Orclever Proceedings of
Research and Development

https://doi.org/10.56038/oprd.v5i1.616

OR CLEVER
Science & Research Group

Conference Article

AI-Enhanced Cybersecurity

Vulnerability-Based Prevention, Defense, and Mitigation using Generative AI

Osman Çaylı[1*]

[1] VBT YAZILIM A.Ş., 0009-0006-9072-9146, Osman.Cayli@vbt.com.tr

Received 15 November 2024
Received in revised form 25 December 2024
In final form 87 December 2024

4th International Conference on Design, Research and Development
(RDCONF 2024)
December 19 - 20, 2024

Reference: Çaylı, O. (2024). AI-enhanced cybersecurity: Vulnerability-based prevention, defense, and mitigation using generative AI. Orclever Proceedings of Research and Development, 5(1), 655-667.

Abstract

The rapid evolution of cyberattacks, driven by increasingly sophisticated techniques and the proliferation of readily available AI tools, presents significant challenges for organizations worldwide. Traditional cybersecurity approaches often prove insufficient in addressing the speed, adaptability, and complexity of modern threats. The VULTURE project directly tackles these challenges by proposing a revolutionary AI-powered cybersecurity platform that leverages the capabilities of generative AI (GenAI) and large language models (LLMs) to enhance vulnerability prediction, automate penetration testing, improve intrusion detection, and enable advanced cyber-physical risk profiling. This paper will examine VULTURE's architecture, key technological innovations, anticipated impact, and future research directions.

The increasing sophistication and frequency of cyberattacks underscore the urgent need for innovative and adaptable cybersecurity solutions. Traditional approaches, often based on static rules and signature-based detection, struggle to keep pace with rapidly evolving threats, particularly the emergence of AI-driven attacks that can bypass conventional defenses and exploit previously unknown vulnerabilities (zero-day exploits). The shortage of skilled cybersecurity *professionals further exacerbates these challenges, limiting organizations' ability to effectively respond to emerging threats.*

The VULTURE project proposes a novel approach to cybersecurity leveraging the power of Large Language Models (LLMs). This paper explores the technical innovations presented in the VULTURE proposal, focusing on the application of LLMs for vulnerability prediction and automated penetration testing. We analyze the proposed methodologies and discuss their potential impact, highlighting opportunities and challenges. Further research is necessary to validate the efficacy and scalability of the proposed methods.

Keywords: Large Language Models (LLMs) in Cybersecurity, Zero-Day Vulnerability Detection, Automated Penetration Testing, Cyber-Physical Risk Assessment, Generative AI for Cybersecurity, Harmonized Threat Assessment

1. Introduction:

Cybersecurity practitioners today face an evolving landscape of sophisticated challenges that demand swift and intelligent responses. Among these, the rise of automated AI-driven cyberattacks poses a formidable threat, as adversaries increasingly harness artificial intelligence to launch targeted, adaptive, and high-frequency attacks. These AI-based attacks can quickly identify weaknesses, bypass standard defenses, and adapt to changing security postures in real time. Practitioners now contend with automation-enabled attack vectors, requiring equally advanced AI-driven defenses to counter these threats. The situation is further complicated by the complexity of modern software systems, which rely on thousands of open-source and third-party components in the supply chain, each potentially harboring vulnerabilities. Notable incidents like the SolarWinds attack1 and the Log4j vulnerability2 have demonstrated how attackers can exploit weak links in the software supply chain, infiltrating systems by compromising widely used components. This extended attack surface is difficult to monitor, creating entry points that put not only individual organizations but entire industries at risk.

