SMART CYBER PROTECTION CHATBOT : A SURVEY

[1]Gangambika G, [2]Sahana C Naik, [3]Samruddhi AP, [4]Shreya C, [5]Srishti Banavath M
[1]Professor, [2,3,4,5]Students
Dept of CSE
East West Institute of Technology
Bengaluru, India

Abstract— The advancement of Generative AI (Gen AI) models has undeniably stood out as a focal point in the digital transformation of 2022. As models like ChatGPT and Google Bard progress in their sophistication and capabilities, it becomes imperative to evaluate their implications within the realm of cybersecurity. Recent instances have showcased Gen AI tools being used in defensive and offensive cybersecurity contexts, shedding light on their social, ethical, and privacy implications. This research paper emphasizes the limitations, hurdles, potential risks, and opportunities associated with Gen AI in cybersecurity and privacy. Specifically, it highlights vulnerabilities in ChatGPT that could be exploited by malicious users to extract sensitive information while bypassing ethical constraints. The paper demonstrates various successful attacks on ChatGPT, such as Jailbreaks, reverse psychology, and prompt injection attacks. Furthermore, it explores how cyber offenders can leverage Gen AI tools for developing cyber-attacks, outlining scenarios where adversaries could utilize ChatGPT for social engineering, phishing, automated hacking, generating attack payloads, creating malware, and polymorphic malware. Additionally, the paper delves into defense techniques leveraging Gen AI, including cyber defense automation, threat intelligence; secure code generation, attack identification, ethical guidelines development, incidence response plans, and malware detection. It also discusses the societal, legal, and ethical implications tied to ChatGPT. Ultimately, the paper concludes by highlighting ongoing challenges and future pathways to ensure the security, safety, trustworthiness, and ethical use of Gen AI within the cybersecurity landscape.

Keywords— Accountability · Anti-Corruption · Artificial Intelligence · Brazil · Corruption · Integrity · Technology

I. INTRODUCTION

As we The chatbot employs advanced AI algorithms for instant scrutiny of security incidents, enabling swift identification and classification of potential threats. Utilizing NLP, it interprets user queries and incident reports in natural language, fostering seamless communication between security personnel and the system. With intelligent algorithms, the chatbot assesses incident severity and prioritizes them based on potential impact, allowing security teams to promptly address critical issues. Through predefined response actions for known threats, it automates tasks, reducing human error and expediting incident resolution. Leveraging machine learning, it continuously learns from past data, adapting responses to evolving threats, enhancing its effectiveness. This AI-powered cybersecurity chatbot embodies a proactive approach, fortifying digital defenses by integrating advanced AI capabilities, promoting continuous learning, and ensuring seamless integration, significantly bolstering cybersecurity effectiveness amidst today's dynamic threat landscape.

Over the past decade, the advancement of Artificial Intelligence (AI) and Machine Learning (ML) has been instrumental in driving digital transformation. These fields have seen remarkable progress, evolving from supervised learning to the more recent developments in unsupervised, semi-supervised, reinforcement, and deep learning. The latest frontier, Generative AI, utilizes deep neural networks to learn from extensive datasets, generating diverse content forms like text, images, and more. Open-AI's ChatGPT, launched in November 2022, marked a significant milestone, showcasing the potential of generative AI to a wider audience, reshaping perceptions about AI/ML. Presently, the tech industry is fervently competing to create sophisticated Large Language Models (LLMs) like Microsoft’s GPT, Google’s Bard, and Meta’s LLaMa, aiming for human-like conversation capabilities. implementation approaches. It's fascinating to observe the distinction between rule-based and AI-based models, particularly in the generative and retrieval-based paradigms. This classification sheds light on the diverse strategies employed in shaping these conversational agents.

