mapping

Readme.md

# Appendix C: Risks

Sprint 1 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Data Availability | Insufficient availability of relevant literature and resources | High | High | Conduct a thorough search across multiple databases and repositories to ensure comprehensive coverage of literature and resources. Collaborate with subject matter experts to identify key sources. |
| Time Constraints | Limited time to conduct a comprehensive literature review | Medium | High | Establish a clear timeline and prioritize tasks. Allocate sufficient time for literature search and review. Utilize efficient research methods and tools to streamline the process. |
| Scope Definition | Unclear definition of project scope and objectives | Medium | Medium | Conduct regular meetings with stakeholders to clarify project goals and objectives. Develop a detailed project plan with defined deliverables and milestones. Continuously reassess and refine the scope as needed. |
| Technical Expertise | Lack of expertise in analyzing LLM applications and attacks | High | Medium | Invest in training and upskilling team members in LLM technology and cybersecurity concepts. Seek guidance from external experts or consultants if needed. Utilize online resources and tutorials for self-learning. |
| Data Quality | Inaccurate or incomplete data in initial datasets | Medium | Medium | Implement data validation and cleaning processes to identify and correct errors. Cross-reference data from multiple sources to ensure accuracy and completeness. Collaborate with data providers to address any issues or discrepancies. |

Table 26: Sprint 1 Risks

During Sprint 1, our team faced challenges such as limited availability of relevant literature on Large Language Models (LLMs) and technical expertise in analyzing LLM applications and cyber threats. We addressed these risks by conducting extensive searches, investing in training, and seeking guidance from external experts. However, gaps in literature coverage and the need for more structured training sessions were identified as areas for improvement.

We successfully identified and mitigated risks related to data quality and time constraints by implementing data validation processes and establishing clear timelines. However, unanticipated risks such as scope definition issues and technical challenges arose, emphasizing the importance of ongoing communication and collaboration.

Overall, our experience highlighted the need for thorough planning, effective risk management, and continuous adaptation. In future projects, we will prioritize early engagement with experts, allocate sufficient time for research and training, and maintain open communication to ensure project success.

Sprint 2 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Data Acquisition | Difficulty in extracting relevant information from VERIS and MITRE ATLAS datasets due to complexity | High | High | Allocate additional resources for data wrangling, leverage automated tools, streamline preprocessing |
| Web Scraping | Limited availability of reliable web scraping tools for real-time data gathering on LLM attacks | Medium | High | Expand search criteria, explore alternative scraping techniques, optimize data collection process |

Table 27: Sprint 2 Risks

During Sprint 2, the team faced risks associated with data acquisition, particularly the challenge of extracting relevant information from the VERIS and MITRE ATLAS datasets. This risk was compounded by the complexity of the datasets and the need for preprocessing and cleaning. To mitigate this, we allocated additional resources for data wrangling and leveraged automated tools where possible to streamline the process.

Another risk was the limited availability of web scraping tools for gathering real-time data on LLM attacks. While we initially anticipated this challenge, the impact was greater than expected due to the scarcity of reliable sources and the need for manual intervention in data collection. To address this, we expanded our search criteria and explored alternative scraping techniques, ultimately improving the efficiency of our data gathering process.

Despite these challenges, the team successfully navigated the complexities of data acquisition and analysis, demonstrating resilience and adaptability in the face of unforeseen obstacles. Looking ahead, we will continue to refine our processes, enhance our technical capabilities, and proactively address risks to ensure the success of future projects.

Sprint 3 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Data inconsistency | Inconsistencies or errors in the data may lead to inaccurate analysis results | Medium | High | Thoroughly clean and preprocess data, validate data sources, conduct exploratory data analysis to identify anomalies |
| Lack of domain expertise | Limited understanding of the dataset and its context may lead to misinterpretation of results | Medium | Medium | Collaborate with domain experts, seek guidance from data analysts or researchers familiar with the dataset, invest time in understanding the data and its nuances |
| Time constraints | Insufficient time for thorough data exploration and analysis may result in incomplete or rushed findings | High | Medium | Prioritize tasks, focus on high-impact areas, allocate sufficient time for data exploration, communicate any constraints or delays to stakeholders |

Table 28: Sprint 3 Risks

During Sprint 3, as we delved into exploratory data analysis, several risks emerged that required attention.