[1] https://attack.mitre.org/campaigns/C0024/

[2] https://nvd.nist.gov/vuln/detail/CVE-2021-44228

Accelerated development cycles using CI/CD pipelines add another layer of risk by increasing the speed of deployment, leaving less time for thorough security testing. With software being released faster and updated more frequently, vulnerabilities can be missed, creating an environment ripe for exploitation. The shortage of specialized cybersecurity professionals further exacerbates these issues, limiting the ability of teams to identify, manage, and respond to emerging threats. This talent gap leaves organizations under-resourced in their efforts to secure systems that are becoming increasingly vulnerable due to rapid release cycles and complex supply chains. Compounding these challenges is the rise of cyber-physical threats, which combine digital attacks with physical consequences. In critical infrastructure sectors, such as energy and transportation, attackers can target connected physical devices, as seen in the 2021 Colonial Pipeline ransomware attack. Hybrid warfare scenarios also leverage cyber-physical threats, using cyberattacks to disrupt essential services and create instability. To address this array of challenges, cybersecurity practitioners must adopt a multi-layered approach that includes AI-driven defense strategies, robust supply chain risk management, and enhanced investment in both cybersecurity talent and infrastructure. Without such efforts, organizations and industries remain at significant risk in an increasingly interconnected world.

These problems can be addressed using artificial intelligence to increase the ability of cybersecurity practitioners to cope with ever-increasing threats without the need to increase their number. The focus must be in developing tools to make the human element more effective, by automating complex tasks like vulnerability research, penetration tests, detection of attacks and cyber-physical risk modelling. VULTURE proposes to address these four areas, as described in the next section.

2. Solution & System Architecture:

VULTURE will develop advanced cybersecurity tools based on LLM with concrete applications in the fields of (1) LLM-based vulnerability identification, (2) LLM-based automated pentesting, (3) Advanced AI-driven IDS to counter GenAI cyber-attacks and (4) GenAI-enhanced Cyber-physical risk modelling.

These four concepts are fully correlated. LLM-based vulnerability identification, which scans systems to uncover weaknesses, supplying critical insights for automated penetration testing to simulate real-world attacks on identified vulnerabilities. The results from automated pentesting enhance AI-driven IDS by equipping it with detailed attack patterns and vectors, enabling it to recognize and respond to sophisticated GenAI-based cyber threats in real time. This dynamic IDS information can be fed into cyber-physical risk modelling, which evaluates potential impacts on interconnected physical and digital systems, particularly valuable for safeguarding critical infrastructure. Together, these interconnected fields create a proactive and resilient security ecosystem that continuously identifies, tests, monitors, and assesses threats, enabling organizations to mitigate both cyber and physical risks effectively.

The following high-level diagram demonstrates how the VULTURE components interact and complement each other:

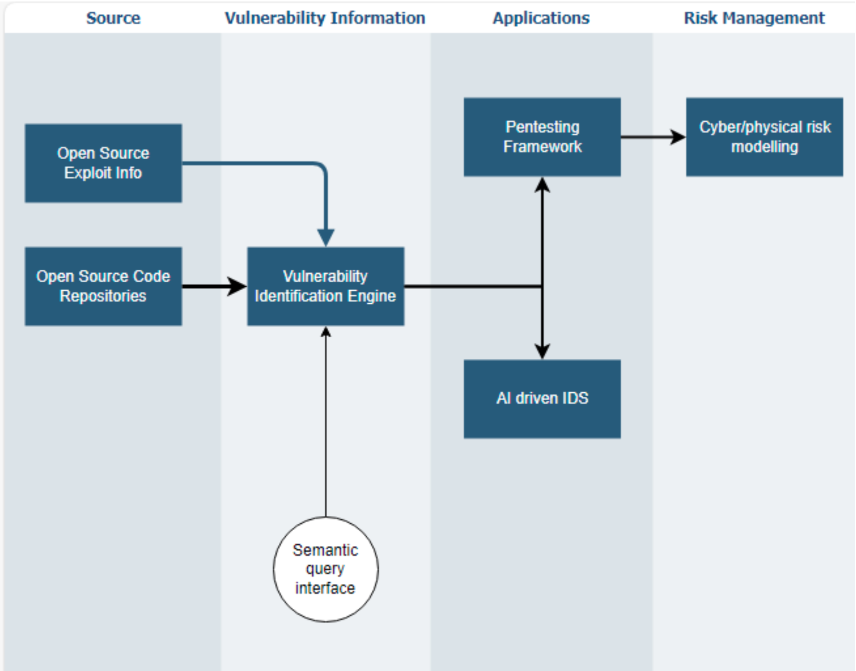
Source
Vulnerability Information
Applications
Risk Management
Open Source Exploit Info
Open Source Code Repositories
Vulnerability Identification Engine
Semantic query interface
Pentesting Framework
AI driven IDS
Cyber/physical risk modelling