The proliferation of Gen-AI on the internet, with ChatGPT reaching 100 million users in two months, highlights its widespread adoption. Chat-bots, serving as conversational agents in voice and text, have transitioned from rule-based to AI-driven systems, incorporating generative and retrieval methods for more precise responses. ChatGPT, a product of iterative development on Generative Pre-trained Transformer (GPT) models, combines supervised and reinforcement learning for refined interactions. These AI-driven chat-bots handle various tasks from knowledge management to customer service and possess diverse capabilities like writing, problem-solving, language translations, and more. However, their widespread use also exposes them to cybersecurity threats, with bad actors exploiting vulnerabilities to execute illicit activities, leading to data breaches and potential system exploitation. This raises critical concerns about the security and integrity of these AI-driven systems, especially ChatGPT, which has been subjected to attempts at malicious code writing, phishing, and data breaches, allowing potential exploitation of system vulnerabilities.

II. LITERATURE SURVEY

The study conducted a literature review, presenting a condensed summary of relevant research in a table format. This table includes essential information like study titles, authors, publication years, research goals, as well as notable pros and cons identified in each piece of work.

Embarking on a captivating exploration of chat-bot

evolution, I traced their path from the early days of ELIZA to the sophisticated landscape of GPT-4. Significant milestones like PARRY, Jabberwacky, A.L.I.C.E, Siri, and others revealed a rich history, showcasing the relentless innovation in chat-bot development.

Title	Authors	Year	Objectives	Advantages	Disadvantages
A Survey on AI in Cybersecurity	Muhammad Usama, et al.	2019	This extensive review article offers a thorough Examination of how AI methods are used across different areas of cybersecurity, encompassing incident response among others	In-depth exploration of AI's role in cybersecurity, delving into the possibilities presented by AI-powered chat-bots specifically in handling incident responses.	Doesn't extensively explore particular implementations of chat-bots.
Artificial Intelligence for Cybersecurity: A Comprehensive Survey	Amir Houmansadr, et al.	2021	This overview Encompasses AI uses across various facets of cybersecurity, such as identifying threats and managing incident responses	Thorough exploration of AI within cybersecurity, Offering a complete perspective on how AI contributes to managing incidents.	Provides limited information regarding the specifics of chat-bot implementations

As I categorized these chat-bots, patterns emerged, revealing insights into their varied interaction modes, application scopes, and underlying implementation approaches. It's fascinating to observe the distinction between rule-based and AI-based models, particularly in the generative and retrieval-based paradigms. This classification sheds light on the diverse strategies employed in shaping these conversational agents.

In essence, this exploration isn't just a recounting of historical events but a personal understanding of chat-bots, unraveling their complexities and capturing the essence of their trans formative journey

Title	Authors	Year	Objectives	Advantages	Disadvantages
An Overview of the State of the Art in the Use of Machine Learning Techniques in the Automated Incident Response Life cycle	Zampunieris, et al.	2018	This article investigates how machine learning is applied across different phases of the incident response process.	Centers on machine learning uses within incident response while offering details on the technical elements involved in developing chat-bots.	Restricted exploration of AI outside the realm of machine learning.
Machine Learning for Cybersecurit-y Incidents Detection and Response:A Survey	Arsh Arora, et al.	2020	This comprehensive review centers on utilizing machine learning for detection purposes	Thorough investigation into machine learning's role within incident response, providing Practical guidance on Implementing chat-bots using ML techniques.	Limited discussion on a wider range of AI methods
A comprehensive study of Artificial intelligence and Cybersecurity on Bitcoin, crypto currency and banking system	Tamanna Choithani, Asmita Chowdhury, Shriya Patel, Poojan Patel, Daxal Patel, Manan Shah	2022	It highlights challenges in AI-based cryptocurrency prediction and banking apps while proposing potential solutions, aiming to shed light on both promising areas and unresolved issues in these fields.	Examines AI's impact on cryptocurrencies and finance, analyzing prediction techniques and practical applications in banking. Identifies future research directions, enhances literature accessibility, and aids decision-making through algorithm performance evaluation.	The paper doesn't quantitatively compare AI methods' prediction accuracies, lacks systematic analysis, and focuses on specific countries. With rapidly evolving domains, its state-of-the-art info might quickly become outdated. It could benefit from integrating tech and finance perspectives while identifying ongoing research issues in more detail
The impact of artificial intelligence technique in augmentation of cybersecurity:a comprehensive review	Binny Naik, Ashir Mehta, Hiteshri Yagnik, and Manan Shah.	2021	The aim is to explore AI's role in cybersecurity by examining its applications, assessing impacts of various AI methods, discussing challenges, and showcasing its potential in strengthening defenses against threats. The objective involves conducting a comprehensive survey through literature review to understand the current landscape and potential trans formative effects of AI in cybersecurity.	The study thoroughly evaluates multiple artificial intelligence methods aimed at bolstering cybersecurity, providing valuable insights into their effectiveness and capacity to improve security protocols.	The literature review about AI in cybersecurity has constraints: it covers a wide range, lacks critical analysis of pros and cons, relies on limited published materials, lacks in-depth exploration of theoretical bases, doesn't address ethical or legal aspects, and provides no best practices or guidance for implementing AI in cybersecurity.

