Automated Threat Modelling Applied to LLM System Integration using NLP and Machine Learning – A Survey

Basavraj Gadade

Dept of Computer Engineering, JSPM University Pune

Prof. G. A. Patil

Dept of Computer Engineering, JSPM University Pune

Abstract: Applications such OpenAI's GPT-4, Google's Bard, and Meta's LLaMA have been including added to various systems, from chatbots and virtual assistants to decision-making systems which are done automatically powered by artificial intelligence. Nonetheless, there is a difficult to solve problem which stems from enormous security concerns posed by the use of these systems, including adversarial attacks, data poisoning, prompt injections, and model inversion attacks. Automated threat Modelling is one of the primary means of tackling the identification, assessment, and mitigation problems related to security risks while ensuring the safety and reliability of these LLM integrated systems. Strategies utilized in traditional cybersecurity approaches like STRIDE or DREAD do not cater to AI driven threat's changing nature, which poses additional problems. Instead, techniques NLP and ML are highly dynamic and effective at dealing with AI or configurable threats of security vulnerabilities. NLP aids in the formulation of inputs that are harmful or malicious through techniques like semantic analysis and anomaly detection, and ML-based security frameworks enhance real time assessments of risk and automated threat detection by utilizing data from incidents of past attacks. This survey details the existing methodologies of automated threat Modelling applied in the integration of LLM systems, their security vulnerabilities, and ways AI is employing these systems to defend against them. It also considers issues of adversarial robustness, regulatory, and the absence of unified security standards for AI models. Subsequent studies must prioritize the development of real-time changeable security strategies, AI ethics, and all-encompassing architectures for upholding the fidelity of systems directed by AI.

Keywords: Threat Modelling, Large Language Models, NLP Security, Machine Learning, AI Security, Adversarial Attacks, Risk Assessment

1. Introduction

The large-scale adoption of Large Language Models (LLMs) has преобразованаостей artificial intelligence, being integrated across various sectors. Having conversational AI and content generation as a few of their applications, LLMs support legal analysis and automated decision making systems due to the vast datasets they are trained on. However, the development of advanced LLMs combined with high security vulnerabilities pose great risks towards data privacy, system integrity, and ethical AI usage [2]. The potential of LLMs being target of adversarial attacks, prompt injections, and data manipulation create astonishing debates on the security of deploying these models into real world applications. Threat Modelling is a strategy used to define, evaluate, and remove security weaknesses of a system. In combination with traditional approaches of cybersecurity like STRIDE and DREAD, these frameworks face challenges when put in the realm of AI models, having been used widely in software security. Unlike traditional software, LLMs process information in a predictive reasoning manner based on existing data which allows great reach of exploitation [2]. This urges a reaction,

automated threat Modeling which employs AI methods such as Natural Language Processing and Machine Learning for real time threat identification and mitigation.

The processes of applying automated threat Modelling techniques to LLMs involve the use of various AI techniques such as adversarial robustness testing, anomaly detection, and behavioral security Modelling. These techniques permit security systems to observe and infer AIs responses to their prompts, isolate harmful prompts, and defend themselves against other counter actions. As part of intrusion detection and prevention systems, NLP-based security solutions like text classification, semantic analysis, and sentiment analysis enable monitoring of user engagements in real time to detect malicious actions or breaches to the system.

Apart from that, Machine Learning models based on threat intelligence data sets can augment the arms of security frameworks by anticipating and reacting to assaults in a progressive manner [3].

The implementation of threat modeling for LLMs still presents challenges, despite advances in AI security. One major issue remains the adversarial aspect of AI attacks where the slightest change to input data can elicit an erroneous or detrimental response from a model. Also, privacy concerns emerge when LLMs are fed a large corpus of content because they can be subjected to data leakage and model inversion attacks. To overcome these issues, active adaptive security strategies, compliance policies, and safety standards across the AI industry are needed [4, 5]. The goal of this paper is to present an in-depth overview concerning automated threat Modelling in LLM system integration, emphasizing on NLP and ML security frameworks. The research analyzes existing approaches, defensive measures, risks, mitigation strategies, and subsequent studies in the field of AI cybersecurity. Focusing on the evolving landscape of threats posed to LLMs alongside artificial defenses will help fortify AI applications while maintaining ethical standards for AI integration.