Figure 1. VULTURE Components Architecture

2.1. LLM-based vulnerability identification

Large Language Models (LLMs) are increasingly leveraged to identify security vulnerabilities in open-source codebases by analyzing code patterns and detecting potential risks that may lead to exploits. These models, trained on vast repositories of code, can recognize vulnerabilities such as SQL injection, cross-site scripting, and buffer overflows by identifying suspicious patterns and anomalies in code logic. In VULTURE, we will use data from public vulnerabilities disclosure channels, including MITRE CVE, NIST NVD, Offensive Security Exploit-DB, OSVDB, Security Advisories from manufacturers (Microsoft, Google, Oracle, Adobe, others), Bug Bounty programs (HackerOne, Bugcrowd, Synack, others), CERT advisories, SecurityFocus Bugtraq, GitHub security advisories and other open-source vulnerability databases (OSV) to identify these vulnerable code segments and train the model to predict other possible security flaws on the public code repository.

This knowledge base will enable the following deliverables:

a) A natural language prompt that can be used to query the vulnerability database.
b) Automated code analysis and vulnerability scoring.
c) Prediction of “zero-day” vulnerabilities.

The current state-of-the-art in predicting vulnerabilities within open-source code leverages advanced machine learning techniques, particularly Large Language Models (LLMs) and Graph Neural Networks (GNNs). LLMs, such as CodeBERT and CodeLlama, are pre-trained on extensive code repositories, enabling them to understand code semantics and identify potential vulnerabilities.

Complementing LLMs, GNNs are employed to capture the structural aspects of code by representing it as graphs that include syntax, control flow, and data dependencies. Models such as Vul-LMGNN integrate pre-trained code language models with GNNs to detect vulnerabilities effectively. This approach constructs a code property graph that merges various code attributes and utilizes a gated code GNN to retain dependency information among these attributes. Evaluations across multiple real-world datasets have demonstrated that such models outperform previous methods, achieving higher F1 scores, especially on smaller datasets.

These advancements indicate a trend towards combining semantic understanding from LLMs with structural insights from GNNs, resulting in more accurate and efficient vulnerability detection in open-source codebases.

2.2. A Novel LLM-based Automated Penetration Testing Benchmark

Generative AI (GenAI), particularly LLMs, is reshaping the landscape of cybersecurity. These technologies have the potential to transform penetration testing, also known as ethical hacking or pen testing, a critical security measure that simulates cyberattacks to identify system vulnerabilities. This approach enables organisations to assess how well their systems can resist real-world attacks, uncovering potential weaknesses that malicious attackers could exploit.

It should be noted that penetration testing requires deep expertise and extensive training, making it challenging to execute effectively. To address this issue, researchers are investigating the automation of penetration testing using LLMs. Recent advancements in LLMs demonstrate promising potential to streamline the penetration testing process, allowing teams to prioritise critical risks and implement countermeasures swiftly. However, as research in this area remains in its early stages, there are currently very few comprehensive, open, end-to-end automated penetration testing benchmarks available to drive progress and evaluate the capabilities of these models in security contexts. This project aims to address this gap by introducing a novel open benchmark for LLM-based automated penetration testing, with a specific focus on identifying and responding to high-impact real-world zero-day vulnerabilities, which are particularly critical since they are unknown to defenders and have no existing patches. By leveraging this benchmark to simulate a wide range of potential attack scenarios, including real-world zero-day vulnerabilities and recent common vulnerabilities, this benchmark will enable security teams to identify systems and networks weakness more efficiently and effectively.