One significant risk was data inconsistency, given the complexity of the datasets. We implemented rigorous data cleaning and preprocessing to mitigate this risk, ensuring the accuracy of our analysis.

Another risk was the lack of domain expertise among team members, which could lead to misinterpretation of results. We addressed this by seeking guidance from domain experts and investing time in understanding the data better.

Time constraints were also a concern, as completing data exploration within a limited timeframe was challenging. While we prioritized tasks and focused on high-impact areas, managing time effectively remained a challenge.

Overall, the team accurately identified major risks, but unanticipated complexities and the need for more domain expertise arose. In the future, a comprehensive understanding of the data, involving domain experts from the start, and better time management strategies could mitigate these risks more effectively.

Sprint 4 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Lack of expertise in risk analysis | Team members may lack experience in risk analysis, leading to inaccurate results | Medium | High | Seek guidance from experts in risk analysis, conduct thorough research on risk assessment methodologies |
| Complexity of implementing SMB Profiling method | Implementing the SMB Profiling method may be complex and time-consuming | High | High | Break down the process into smaller tasks, allocate sufficient time for each task, and seek assistance from experts if needed |

Table 29: Sprint 4 Risks

During Sprint 4, the focus was on forming a framework to calculate risk metrics using the SMB Profiling method and Lorenz's curve. Several risks were identified during this phase.

One significant risk was the team's lack of expertise in risk analysis. To mitigate this, we sought guidance from experts in the field and conducted thorough research on risk assessment methodologies. However, this risk could have been better anticipated, and more emphasis could have been placed on training team members in risk analysis techniques from the beginning of the project.

Another challenge was the complexity of implementing the SMB Profiling method. While we anticipated this risk, it proved to be more time-consuming than initially expected. Breaking down the process into smaller tasks and seeking assistance from experts helped mitigate this risk to some extent. However, better time management and allocation of resources could have improved the efficiency of this process.

In conclusion, while the team accurately identified major risks during Sprint 4, unanticipated complexities and the need for more expertise in risk analysis were encountered. To improve similar projects in the future, it is essential to allocate resources effectively, seek guidance from experts early on, and prioritize data privacy and compliance throughout the project lifecycle.

Sprint 5 Risks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Risk | Description | Probability | Impact | Mitigation |
| Data inconsistency | Inconsistent or incomplete data may affect mapping accuracy | Medium | Medium | Conduct thorough data validation and cleaning processes |
| Technical challenges with visualization | Developing the visualization dashboard may encounter technical difficulties | High | High | Allocate sufficient time for development and testing, seek assistance from visualization experts if needed |
| Stakeholder expectations not met | Stakeholders may have high expectations for the visualization dashboard, leading to dissatisfaction | Medium | High | Ensure clear communication with stakeholders regarding project scope, timelines, and limitations |

Table 30: Sprint 5 Risks

During Sprint 5, the team focused on mapping LLM attack techniques and developing a visualization dashboard to showcase risk across industries. Several risks were identified during this phase.

One significant risk was data inconsistency, which could affect the accuracy of the mapping process. To mitigate this, the team conducted thorough data validation and cleaning processes. While this risk was accurately identified and addressed, better data management practices from earlier sprints could have minimized the likelihood of encountering data inconsistencies.

Another challenge was the technical complexity of developing the visualization dashboard. While the team anticipated this risk, there were unexpected technical difficulties encountered during development. To mitigate this, the team allocated sufficient time for development and testing and sought assistance from visualization experts. However, better planning and coordination with the development team could have reduced the impact of these challenges.

Additionally, there was a risk of stakeholder expectations not being met, as stakeholders may have had high expectations for the visualization dashboard. To mitigate this, the team ensured clear communication with stakeholders regarding project scope, timelines, and limitations. However, better stakeholder engagement and management throughout the project could have helped align expectations more effectively.