Title	Authors	Year	Objectives	Advantages	Disadvantages
Explainable artificial intelligence in Cybersecurity	N. Capuano, W. J. Murdoch, O. Loyola-Gonzalez, G. Vilone, L. Longo, and M. A. Ferrag	2022	The objectives are to examine Explainable AI in CyberSecurity, providing a detailed overview based on a review of 300+ papers. It aims to explore concepts, compare methods, evaluate model interpret ability, identify gaps in CyberSecurity literature, emphasize transparency in security fields, and highlight AI's risks and benefits, proposing risk mitigation strategies.	1. Offers a comprehensive understanding of Explainable Artificial Intelligence (XAI) in CyberSecurity, exploring its concepts, consequences, and applications. 2. Provides an organized overview of existing XAI approaches in CyberSecurity, comparing methods, evaluating model interpret ability, rency and proposing risk mitigation strategies.	ML-based IDS struggles with bias, outliers, and large datasets, while DL-based methods lack flow information, are vulnerable to evasion, require deep data understanding for relevant features, and suffer from a shortage of domain experts. These limitations underscore the critical need for explainable AI in CyberSecurity.
An overview of artificial intelligence ethics	Changwu Huang, Zeqi Zhang, Bifei Mao, and Xin Yao	2023	This paper aims to offer an all-encompassing view of AI ethics. It includes summarizing and analyzing ethical concerns, principles, guidelines, and approaches in AI. Additionally, it explores methods for assessing the morality of AI systems. The goal is to support future research in AI ethics for scholars and professionals interested in the field.	1. Comprehensive Overview: The work thoroughly covers AI ethics, encompassing ethical concerns, guidelines, and approaches, fostering a broad understanding of the field. 2. Research Support: By summarizing AI ethical issues and encouraging further exploration, the work aims to facilitate ongoing research in AI ethics.	1. Lack of Depth: The broad nature of the work might limit detailed analysis on specific AI ethical issues, potentially reducing depth for readers seeking in-depth insights. 2. Limited Focus: The comprehensive coverage of AI ethics might overlook specific niche areas or emerging ethical challenges within the field, limiting depth in certain aspects.
"Machine Learning for Security and the Internet of Things: the Good, the Bad, and the Ugly"	Fan Liang, William G. Hatcher, Weixian Liao, Weichao Gao, and Wei Yu.		The survey explores machine learning in cybersecurity and Cyber Physical Systems (CPS), examining its positive and negative impacts. It identifies areas for additional research to enhance efficiency, transfer-ability, audit ability, and defensive strategies against potential attacks in CPS and cybersecurity.	The survey examines machine learning in cybersecurity and Cyber Physical Systems (CPS), evaluating its positive and negative impacts The importance of developing defensive strategies against potential attacks in CPS and cybersecurity is also underscored.	Machine learning's drawbacks include uncertainties in security and privacy due to risks in data handling and the potential for adversarial attacks impacting decision-making. Vulnerabilities in ML systems create unprotected attack surfaces, challenging detection and defense against replicated user actions used for attacks.
Using AI to Augment Human Intelligence in Cybersecurity	Hayajneh, Thair, et al.	2020	This article explores the collaboration between AI and human intelligence in cybersecurity, particularly within incident response contexts.	Spotlights the collaboration between AI and human knowledge while offering a strategic viewpoint on the role of chat-bots in managing incident	.Doesn't provide in-depth technical information about implementing chat-bots.