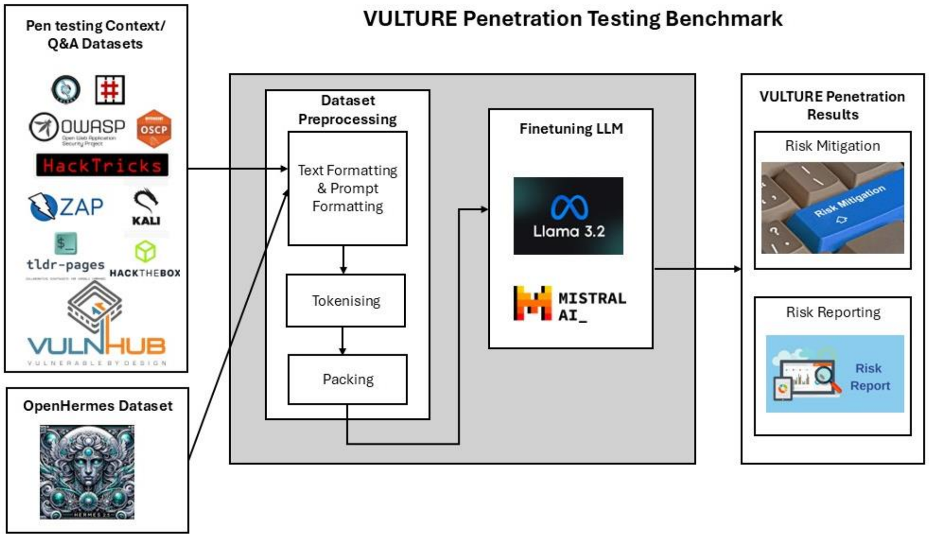

Figure 2. VULTURE Penetration Testing Benchmark Pipeline

In details, the VULTURE Penetration Testing Benchmark Pipeline, illustrated in Figure , aims to advance the automation of penetration testing through finetuning of open-source LLMs. It utilises the OpenHermes 2.5 dataset to enhance both general knowledge and conversational capabilities. By aggregating comprehensive penetration testing context datasets from diverse sources including those generated by VULTURE and open-source platforms, it provides extensive, specialised information in various formats critical for high-quality penetration testing. After data aggregation, the pipeline preprocesses the information through formatting, tokenization, and optimisation to prepare it for fine-tuning advanced LLMs, such as Mistral AI and Llama 3.2. This finetuning process tailors the models to detect vulnerabilities, assess risks, and deliver detailed security insights. Once deployed, the model operates in two core functions: risk mitigation, where it identifies potential vulnerabilities and recommends solutions, and risk reporting, where it outlines identified vulnerabilities and potential impacts. This empowers security professionals with actionable insights to enhance organisational security.

The LLM-based Automated Penetration Testing Benchmark will enhance the efficiency and speed of security teams in performing penetration tests. It will act as a digital assistant, aiding the penetration testers in every step of the way:

- Pen-Test Assistant: It will assist security professionals in both the planning and execution phases of penetration tests. The tool will provide technical guidance and real-time support by analyzing the outcomes of each test step. This dynamic interaction will enable the system to offer more accurate and context-aware advice throughout the testing process.
- Post-Execution Summary: Upon completion of the penetration tests, it will automatically generate a comprehensive report summarizing the results. This report can be tailored to meet specific user needs, ensuring flexibility in format and content as per the project's requirements.

2.3. Advanced Lightweight AI-driven Intrusion Detection System (IDS)

As GenAI technology advances, cyber attackers are also evolving, employing sophisticated GenAI and LLM tools to develop increasingly advanced methods for exploiting system vulnerabilities. This shift is driving demand for cutting-edge intrusion detection technologies, making it a crucial focus for cybersecurity providers and businesses alike. Consequently, the cybersecurity field is rapidly evolving to counter these complex, AI-driven threats, marking a new era in the ongoing battle between cyber defense and cybercrime.

To address this ongoing battle and enhance security, another core innovation of this project lies in the advanced AI-driven Intrusion Detection System (IDS). Unlike traditional IDS, VULTURE system leverages state-of-the-art deep learning algorithms to process vast amounts of complex network data in real-time, enabling precise anomaly detection and threat identification. This system operates as a virtual security analyst, providing continuously system monitoring, promptly identifying malicious activities, and generating alerts. The AI-driven IDS not only enhances security but also adapts to evolving cyber threats, providing a dynamic and robust detection mechanism. It is lightweight, promising enhanced accuracy, and has the potential for real-time detection of zero-day suspicious threats.