2. Background & Related Work

The explosion of applications of artificial intelligence large language models (LLM) in areas such as customer service, content creation, and even automated decision making have permeated all industries. Simultaneously, dependency on LLMs has created new security challenges which need sophisticated threat Modelling to defend against adversarial attacks, data leaks, and model attacks. Traditional cybersecurity approaches like a rule-based security policy and heuristic based anomaly detection have not been successful in dealing with AI model driven systems. In this document, I discuss the basic concepts of threat Modelling, analyze the existing security policies, and evaluate recent works on AI security challenges relevant to LLM system integration.

2.1 Traditional Threat Modelling Approaches

Since the early days of software security, threat modeling has been important for organizations to identify and mitigate potential threats. Some of the traditional threat modeling frameworks include methodologies like STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privileges) and DREAD (Damage Potential, Reproducibility, Exploitability, Affected Users, Discoverability) [7]. While both of these have been influential in cybersecurity, they fail to address the constantly evolving nature of LLM vulnerabilities. STRIDE, designed by Microsoft, is systematic in that it organizes threats into six distinct groupings. It has been helpful for identifying glaring security risks in conventional software systems, but is far too rigid for anything AI-driven. Equally, the DREAD framework offers a calculation-based scoring system for assessing risks but neglects the fact that AI responses for posed questions are probabilistic; the model's behavior

can be insidiously changed by adversarial inputs. PASTA (Process for Attack Simulation and Threat Analysis) is another popular threat modeling framework. PASTA approaches threat modeling by examining the risks from the viewpoint of an attacker [8]. Its focus on continuous monitoring and evaluation of associated risks makes it more applicable in dynamic environments such as LLMs. Regardless of this adaptability, PASTA continues to depend on given attack scenarios which restrict its capability to identify new AI-specific dangers.

2.2 Limitations of Traditional Threat Modelling in AI Security

When it comes to LLM security, the major drawback with standard threat Modelling frameworks is the reliance on rigid algorithms and set threats. Unlike other forms of software, LLMs do not follow rigid logic systems. They process enormous amounts of data containing records and formulate responses based on probabilities. This unique aspect makes AI-based models vulnerable to prompt injections, adversarial attacks, model inversion attacks, and data poisoning. Adversarial attacks, for instance, consist of changing a model's inputs so that it generates biased and malicious responses. Research indicates that even slight changes in data can produce significant shifts in the results obtained from AI models, thus nullifying the effectiveness of traditional threat Modelling frameworks. The same effect is produced with prompt injections, where the prompts with the intent of influencing a model into providing set outputs. This is another security risk standard threat Modelling frameworks do not resolve adequately [9]. Data poisoning is one of the more critical threats and it involves adding false data to an LLM's training set with the intent of changing the model's responses in the future. The traditional frameworks fail to account for such an attack because they primarily focus on static software- not self-learning AI. As well, and perhaps more importantly, the attempts of attackers to retrieve confidential training data from an LLM through model inversion attacks underscore the weaknesses of classical security strategies concerned with perimeter defenses instead of the AI's inner workings.

2.3 AI-Specific Threat Modelling Approaches

When it comes to LLM security, the major drawback with standard threat Modelling frameworks is the reliance on rigid algorithms and set threats. Unlike other forms of software, LLMs do not follow rigid logic systems. They process enormous amounts of data containing records and formulate responses based on probabilities. This unique aspect makes AI-based models vulnerable to prompt injections, adversarial attacks, model inversion attacks, and data poisoning. Adversarial attacks, for instance, consist of changing a model's inputs so that it generates biased and malicious responses. Research indicates that even slight changes in data can produce significant shifts in the results obtained from AI models, thus nullifying the effectiveness of traditional threat Modelling frameworks. The same effect is produced with prompt injections, where the prompts with the intent of influencing a model into providing set outputs. This is another security risk standard threat Modelling frameworks do not resolve adequately [9]. Data poisoning is one of the more critical threats and it involves adding false data to an LLM's training set with the intent of changing the model's responses in the future. The traditional frameworks fail to account for such an attack because they primarily focus on static software- not self-learning AI. As well, and perhaps more importantly, the attempts of attackers to retrieve confidential training data from an LLM through model inversion attacks underscore the weaknesses of classical security strategies concerned with perimeter defenses instead of the AI's inner workings.