In conclusion, while the team accurately identified major risks during Sprint 5, unanticipated technical challenges and stakeholder expectations were encountered. To improve similar projects in the future, it is essential to prioritize data quality, allocate sufficient time for development and testing, and maintain clear communication with stakeholders throughout the project lifecycle.

# Appendix D: Agile Development

Scrum Methodology

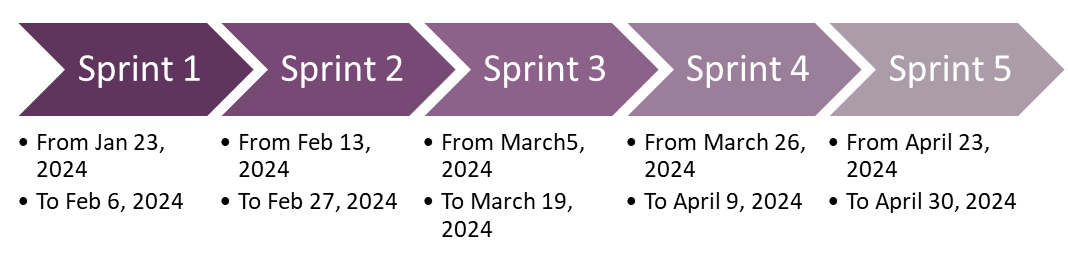


Figure 30: Sprint project dates.

The team made a deliberate effort to adopt the Scrum methodology for our data analytics engineering project. Initially, adapting to Scrum was challenging as it required a shift in mindset and work approach for some team members. However, as we progressed through the project, we found that Scrum provided a structured framework that helped us manage our tasks more efficiently.

We conducted daily scrum meetings to discuss progress, challenges, and plans for the day. These meetings were instrumental in keeping everyone aligned and informed about the project's status. While initially challenging to maintain consistency with daily scrums, the team eventually found a rhythm and appreciated the regular communication it provided.

Using the YouTrack tool to manage the project was both beneficial and challenging. It offered a centralized platform for task management, issue tracking, and collaboration, which helped keep the team organized. However, some team members found the tool complex to navigate and required additional time to become proficient in its use. Despite this initial learning curve, the team recognized the value of YouTrack in streamlining project management and facilitating communication.

One key lesson learned from using Scrum and YouTrack was the importance of ongoing communication and collaboration. Regular scrum meetings ensured that everyone was on the same page and allowed us to quickly address any issues or roadblocks. Additionally, thorough documentation and organization within YouTrack helped us stay organized and track progress effectively.

Overall, while there were challenges in adapting to Scrum and using YouTrack, the experience was valuable in improving project management practices and fostering collaboration within the team. Moving forward, we would recommend new project teams invest time upfront in understanding Scrum principles and familiarizing themselves with project management tools to maximize their effectiveness.

Sprint 1 Analysis

During Sprint 1, the team focused on conducting a literature review to understand LLM (Language Model Manipulation) applications in various fields and LLM-based attacks. Initially, the team identified user stories by breaking down the research objectives into manageable tasks, such as identifying key papers on LLM applications, understanding the types of LLM-based attacks, and summarizing findings from academic sources.

Overall, the team performed well in gathering relevant literature and extracting valuable insights from various research papers. However, managing activities during the sprint proved to be somewhat challenging, as some team members found it difficult to prioritize tasks effectively and stay organized.

One thing the team did well was establishing clear objectives for the literature review and breaking down tasks into smaller, actionable items. This approach helped maintain focus and ensure progress throughout the sprint. Additionally, the team maintained open communication channels to discuss findings, share resources, and address any challenges that arose.

In terms of improvements, the team could have allocated more time for initial planning and task estimation to better manage the workload. Some team members struggled with time management and underestimated the effort required for certain tasks, leading to some tasks being rushed or incomplete by the end of the sprint.

Moving forward, the team recognized the importance of setting realistic expectations and allocating sufficient time for each task. Additionally, better coordination and communication during sprint planning could help ensure that all team members are aligned on priorities and timelines. Overall, Sprint 1 provided valuable lessons in effective task management and collaboration, which will inform our approach in future sprints.