Figure presents an overview of the VULTURE Lightweight AI-driven IDS architecture. It leverages synthetic data augmentation and lightweight transformer-based modelling to provide efficient and effective intrusion detection. This IDS model will use VULTURE generated datasets along with public datasets, such as CIC-IDS2018 dataset, a widely recognised benchmark for intrusion detection in cloud environment. Since most of those datasets are usually highly imbalanced, synthetic data generation will be employed to balance the dataset and enhance model robustness and performance. The system plans to employ different models, such as a lightweight Vision Transformer (ViT) model, specifically designed to handle sequential features of network data. The models will output a classification that flags traffic as either normal or malicious traffic. If malicious activity is detected, a "Threat Alert" will be generated, indicating a potential security breach.

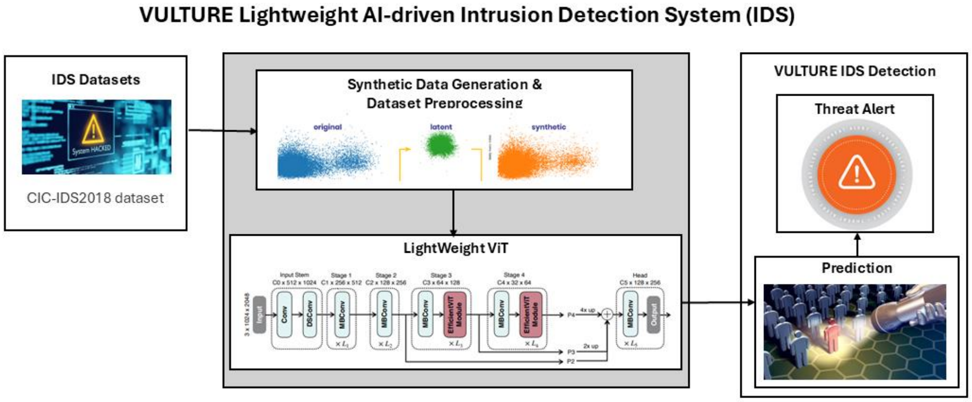

Figure 3. Advanced Lightweight AI-driven Intrusion Detection System (IDS)

2.4. Cyber-Physical threats via harmonised threat assessment

In tandem with the LLM cyber risk assessment tools, we shall also develop the ability to profile Cyber-Physical risks – which allows for the physical access and challenge to key systems and vulnerable cyber systems to be considered from the physical pen testing domain. This is key as increasingly distributed compute is developed. Additionally certain cyber protections can be undermined by physical penetration, which allows for the scoping of internal, insider and physical pen testing threats to be scored and evaluated within a representative volume of the physical and cyber-physical domain. Proliferation of cyber tools has been mirrored in Physical penetration capabilities; with harmonised approaches (rather than siloed) allowing for holistic threat and risk reduction.

This exploits existing architectures for Cyber Risk Profiling, but applies risk profiling methodologies from Physical Penetration methodologies, anchored within a representative volumetric space. This allows for considerations, such as visibility, noise and time to defeat a security measure – allowing for appropriate risk mitigation measures and allowing for the simulation of increasingly changing work and operating environments, such as co-located businesses, incubators and federated or distributed systems.