2.4 Recent Research on AI Security Threats

Integrating Language Learning Model systems has highlighted the need for hitherto advanced adaptive threat modeling frameworks as very recent research has pointed out. For instance, the work

done by [11] on adversarial perturbations in deep learning illustrates the capability of small changes to affect the intended outcome of neural networks – deep learning ninjas often get sidetracked by the silliest input-related issues! Adversarial robustness testing has become a prime requirement of AI ethical assessment and audit, thanks to the pioneering research done in this field. Later on, [11] studied the attack transferability and found that examples adversarially crafted targeting a single architecture AI model could, indeed, successfully mislead a model employing a completely different architecture. The exploit showcased the necessity of having security evaluation models across language model family integration. Further work done by [12] concerning poisoning data uncovered that attackers stood a chance of planting some hostile intended training data into AI models, jeopardizing sustained security. The need for secured curated datasets, observing the undisturbed integrity of the data provided during training in the context of LLMs posits the claim put forth by the findings. Additionally, work done by [13] on model inversion attack illustrates the possible extraction of private sensitive training data from AI models, whereby increasing the vulnerability concern on data privacy regarding the LLM subject. This trend of research has stemmed from concerns about the exploitation of data and has focused on shielding AI through the application of differential privacy, federated learning, and other privacy-preserving methods.

2.5 Integration of NLP and Machine Learning in Threat Modelling

The overlapping realms of Natural language processing and Machine Learning (ML) NLP techniques) with Threat Modelling face new challenges for AI empowered risk assessment. The LLM is vulnerable to security threats in user interaction; therefore, semantic analysis, named entity identification, and sentiment detection can be utilized to flag and block malicious intents. As an example, sentiment analysis can be hostile or manipulative attempts targeted towards exploiting LLM response generation capabilities [10]. Named entity recognition can categorize sensitive user provided information, and activate security protocols to block potential data leaks. Furthermore, ML-powered anomaly detection systems can ensure real-time monitoring of the LLM output to capture any sharp changes indicative of possible targeted undermining changes or adversarial attempts. Through the application of threat Modelling automation, the fusion of ML and NLP shifts security systems from rigid deterministic structures towards dynamic, algorithmic adaptive heuristics that are capable of learning. These systems allow real-time identification of sophisticated attacks eliminating the potential for adversarial exploitation of LLMs.

2.6 Summary of Key Insights

Shifting from Modelling traditional Methods of Threats to AI-specific security frameworks is critical for securing systems with LLM integration. Sent software frameworks might work for standard security, but lacking the needed agility for handling AI-induced risks makes them obsolete. There has been progress in adversarial robustness exploring AI-based threat Modeling, including automated anomaly search and NLP monitoring, showing great promise in mitigating emerging challenges [14]. The study aims to illustrate the need for proactive, real-time adaptive security for LLM applications. The incorporation of NLP and ML along with AI-powered threat analytics enables more efficient preemptive risk evaluation. Further work in the field should concentrate on broader resilience to adversarial attacks, improving real-time surveillance systems, and developing uniform security protocols for LLM systems.

3. Automated Threat Modelling for LLMs

The rapid advancement of Large Language Models (LLMs) has dramatically changed various aspects of the world such as automated reasoning, conversation AI, and decision support systems. However, this shift has raised security risks. Their training on massive datasets makes them susceptible to many

adversarial attacks like prompt injection, data poisoning, and model inference attacks. To cope with these problems, Automated Threat Modelling (ATM) has surfaced as a proactive framework that focuses on the identification, evaluation, and risk mitigation concerning threats on LLM - integrated systems. Unlike other threat-centric cybersecurity elements, Automated Threat Modeling implements Artificial Intelligence systems, such as Natural Language Process and Machine Learning, to actively monitor for potential threats, vulnerability analysis, and mitigate those threats to increase the model's robustness. This chapter examines the methodologies and techniques of ATM focusing on the security concerns of LLMs [15].

3.1 Need for Automated Threat Modelling in LLMs

Like many sectors, cybersecurity also has its set of best practices, and threat modelling is one of them. It includes examining the architecture of a given system and determining its attack surfaces and vulnerabilities. STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, and Elevation of Privileges) and DREAD (Damage, Reproducibility, Exploitability, Affected Users, and Discoverability) have provided advantages in perception-based methodologies like soft computing for computer systems security (software systems) [16]. These frameworks, however, do not work well with LLMs because interactions with language models and AI responses are far too complex and unpredictable. Language models do not operate as controlled software systems; they are probabilistic frameworks, thus making them vulnerable to adversarial inputs and unpredictable behavior.

Automated Threat Modelling is essential for securing LLMs because:

1. Real-time Threat Detection: LLMs process large volumes of user queries, making it crucial to identify security threats in real time rather than relying on manual security audits.
2. Dynamic Risk Assessment: Unlike traditional software systems, where threats can be predefined, LLMs are exposed to constantly evolving threats. ATM continuously updates its risk models based on newly detected attack patterns.
3. Adversarial Attack Mitigation: ATM enhances model robustness against adversarial attacks by identifying vulnerabilities before they are exploited.
4. Scalability and Automation: Since LLMs are deployed across various industries, manual security assessment is impractical. Automated Modelling ensures scalable and continuous monitoring of security risks.