Sprint 2 Analyis

During Sprint 2, the team focused on gathering data resources for the project, including VERIS, MITRE ATLAS, and the US Census. Additionally, the team conducted web scraping to collect articles related to LLM (Language Model Manipulation). To manage these tasks, the team identified user stories that encompassed different aspects of data collection, such as identifying relevant datasets, scraping articles, and organizing the gathered data.

Overall, the team performed well in accessing and utilizing the selected data resources. However, managing activities during the sprint presented some challenges, particularly in coordinating the collection and organization of various datasets. Some team members found it difficult to navigate and extract relevant information from the datasets, leading to delays in completing certain tasks.

One aspect the team did well was leveraging web scraping techniques to gather real-time data from news articles. This approach helped supplement the structured datasets with current information and insights related to LLM. Additionally, the team maintained regular communication and collaboration to share progress, discuss challenges, and provide support to each other.

However, there were areas for improvement in task estimation and prioritization. Some tasks took longer than anticipated, resulting in incomplete work by the end of the sprint. In hindsight, the team could have allocated more time for data preprocessing and cleaning, as well as for training team members on effective web scraping techniques.

Looking ahead, the team recognized the importance of refining their data collection processes and ensuring clearer task prioritization in future sprints. Additionally, providing additional training and support for team members in areas such as data preprocessing and web scraping will be beneficial for improving efficiency and effectiveness. Overall, Sprint 2 provided valuable insights into the complexities of data gathering and highlighted areas for improvement in project management and execution.

Sprint 3 Analysis

During Sprint 3, the team focused on conducting exploratory analysis on the gathered datasets and began exploring risk analysis techniques. The team identified user stories related to exploring the datasets, analyzing patterns and trends, and gathering information for risk analysis.

Overall, the team performed well in conducting exploratory analysis and gathering insights from the datasets. However, managing activities during the sprint presented some challenges, particularly in balancing the exploration of datasets with the need to start risk analysis. Some team members found it difficult to prioritize tasks and allocate time effectively between exploration and analysis.

One aspect the team did well was collaborating to share findings and insights from the datasets. Regular discussions helped identify interesting patterns and trends, laying the groundwork for further analysis in subsequent sprints. Additionally, the team made progress in understanding the complexities of risk analysis and began laying the foundation for developing risk metrics.

However, there were areas for improvement in task prioritization and time management. Some team members felt overwhelmed by the volume of data and struggled to focus on specific analysis tasks. In hindsight, the team could have benefited from more structured planning and clearer goals for the exploratory analysis phase.

Looking ahead, the team recognized the importance of refining their analysis techniques and establishing clearer guidelines for task prioritization. Incorporating feedback from Sprint 3, the team aimed to streamline their approach to data exploration and analysis in future sprints. Additionally, providing additional support and training for team members in data analysis techniques would help improve overall performance and efficiency. Overall, Sprint 3 provided valuable insights into the challenges of data exploration and laid the groundwork for more focused analysis in subsequent sprints.

Sprint 4 Analysis

During Sprint 4, the team focused on risk analysis, particularly on forming a framework to calculate risk metrics using the SMB Profiling method and Lorenz's curve. The team identified user stories related to understanding the SMB Profiling method, implementing the Lorenz curve, and developing risk metrics based on the analysis.

Overall, the team performed well in understanding the concepts and methodologies behind risk analysis. They effectively divided tasks among team members and collaborated to ensure smooth progress. However, managing activities during the sprint presented some challenges, particularly in integrating the various components of risk analysis into a cohesive framework.

One aspect the team did well was conducting thorough research on the SMB Profiling method and Lorenz's curve. They demonstrated a clear understanding of the principles and were able to apply them effectively to the project. Additionally, the team successfully implemented the Lorenz curve to visualize risk distributions across industries, providing valuable insights into the data.

However, there were areas for improvement in integrating the SMB Profiling method with the risk analysis framework. Some team members found it challenging to translate the theoretical concepts into practical applications within the project context. In hindsight, the team could have benefited from more structured guidance and hands-on training in applying the SMB Profiling method.