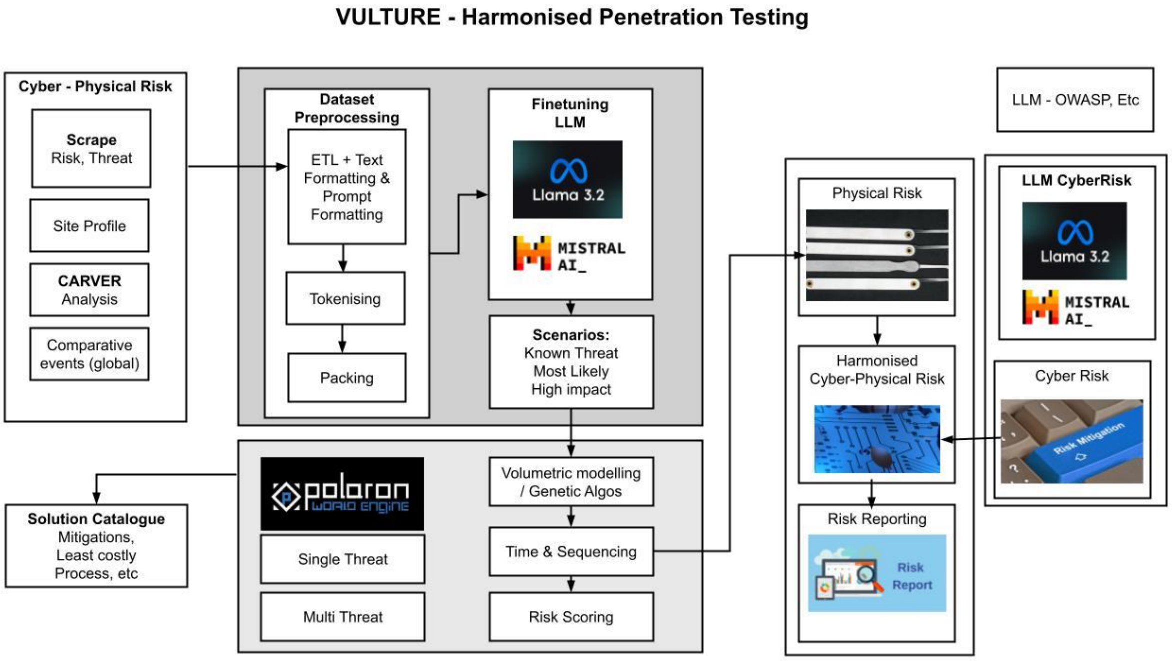

Figure 4. Vulture Harmonised Penetration Testing

This works by using the Cyber LLM and tuning models to produce a dynamic appraisal of adversary capability, intent, visibility and then profile this against representative security measures. This is considered within the scope of layered protection (e.g. https://www.npsa.gov.uk/physical-security) as well as a direct volumetric model of the facility. This permits varying levels of detail to be produced within an opportunistic digital twin, onto which the scenario and physical threats are estimated – resulting in a score of different compartments within a volume, and the infrastructure contained there. This can be cross-referenced against the Cyber Threat profiling – allowing for cost effective and balanced physical security considerations to be made in context of Cyber threats – and the location of key systems of infrastructure access points. This is modelled after the Canadian Harmonised Threat Reduction Analysis process; but exploits LLMs to provide both visibility and risk scores against the cyber-physical domain and appropriate measures taken.

Similar to zero-day errors this domain is an evolving field (with new technologies disseminating quickly) which allows for a dynamic risk scoring and re-scoring against the threat landscape. Additionally some Cyber Exploits and systems are only vulnerable via physical interactions, as well as recovery from certain cyber events (e.g. CloudStrike) requiring direct physical access. Thus this solution can also quantify recovery and physical access requirements in both defence and recovery stages of a Cyber Incident or Threat Assessment.

In summary, VULTURE proposes the following key innovations to tackle challenges in modern AI-driven cyber defense: (a) LLM-based vulnerability identification; (b) the development of a novel LLM Benchmark for automated penetration testing and zero-day vulnerability exploitation; (c) the creation of a lightweight AI-driven state-of-the-art IDS with the potential for real-time detection of zero-day suspicious threats, and (d) Cyber-Physical threats via harmonised threat assessment (Figure). To evaluate the potential effectiveness of these innovations, the project will implement statistical evaluations to assess the performance of developed LLMs in various penetration testing scenarios. Additionally, a controlled environment will be established in a green distributed data centre as a use case to simulate real-world cyber threats, allowing for conducting extensive testing of both the LLM Benchmark and the IDS under various attack scenarios. Alongside these testing, the project will also provide recommendations and mitigation strategies aimed at enhancing cybersecurity resilience.