3.2 Threat Vectors in LLM-Based Systems

To build effective threat models, it is essential to understand the primary attack vectors that can compromise LLM security:

1. Prompt Injection Attacks: Malicious users craft specific queries designed to manipulate the model’s behavior. This can lead to unintended information leaks, biased responses, or policy violations.
2. Data Poisoning: Attackers inject malicious training data into publicly available datasets, causing the LLM to learn and replicate harmful or biased patterns.
3. Adversarial Attacks: Small perturbations in input text can drastically alter the model’s predictions. Attackers use adversarial examples to mislead LLMs into generating incorrect or harmful outputs.

4. Model Inference Attacks: Attackers attempt to reconstruct training data from model outputs, leading to potential privacy violations and sensitive data leaks.
5. Denial of Service (DoS) Attacks: LLMs, especially when deployed in cloud-based APIs, are susceptible to overwhelming traffic that can disrupt service availability.

Each of these threats requires proactive detection and mitigation through automated security frameworks tailored to the unique operational characteristics of LLMs.

3.3 AI-Driven Techniques for Automated Threat Modelling

To effectively mitigate the aforementioned threats, automated threat Modelling for LLMs integrates AI-driven techniques, particularly NLP-based anomaly detection and ML-powered security analytics [17].

1. Anomaly Detection in LLM Interactions

Anomaly detection leverages NLP techniques to analyze incoming user queries and identify potentially harmful inputs. By comparing new queries with historical data, the system can detect anomalies that indicate adversarial intent. Key techniques include:

- Semantic Analysis: Evaluating the context and intent of user queries to detect unusual patterns.
- Outlier Detection: Identifying rare or suspicious prompts using unsupervised learning models such as Isolation Forest and One-Class SVM.
- Linguistic Fingerprinting: Recognizing adversarial attack patterns based on linguistic style and token distribution.

2. Machine Learning for Threat Identification

ML techniques are used to train security models capable of recognizing new and evolving threats. These models analyze past security incidents and classify incoming requests as benign or malicious. Common approaches include:

- Supervised Learning: Training classifiers such as Random Forest, Gradient Boosting, and Transformer-based BERT models on labeled datasets containing known security threats.
- Unsupervised Learning: Applying clustering algorithms (K-Means, DBSCAN) to detect anomalous queries that deviate from typical user interactions.
- Reinforcement Learning: Deploying self-learning security agents that adapt to evolving attack strategies by continuously refining their threat detection policies.

3. Automated Threat Intelligence & Attack Simulation

A critical component of ATM is the simulation of adversarial attacks to test the robustness of LLM security measures. By deploying AI-generated attack scenarios, organizations can assess their model's vulnerabilities and proactively implement mitigation strategies. Automated Red Teaming uses generative adversarial networks (GANs) to craft adversarial inputs and evaluate model responses [18].

3.4 Case Study: Applying Automated Threat Modelling to LLM Security

To demonstrate the practical application of ATM, consider a case study where a financial institution deploys an LLM-powered chatbot to handle customer inquiries. Without a proper security framework,

attackers could exploit the system to extract sensitive customer information through prompt injections. By integrating automated threat Modelling, the system [19]:

- Monitors real-time interactions for signs of social engineering attacks.
- Detects adversarial queries using semantic similarity matching and anomaly detection.
- Prevents unauthorized data access by implementing role-based response filtering.

Through this AI-driven security approach, the institution enhances the chatbot's resilience against data breaches, adversarial exploits, and unauthorized access attempts.

3.5 Future of Automated Threat Modelling for LLMs

As AI systems continue to evolve, the future of automated threat Modelling will focus on:

- Adaptive Security Mechanisms: Continuous learning-based models that update their threat intelligence dynamically.
- Regulatory Compliance & Ethical AI: Ensuring that LLM security aligns with global AI governance frameworks.
- Explainable AI (XAI) in Security: Enhancing transparency in threat detection models to improve trust and accountability.

Automated Threat Modelling is a crucial component of LLM security, offering a proactive and AI-driven approach to identifying, assessing, and mitigating security threats. By leveraging NLP and ML techniques, ATM enhances real-time risk detection, adversarial defense, and cybersecurity resilience. However, continuous research and development are required to address evolving threats, enhance model robustness, and establish standardized security frameworks for AI-based systems. Future advancements in AI security, regulatory policies, and adaptive threat Modelling will play a vital role in shaping the secure deployment of LLMs across various industries [20].