Looking ahead, the team recognized the importance of refining their understanding of risk analysis methodologies and their application to the project. They planned to seek additional support and resources to deepen their knowledge and improve their ability to integrate complex concepts into the project framework. Overall, Sprint 4 provided valuable insights into the complexities of risk analysis and highlighted areas for further development and improvement in future sprints.

Sprint 5 Analysis

During Sprint 5, the team focused on two main tasks: mapping LLM attack techniques and creating a visualization dashboard showcasing risk across industries. The team identified user stories related to researching LLM attack techniques, defining data visualization requirements, and developing the dashboard.

Overall, the team performed well in completing the tasks for this sprint. They effectively researched and mapped LLM attack techniques, identifying common patterns and trends across different industries. Additionally, they successfully developed a visualization dashboard to present risk assessments for various industries based on the data analysis conducted in previous sprints.

However, managing activities during this sprint presented some challenges, particularly in coordinating the development of the visualization dashboard. The team encountered difficulties in integrating the data analysis results with the dashboard design, resulting in delays and some confusion among team members.

One aspect the team did well was their collaboration in researching and mapping LLM attack techniques. They leveraged their understanding of the data and methodologies from previous sprints to create insightful visualizations that effectively communicated the risk landscape to stakeholders.

However, there were areas for improvement in project management and communication during this sprint. The team could have benefited from clearer task assignments and more frequent check-ins to ensure that everyone was on the same page regarding the dashboard development process.

Looking ahead, the team recognized the importance of refining their project management processes and communication strategies to ensure smoother coordination in future sprints. They planned to implement more structured task tracking and communication protocols to avoid similar challenges in upcoming iterations.

Overall, Sprint 5 provided valuable lessons in data visualization and project coordination, highlighting the importance of clear communication and effective collaboration in achieving project goals.

Reference

# Works Cited

[1] J. Yang *et al.*, “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.” arXiv, Apr. 27, 2023. doi: 10.48550/arXiv.2304.13712.

[2] OpenAI *et al.*, “GPT-4 Technical Report.” arXiv, Mar. 04, 2024. doi: 10.48550/arXiv.2303.08774.

[3] “Introducing LLaMA: A foundational, 65-billion-parameter language model.” Accessed: Apr. 16, 2024. [Online]. Available: https://ai.meta.com/blog/large-language-model-llama-meta-ai/

[4] “Free Dolly: Introducing the World’s First Truly Open Instruction-Tuned LLM,” Databricks. Accessed: Apr. 16, 2024. [Online]. Available: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm

[5] “Number of ChatGPT Users (Apr 2024).” Accessed: Apr. 16, 2024. [Online]. Available: https://explodingtopics.com/blog/chatgpt-users

[6] “Large language models in medicine | Nature Medicine.” Accessed: Apr. 16, 2024. [Online]. Available: https://www.nature.com/articles/s41591-023-02448-8

[7] S. Wu *et al.*, “BloombergGPT: A Large Language Model for Finance.” arXiv, Dec. 21, 2023. doi: 10.48550/arXiv.2303.17564.

[8] “ChatGPT passing USMLE shines a spotlight on the flaws of medical education | PLOS Digital Health.” Accessed: Apr. 16, 2024. [Online]. Available: https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000205

[9] H. Pearce, B. Tan, B. Ahmad, R. Karri, and B. Dolan-Gavitt, “Examining Zero-Shot Vulnerability Repair with Large Language Models.” arXiv, Aug. 15, 2022. doi: 10.48550/arXiv.2112.02125.

[10] C. S. Xia, M. Paltenghi, J. L. Tian, M. Pradel, and L. Zhang, “Fuzz4All: Universal Fuzzing with Large Language Models.” arXiv, Jan. 15, 2024. doi: 10.48550/arXiv.2308.04748.

[11] S. Y. Feng *et al.*, “A Survey of Data Augmentation Approaches for NLP.” arXiv, Dec. 01, 2021. doi: 10.48550/arXiv.2105.03075.