3. Technological Innovations:

The VULTURE project outlines four key technological innovations centered around leveraging Large Language Models (LLMs) for enhanced cybersecurity:

1. Efficient Vulnerability Prediction and Identification: This involves training LLMs on extensive vulnerability datasets (MITRE CVE, NIST NVD, etc.) to identify known vulnerabilities and predict similar patterns in unclassified code. Natural language querying allows user-friendly access to this vulnerability database, and

an automated code analysis system provides vulnerability scoring. The system is further enhanced by incorporating Graph Neural Networks (GNNs) for structural code analysis, improving accuracy and efficiency, particularly for zero-day vulnerability prediction.

2. Automating Penetration Testing with Generative AI: This innovation focuses on developing an open benchmark for evaluating LLM-based automated penetration testing, addressing the current lack of comprehensive benchmarks in this area. A structured pipeline aggregates diverse penetration testing datasets and optimizes them for LLM fine-tuning. Fine-tuned advanced LLMs, like Mistral AI and Llama 3.2, enhance both general knowledge and conversational capabilities for penetration testing. A pen-test assistant tool provides real-time guidance and support, and automated post-execution reporting streamlines the process.
3. Next-Generation AI-Driven Intrusion Detection: A lightweight, real-time threat detection system using deep learning algorithms processes complex network data to identify malicious activities. Synthetic data augmentation improves model robustness and performance. This AI-driven IDS (Intrusion Detection System) adapts dynamically to evolving threats, enhancing accuracy and enabling real-time detection of zero-day threats.
4. Integrated Cyber-Physical Risk Profiling and AI-Driven Threat Detection: This combines cyber risk assessment tools with the ability to profile cyber-physical risks, considering how physical access impacts vulnerable cyber systems. A harmonized threat assessment framework integrates cyber and physical penetration risk assessments, using a 3D volumetric model that maps threats to specific locations. Tuned LLMs dynamically appraise adversarial capabilities and intent, producing real-time risk scores. Opportunistic digital twin simulations assess scenarios in various environments (co-located businesses, incubators, etc.). This approach adapts to evolving environments and incorporates the Canadian Harmonized Threat Reduction Analysis (HTRA) model.

In essence, VULTURE aims to create a proactive and adaptable cybersecurity ecosystem by combining the strengths of LLMs, GNNs, and other AI techniques to efficiently identify, assess, and mitigate a wider range of threats, including zero-day vulnerabilities and cyber-physical risks.

IV. Discussion and Conclusion

The VULTURE project aims to revolutionize cybersecurity by leveraging the power of Large Language Models (LLMs) and other AI techniques to address the increasing sophistication and frequency of cyberattacks, particularly those employing AI. The project's core innovations include:

- LLM-based Vulnerability Identification: Utilizing LLMs trained on extensive vulnerability datasets to predict vulnerabilities, including zero-day exploits, and automate code analysis for risk scoring.
- Automated Penetration Testing: Developing an open benchmark and automated tools to streamline penetration testing, enabling more efficient and effective identification of vulnerabilities.
- AI-driven Intrusion Detection System (IDS): Creating a lightweight, real-time IDS capable of detecting both known and unknown (zero-day) threats, adapting dynamically to evolving attack techniques.
- Harmonized Cyber-Physical Risk Profiling: Integrating cyber and physical risk assessments into a unified framework, using LLMs and volumetric modeling to assess threats and vulnerabilities in complex, interconnected systems.

The project employs a multi-layered approach, combining LLMs with GNNs and advanced deep learning algorithms to enhance accuracy, efficiency, and adaptability. A key component is the development of an open benchmark for LLM-based penetration testing, fostering collaboration and driving innovation in this field. The project considers a range of business outcomes, including improved defense against AI-powered attacks, early detection of zero-day vulnerabilities, automated penetration testing, and comprehensive cyber-physical risk modeling. The VULTURE project anticipates significant market impact, enhancing the capabilities of cybersecurity professionals and enabling organizations to better protect their systems and data against increasingly sophisticated threats. The consortium comprises organizations with expertise in various areas, from AI research and development to cybersecurity consulting and implementation, ensuring a holistic and collaborative approach. Finally, the project emphasizes the need for proactive, adaptive, and resilient cybersecurity strategies in response to the ever-evolving threat landscape.

References

[1] https://attack.mitre.org/campaigns/C0024/
[2] https://nvd.nist.gov/vuln/detail/CVE-2021-44228