4. Role of NLP & Machine Learning in Security

The development of different fields of Natural Language Processing (NLP) and Machine Learning (ML) had further enhanced the automation of work and the security of systems, both of which are vital aspects of modern AI. Nowadays, practically every service, including chatbots, employs Large Language Models (LLMs). This widespread use raises various security challenges. NLP and ML are in the forefront of the defensive strategies for identifying, preventing, and minimizing the cyber threats directed at LLMs. These AI powered approaches are capable of dynamically adjusting techniques for safeguarding language models from assault, data alteration, and unauthorized exposure.

4.1 NLP-Based Security Mechanisms for LLMs

Maintaining a LLM-integrated system's security requires real-time monitoring for ill intent, which is facilitated by Natural Language Processing (NLP) technology. One major security challenge in LLM implementations is adversarial prompting, which involves modifying input prompts to change and misuse AI responses. NLP strategies like semantic scrutiny and context-sensitive filtering seek out and thwart adversarial attempts. For instance, some query-based anomaly detection models can identify querying patterns that are aimed at retrieving overly sensitive data from LLMs, which mitigates the chances of these systems being abused for model inversion or data leakage attacks [21].

Ner and Sentimental Analysis could provide evaluation of users' input for trustworthiness together with intent and tone so that the AI generated responses are ethically right and secure. According to [22], NLP also performs real-time moderation. NLP has multiple applications in security which include prompt injection detection. Prompt injection is where attackers design untrustworthy prompts with the aim of coercing the LLM their desired outputs. Such outputs could be vulgar, hate speech among others. NLP based models use Text classifying algorithms on the user queries to identify distinct patterns of malicious intent. NLP powered models that automate content moderation check responses from LLMs integrated in customer service, social networks, and commercial applications to control the dissemination of vulgar, false, or misleading information developed by or directed at genuine users and through untrustworthy systems with malicious motives. Moreover, AI based response filters are designed to block preprogrammed keywords that require AI systems to respond according to set security policies. The need for increased accuracy to intercept and prevent security risks in conversational AI heightens the need for automated language modeling and keyword spotting.

4.2 Machine Learning for LLM Security

Within the context of LLM environments, Machine Learning (ML) algorithms offer tailored adaptive learning capabilities, anomaly detection, and risk assessment logic which strengthen security architecture. Security systems dependent on past heuristics are less effective at novel, emerging threat detection. In contrast, ML-powered security systems leverage patterns recognized through real-time interactions and dynamically adapt to emerging threats [23]. One of the primary uses of ML in LLM security is anomaly detection. Security systems trained on normal, benign, and malicious interactions utilize supervised ML models for behavior deviation detection. Unsupervised clustering techniques such as K-Means and DBSCAN, as well as Autoencoders, provide advanced anomaly detection for user inputs and LLM-generated outputs. These systems are very effective against advanced persistent phishing, adversarial prompt manipulations, and other language model prompt injections. Adversarial robustness training is yet another important ML security feature. Some attackers use adversarial inputs to bias or manipulate the outputs of LLMs, often resulting in dangerous and misleading content. In ML, this type of training is known as adversarial training and consists of training LLMs on modified input samples to make them more robust to adversarial attacks. Other techniques that augment the severity of adversarial attacks against LLMs include adversarial data augmentation and security tuning through reinforcement learning. Such transformations ensure that LLMs are capable of producing dependable and secure results even when faced with hostile conditions. LLM-based services are also subjected to phishing, social engineering, and financial fraud attempts which are detected through Random Forest, Support Vector Machine, and Deep Neural Network classifiers. ML models also enhance the assessment of risk and fraud in AI systems. Supervised learning models based on a cybersecurity dataset can evaluate the user's session and mark them as suspicious or harmless depending on their activity. These user interaction networks are also evaluated by graph-based ML models which identify suspicious patterns indicative of coordinated attacks.

4.3 AI-Powered Access Control and Authentication

Control of access to LLM-Integrated systems is critical in guaranteeing technology misuse and data breach prevention. Biometric systems like behavior-based access control, and anomaly-based login credentials verification falls under ML-Powered Authentication System and strengthens security of AI driven applications. For instance, behavioral biometrics uses ML to authenticate users by checking if they fall with the typing speed, voice and interaction history of the user. Also, anomaly-based authentication models check if the user is behaving abnormally with the login and hence termed as login anomaly detection which protects accounts from take over and brute force attacks targeting LLM based services [25]. NLP contributes a lot in verifying the identity of users in LLM integrated

applications. Challenge-response protocols and text-based security questions which use natural language pose verification challenges employ NLP in validating responses. Furthermore, algorithms that recognize speech protect voice-operated AI assistants against manipulation by unauthorized individuals who may attempt to issue commands intending to alter the output of LLM.