[12] “What Are Large Language Models (LLMs)?,” Palo Alto Networks. Accessed: Apr. 16, 2024. [Online]. Available: https://www.paloaltonetworks.com/cyberpedia/large-language-models-llm

[13] “Wenxuan Wang’s Homepage @ The Chinese University of Hong Kong.” Accessed: May 02, 2024. [Online]. Available: https://jarviswang94.github.io/

[14] “Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory.” Accessed: May 02, 2024. [Online]. Available: https://confaide.github.io/

[15] A. Esmradi, D. W. Yip, and C. F. Chan, “A Comprehensive Survey of Attack Techniques, Implementation, and Mitigation Strategies in Large Language Models.” arXiv, Dec. 18, 2023. doi: 10.48550/arXiv.2312.10982.

[16] B. Vidgen *et al.*, “SimpleSafetyTests: a Test Suite for Identifying Critical Safety Risks in Large Language Models.” arXiv, Feb. 16, 2024. doi: 10.48550/arXiv.2311.08370.

[17] “WormGPT - The Generative AI Tool Cybercriminals Are Using to Launch BEC Attacks | SlashNext.” Accessed: May 02, 2024. [Online]. Available: https://slashnext.com/blog/wormgpt-the-generative-ai-tool-cybercriminals-are-using-to-launch-business-email-compromise-attacks/

[18] M. Bano, D. Zowghi, and J. Whittle, “Exploring Qualitative Research Using LLMs.” arXiv, Jun. 23, 2023. doi: 10.48550/arXiv.2306.13298.

[19] “Enabling conversational interaction on mobile with LLMs.” Accessed: May 02, 2024. [Online]. Available: https://research.google/blog/enabling-conversational-interaction-on-mobile-with-llms/

[20] “Weaver: Foundation Models for Creative Writing,” ar5iv. Accessed: May 02, 2024. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2401.17268

[21] “Translating Web Search Queries into Natural Language Questions,” ar5iv. Accessed: May 02, 2024. [Online]. Available: https://ar5iv.labs.arxiv.org/html/2002.02631

[22] “Understanding LLM Hallucinations and how to mitigate them.” Accessed: May 02, 2024. [Online]. Available: https://kili-technology.com/large-language-models-llms/understanding-llm-hallucinations-and-how-to-mitigate-them

[23] “Where large language models can fail in business and how to avoid common traps,” Thoughtworks. Accessed: May 02, 2024. [Online]. Available: https://www.thoughtworks.com/en-us/insights/blog/generative-ai/where-large-language-models-fail-in-business-and-how-to-avoid-common-traps

[24] “ChatGPT: A Case Study on Copyright Challenges for Generative Artificial Intelligence Systems | European Journal of Risk Regulation | Cambridge Core.” Accessed: May 02, 2024. [Online]. Available: https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965

[25] “(PDF) Evaluating Local Language Models: An Application to Bank Earnings Calls.” Accessed: May 02, 2024. [Online]. Available: https://www.researchgate.net/publication/375440835\_Evaluating\_Local\_Language\_Models\_An\_Application\_to\_Bank\_Earnings\_Calls

[26] “List of victimized companies of MOVEit Cyber Attack - Cybersecurity Insiders.” Accessed: May 02, 2024. [Online]. Available: https://www.cybersecurity-insiders.com/list-of-victimized-companies-of-moveit-cyber-attack/

[27] “Statistics of U.S. Businesses (SUSB).” Accessed: May 02, 2024. [Online]. Available: https://www.census.gov/programs-surveys/susb.html

[28] “North American Industry Classification System (NAICS) U.S. Census Bureau.” Accessed: May 02, 2024. [Online]. Available: https://www.census.gov/naics/

[29] J. Jordon *et al.*, “Synthetic Data -- what, why and how?” arXiv, May 06, 2022. doi: 10.48550/arXiv.2205.03257.

[30] “MITRE ATLASTM.” Accessed: May 02, 2024. [Online]. Available: https://atlas.mitre.org/

[31] “Methods of Measuring the Concentration of Wealth: Publications of the American Statistical Association: Vol 9, No 70.” Accessed: May 02, 2024. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/15225437.1905.10503443

**This page intentionally left blank**