¨Machine and Deep Learning Solutions for Intrusion Detection and Prevention in IoTs: A Survey"	P.L.S. Jayalaxmi, Rahul Saha, Gulshan Kumar, Mauro Conti, and Tai-Hoon Kim.	2022	1. Develop an IoT-focused IDPS classification using ML/DL methods. 2. Evaluate recent IDPS models' performance through ML/DL. 3. Investigate IPS prevention methods and strategies for IoT while also proposing a risk analysis method and a hybrid IDPS framework, emphasizing the significance of ML/DL techniques, comparing their viability and challenges, and outlining future research paths for ML/DL-driven IDPS in IoT.	1. Explores intrusion detection concepts and adaptable ML/DL models for chat-bot security while discussing proactive threat response techniques and risk analysis for threat prioritization. 2. Proposes a hybrid framework to balance chat-bot detection approaches, utilizes ML/DL analysis for algorithm selection, identifies research gaps, and provides foundational knowledge to innovate chat-bot security using AI/ML/DL.	The literature survey does not provide a comprehensive comparison of existing surveys in the field of intrusion detection and prevention systems. Additionally, it lacks a detailed analysis of the limitations and challenges faced by the current security models.

III. CONCLUSION

The advent of AI-driven tools like Gen-AI, such as ChatGPT and other LLM tools, has significantly impacted society. Humans have embraced these technologies openly, using them in diverse and innovative ways, from crafting images to composing music and generating text. Their influence spans nearly every domain, including cybersecurity, where Gen-AI, particularly ChatGPT, has significantly affected how organizations approach their cybersecurity strategies. This paper aims to methodically investigate and present the challenges, limitations, and opportunities that Gen-AI presents within the cybersecurity landscape. Initially, using ChatGPT as the primary focus, the paper showcases how it could be vulnerable to attacks aimed at bypassing its ethical and privacy safeguards through reverse psychology and jailbreak techniques. Subsequently, the article explores various cyber-attacks that could be devised and unleashed using ChatGPT, demonstrating the use of Gen-AI in cyber offenses. It then delves into experiments involving different cyber defense mechanisms supported by ChatGPT, followed by a discussion on the social, legal, and ethical concerns related to Gen-AI. Additionally, the paper highlights the distinguishing features between two prominent LLM tools, ChatGPT and Google Bard, showcasing their respective capabilities in the realm of cybersecurity. Finally, it outlines numerous ongoing challenges and research problems specific to cybersecurity and the performance of Gen-AI tools. The aim is to inspire further research and the development of innovative approaches to harness the potential of Gen-AI within the cybersecurity domain.

REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. Communications of the ACM, 63(11):139–144, 2020.

[2] Generative AI – What is it and How Does it Work? https://www.nvidia.com/enus/glossary/data science/generative-ai/. (Accessed on 06/26/2023).

[3] Open AI. Introducing ChatGPT. https://openai.com/blog/ chatgpt, 2023. Accessed: 2023-05-26.

[4] Do ChatGPT and Other AI Chat-bots Pose a Cybersecurity Risk?: An Exploratory Study: Social Sciences & Humanities Journal Article. https://www.igi-global.com/article/ do chat-GPT and-other-ai-chatbots-pose-a- cybersecurity-risk/ 320225. (Accessed on 06/26/2023).

[5] Models - Open AI API. https://platform.openai.com/docs/ models. (Accessed on 06/26/2023). [6] Google Bard. https://bard.google.com/. (Accessed on 06/26/2023).

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie- Anne Lachaux, Timothee Lacroix, Baptiste Rozi ´ ere, Naman ` Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[7] Number of ChatGPT Users (2023). https://explodingtopics.com/ blog/chatgpt-users. (Accessed on 06/26/2023)..

[8] A History of Generative AI: From GAN to GPT4. https://www.marktechpost.com/2023/03/21/ a-history-of-generative-ai-from- gan-to-gpt-4/. (Accessed on 06/27/2023).