4.4 Cybersecurity Applications of NLP and ML

Besides the security of LLMs, NLP and ML work together in their broader application to cybersecurity in areas such as cyber threat intelligence, network security, and automatic attack detection. Various types of analysis help mining of the text, sentiment, and social media forensics, which assists in preparing red flag reports, emerging threat documents, and social media intelligence reports, and also enables discussions on the dark web. ML-based intrusion detection systems (IDS) rely on classification algorithms to monitor networks for anomalies and possible cyberattacks, attempting to address these issues dynamically. Another important area is policy automation, where AI models compose adaptive security policies from compliance documents, policy documents, regulations, security best practices, and industry standards applicable to organizations. Further policy enforcement is achieved with reinforcement learning-based security optimization models which make these policies responsive to changes in the threat environment.

4.5 Future Prospects of AI-Driven Security Mechanisms

The merging of natural language processing and machine learning into cybersecurity is likely to change with the onset of explainable AI (XAI) and AI-federated learning security models. XAI improves communication so that cybersecurity professionals can make sense of the results generated by highly sophisticated automated threat analysis. In addition, federated learning allows separate organizations to participate in cultivating international intelligence and training security features in a decentralized manner, so the collective risk of losing information is mitigated. Furthermore, the adoption of AI-powered zero-trust security frameworks is on the rise. In zero-trust, systems verify and validate every user and device through continuous authentication and AI-powered anomaly recognition prior to placing access privileges on LLM-based systems. These advancements are essential for addressing AI-centric security vulnerabilities and risks while seamlessly integrating language models into numerous systems and areas of use [27]. Furthermore, the impact of NLP and ML on security provides the necessary measures to defend LLM-based systems from adversarial attacks, data poisoning, and illegal user access. These NLP techniques are also known as automatic detection of adversarial inputs, content control, and semantic analysis while the dynamic change ML frameworks use for countermeasures include anomaly detection, adversarial machine learning, and fraud control. AI-powered sophisticated access mechanisms provides even tighter control safeguarding LLM applications within perennially changing contexts [28]. As challengers of AI security continue to advance, future studies remain focusing on providing LLMs with explainable, real-time, and on-demand scaling security measures. Advancements in security frameworks will stem from the fusion of federated learning with zero-trust security models and threat mitigation based on reinforcement learning. Combining NLP and ML allows employing new organizational frameworks which guard over data, ensuring privacy while defending ethical AI practices and compliance to constantly evolving digital threats.

5. Challenges & Future Directions

The application of Large Language Models (LLMs) into real-life scenarios poses new security problems that require sophisticated methods to mitigate them. While there are efforts regarding automated threat Modelling, there still remain gaps to address such as the model's sturdiness, privacy, flexibility, and responsiveness to emerging threats. These gaps need to be filled so that AI systems

can be safely and ethically implemented [29]. One of the most important problems within securing LLMs is their exposure to adversarial attacks. Intruders have the ability to alter the model's input prompts to yield outputs that are incorrect, biased and even prejudicial. Such misleading prompt injection assaults derive from exploitation of LLM's probabilistic mechanisms which can trigger unpredictable outcomes. Worrying as well is the continued danger posed by adversarial perturbation, minor but critically perception-altering changes in the text. Defenses like adversarial training and incorporating robust embeddings are still too rigid to be effective against multiple forms of tailored threats.

Data privacy and model inversion attacks pose a great threat. Memorized data in LLMs trained on massive datasets gets sensitive information mixed up. Using model outputs enables attackers to reconstruct training data, which compromises privacy. Boston area faculty advocate for the application of differential privacy, homomorphic encryption, and secure multi-party computation, but these methods come with lowered computational efficiency and decreased performance of the model. Further enhancement in private preserving methods paired with needed responsiveness in models makes for better research focus. A lack of standardized security frameworks designed for LLM systems makes for a significant obstacle. AI as a form of technology has not optimized traditional cybersecurity methodologies such as STRIDE and DREAD. Advanced AI covers a few security protocols, but assessing LLM-specific threats still has no universally accepted standard. Security benchmarks needed for LLM deployment can be created by governance frameworks and policies regulating AI. There is an industry and government collaboration needing forming that addresses risk concerning ethics and guiding the responsible use of AI.

Another key challenge is the inability to identify and mitigate hallucinations in LLM outputs. LLMs tend to answer prompts in ways that are incorrect, but plausible and deeply misleading, leading to misinformation. Factual consistency as well as detecting lies require more sophisticated knowledge grounding and verification developments. Also, employing fact-checking models and RAG, together with RLHF, may diminish this danger.

The computational complexity of detecting threats in real-time is another challenge. There is little doubt that AI-powered threat modelling entails intensive computational demands, particularly for large-scale implementations. Older rule-based security paradigms tend to be inadequate, giving rise to the need for self-learning, ML-based anomaly detection systems. Efficient threat detection architectures using lightweight transformers, knowledge distillation, and distributed security models make for easier, real-time security monitoring without burdening the system's computational resources. Future work should emphasize adaptive, self-healing security systems for LLMs. Advances in federated learning, zero-trust frameworks, and continuous model auditing will be essential spelling emerging threats. There is also great need for interdisciplinary work in AI security, cryptography, and explainable AI (XAI) to protect the evolution of AI-driven solutions. The evolution of proactive, AI-driven cyber defense infrastructures will mark the new epoch of secure, reliable, and responsible AI systems.

6. Conclusion

The adoption of Large Language Models (LLMs) in applications like chatbots and virtual assistants has significantly impacted how AI works through the automation of content creation and decision

support services. At the same time, the security weaknesses created by LLMs and their associated technologies pose significant problems that require new threat modeling approaches. This survey focused on the application of automated threat modeling using NLP and ML to secure LLM systems. Following the investigation of existing frameworks, security risks, and their remedies, we argue against emerging AI threats that apply proactive security measures that attempt to mitigate the risks. One such risk that poses significant consequences to security is the vulnerability of LLMs to adversarial attacks that span through prompt injections, adversarial perturbations, and data poisoning that can lead to disinformation, exploitation, or biased decisions. Traditional security approaches do not account for the continuously shifting nature of AI-dominated spaces, as seen with STRIDE and DREAD. This makes the application of automated threat modeling with real-time, anomaly-driven intrusion detection systems, and adversarial resilience critical for LLM security. In combination with algorithms operating through machine learning enhancing risk evaluation, threat avoidance, and prevention, NLP-based security features like threat detection, text classification, and semantic analysis aid in distinguishing harmful inputs. Another great issue of concern in LLM security is data privacy as well as model inversion attacks. Since LLMs use large datasets, it is common for them to work on sensitive data that can be exploited. Model inversion attacks enable the attacker to extract sensitive information from the trained AI models which poses serious privacy concerns. To address these issues, AI techniques that guarantee privacy such as differential privacy, homomorphic encryption, and federated learning can be applied to the architectures of LLMs to improve their privacy. Secure AI auditing and explainability frameworks can increase the transparency and accountability of LLMs which improves responsible AI deployment. The contribution of NLP and ML in automated threat Modelling has remarkably improved AI security. NLP methods help models capture the presence of anomalous behavior, sentiment related threats, and harmful content marking, which prevents their exploitation. On the other hand, ML based security frameworks apply supervised learning for classifying attacks, unsupervised learning for anomaly detection, and reinforcement learning for dynamic response to security threats. These techniques improve the ability of the system to withstand changing threats aimed at applications that use LLMs. Even with the progress made in AI security, some issues are still outstanding. Different implementations of LLMs lack unified practices which leads to inconsistency in security coverage due to absence of standardized threat Modelling frameworks for AI systems. Moreover, continuous improvements to security mechanisms are required due to the growing sophistication of adversarial risks. Equally, AI systems do not remain free of critical issues such as ethical bias, AI's fairness, and the accountability of the decisions made by the system. For the security and robustness of LLM systems to be guaranteed, further research that focuses on adaptive security mechanisms, industrial standards, and explainable AI frameworks is needed. To address the continuous cyber threat evolution targeting LLM systems, this research highlights the need for Automated Threat Modeling. Real-time threat analysis and mitigation enabled by NLP, ML, and AI-based security frameworks allow for proactive vulnerability detection and resolution. Ongoing multi-sector collaborations focused on establishing robust policies are essential to protect AI's evolving uses. Adopting LLMs requires an integrated approach to AI security and compliance with regulatory frameworks and ethical AI development to instill reliability and trust in the AI systems.

7. References

[1.] Abusnaina, A., Wang, S., Sarker, I. H., & Kamhoua, C. (2022). Adversarial attacks and defenses on AI in cybersecurity. IEEE Transactions on Dependable and Secure Computing, 19(5), 3204-3217. https://doi.org/10.1109/TDSC.2022.3151628

[2.] Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185-5198. https://doi.org/10.18653/v1/2020.acl-main.463

[3.] Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

[4.] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Erlingsson, U. (2023). Extracting training data from large language models. Proceedings of the USENIX Security Symposium, 2633-2650.

[5.] Chen, X., He, K., Fan, Y., et al. (2021). Towards trustworthy AI: Threat Modelling and mitigation. Journal of Machine Learning Research, 22(1), 6796-6831.

[6.] Chia, Y., Weng, L., He, P., et al. (2022). On the risks of large language models serving as code assistants. NeurIPS 2022 Workshop on Security and Privacy in Machine Learning.

[7.] Goodfellow, I., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. Communications of the ACM, 61(7), 56-66. https://doi.org/10.1145/3134599

[8.] Hume, J., & Macdonald, C. (2022). Evaluating security risks in NLP-based AI systems: A survey. Artificial Intelligence Review, 55(2), 1105-1138.

[9.] Johnson, M., & Shmatikov, V. (2022). Privacy and security challenges in large-scale AI models. IEEE Transactions on Neural Networks and Learning Systems, 33(4), 1762-1775.

[10.] Kurita, K., Michel, P., & Neubig, G. (2020). Weight poisoning attacks on pre-trained models. Proceedings of the Annual Conference on Empirical Methods in Natural Language Processing (EMNLP), 7971-7984.

[11.] Li, X., & Wang, S. (2023). Threat Modelling for AI-powered applications: A systematic review. Computers & Security, 128, 102843.

[12.] Lin, X., & Liu, Y. (2021). A survey on adversarial robustness in NLP. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, 2123-2139.

[13.] Liu, Z., Yang, Y., & Xu, K. (2023). A systematic review of security risks in deep learning-based natural language processing. ACM Computing Surveys, 55(6), 129-159.

[14.] McKinney, M., & Garg, P. (2021). Automated threat Modelling for AI systems: Applications in cybersecurity. Journal of Cybersecurity, 7(1), 5-19.

[15.] Mikolov, T., Joulin, A., & Baroni, M. (2021). A roadmap towards machine intelligence. arXiv preprint arXiv:2006.06850.

[16.] Mitchell, M., & Lapuschkin, S. (2022). AI model auditing and explainability in NLP applications. Journal of Artificial Intelligence Research, 74(3), 45-69.

[17.] Papernot, N., Song, S., Mironov, I., et al. (2022). Deep learning with differential privacy. Proceedings of the ACM Conference on Computer and Communications Security (CCS), 65-80.

[18.] Peng, H., Zhang, J., & Jiang, Y. (2021). Security vulnerabilities in large-scale NLP models. IEEE Transactions on Information Forensics and Security, 17, 1023-1038.

[19.] Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(1), 5485-5523.

[20.] Rajput, S., & Sinha, R. (2023). Securing LLM systems: Threat Modelling and risk mitigation. AI & Society, 38(2), 563-578.
[21.] Ramesh, A., Pavlov, M., Goh, G., et al. (2021). Zero-shot text-to-image generation. Proceedings of the International Conference on Machine Learning (ICML), 8821-8834.
[22.] Ren, J., Lyu, C., & Li, X. (2022). NLP security in the era of large-scale AI: A comprehensive survey. ACM Computing Surveys, 54(8), 1-39.
[23.] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
[24.] Schmidt, F., Liu, J., & Jones, T. (2022). AI-driven cybersecurity: Leveraging NLP for automated threat detection. Cybersecurity Journal, 45(3), 1001-1015.
[25.] Stiennon, N., Ouyang, L., Wu, J., et al. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems (NeurIPS), 2020(33), 3008-3019.
[26.] Tramer, F., Carlini, N., Brendel, W., & Madry, A. (2023). Machine learning security: A retrospective on adversarial attacks and defenses. Journal of Machine Learning Research, 24(3), 311-339.
[27.] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998-6008.
[28.] Wallace, E., Feng, S., Kandpal, N., et al. (2020). Universal adversarial triggers for attacking and analyzing NLP models. Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2153-2165.
[29.] Zhang, M., Song, D., Chen, Y., & Liu, S. (2023). Privacy-preserving AI: A survey of techniques for protecting NLP models. IEEE Transactions on Information Security, 12(4), 567-590.
[30.] Zhu, C., Tan, Y., & Hu, R. (2022). AI security frameworks for large-scale NLP systems. Journal of Artificial Intelligence Security, 10(2), 145-167.