Applied Sciences, Computing and Energy, 2024, 1(1), 140-156

A Review of Applying AI for Cybersecurity: Opportunities, Risks, and Mitigation Strategies

Ogochukwu Susan Ndibe[*] and Precious Ogechi Ufomba[2].
Received: 12 July 2024/Accepted: 18 December 2024/Published :31 December 2024

Abstract:*The issue with the evolving rapid development of complicated cyber threats has encouraged organizations to implement Artificial Intelligence (AI) and large language models (LLMs) as the revolutionary characteristics of contemporary cybersecurity development. These systems, through the use of machine learning, natural language processing and predictive analytics are able to perform automated code reviews, anomaly detection in real time, AI-based vulnerability assessments, intelligent analysis of threat intelligence. The potential of AI to handle huge amounts of information assists organizations in becoming even more proactive in detecting weaknesses, shortening the time spent responding to incidence, and even becoming more resilient. But at the same time, AI stands to pose a dual-use problem, encompassing such issues as adversarial attacks, insecure AI-generated code, and automated phish campaigns. This paper looks into mitigation measures including the human-in-the-loop systems, adversarial techniques, and governance frameworks like the NIST AI Risk Management Framework that maintain a balance between innovativeness and ethical governance. The paper concludes that the introduction of AI can vastly enhance cybersecurity even when carried out more judiciously and reinforced with robust governance that does not present an unmanageable.*

Keywords: *Artificial intelligence, cybersecurity, large language models, adversarial attacks, anomaly detection, governance, human-in-the-loop*

Ogochukwu Susan Ndibe
Cybersecurity & Information Assurance, University of Central Missouri, Warrensburg, Missouri, United States
Email: sndibe7@gmail.com
Orcid id: https://orcid.org/0009-0004-1133-7036

Precious Ogechi Ufomba
Cybersecurity, Katz School of Science and Health, Yeshiva University, New York, United States
Email: ufombapreciousoge@gmail.com
Orcid id: https://orcid.org/0009-0009-7932-0034

1.0 Introduction

Cybersecurity is one of the pillars of operational resilience and trust by governments, companies, and individuals in the age of digital transformation. The cloud computing and mobile technologies, as well as IoT, made the digital namescape far more extensive (Conti *et al*., 2018: Ademilua, 2021). As of 2023 to date, 15 billion connected IoT gadgets were estimated to be globally, each being a possible entry point of the cybercriminals (Arnold, 2023). These tendencies have been accompanied by an explosive rise in the number of advanced threats, which include ransomware-as-a-service, supply chain defeat, and nation-state-sponsored cyber espionage operations against critical infrastructure and personal data (Rahman *et al*., 2022). These vulnerabilities have been further augmented by the move towards remote and hybrid working that was compounded by the COVID-19 pandemic, which has decentralised security control and created more reliance on potentially insecure networks and endpoints (Rahman *et al*., 2022).

The existing rule-based and signature-based security mechanisms, although fundamental systems, have been unable to cope with the fast-changing threat environment as well as the volume of data produced by the contemporary networks (Buczak & Guven, 2015). This shortcoming has propelled the increased use of Artificial Intelligence (AI) and machine learning (ML) approaches to cybersecurity with the hope of more dynamic, data-centered defenses. The development of AI in recent years (transformation of primitive expert systems to current deep learning architectures) has provided the security practitioner with intrusion detection automation, detecting anomalies, classification of malware, and predictive threat assessments (Apruzzese *et al*., 2023; Arnold, 2023). These Artificial Intelligence systems are more efficient in identifying more complicated, hidden patterns in highly diverse and high-dimensional data, facilitating quicker and more efficient identification of emerging threats.

The development of large language models (LLMs), including Gemini, GPT-3, GPT-4 and the code-based models like Codex (Brown *et al*., 2020; Chen *et al*., 2021) can be considered one of the most revolutionary shifts in AI that have occurred in the recent past. Transformer-based and other models can produce natural language with virtually human fluency and can be used in threat intelligence analysis in real-time, automated documentation, vulnerability scanning, and code review (Chen *et al*., 2021). Citing an example, LLMs have the potential to analyze big unstructured data of threats and summarize reports as well as even create playbooks of incident response (Shayegani *et al*., 2023). Nevertheless, the later models face substantial dual-use threats: a malicious user can use them to automatically create effective phishing messages, make malware scripts, or automate social engineering activities (Brundage *et al*., 2018; Shayegani *et al*., 2023). This contradiction highlights the issue of innovation and security when applying AI. Understanding the potential of the opportunities and challenges of AI in cybersecurity, the following three goals will be fulfilled in this paper:

(a) to critically discuss the present and future use of AI, including LLMs in the improvement of cybersecurity practices;

(b) to assess the emerging categories of risks and weaknesses brought about by the misuse of or active attack on AI technologies.

(c) to suggest feasible mitigation measures, ethics, and policy suggestions to aid the responsive implementation of AI in the cybersecurity realms (Brundage *et al*., 2018; Sarker *et al*., 2020).

The importance of the research is based on the holistic approach: this work addresses AI not only as a defense tool, but also as a possible threat channel. On the one hand, AI-based solutions will be more efficient, can scale and are expected to provide higher detection rates compared to the current solution, though it also comes with challenges and risks, such as bias, adversarial attacks, and the possibility of allowing more malicious forms of cybercriminal activity (Arnold, 2023). Since AI is transforming the next generation of cybersecurity products, it is important to learn the cyber limitations and create resilient governance mechanisms so that the utilization of AI does not create insecure and unethical outcomes of its implementation.

A narrative literature review approach is used in this project, and the relevant peer-reviewed scholarly articles, technical reports, and industry white papers were reviewed. The information was obtained in major databases and repositories such as IEEE Xplore, ScienceDirect, Scopus, and arXiv, covering both abstract theories as well as up to date empirical research. The following scope is the AI uses in threat detection, network security, incident response, AI misuse risks (adversarial machine learning, social engineering) and, mitigating techniques (explainable AI, adversarial training, and regulations)

(Apruzzese *et al*., 2022; Sarker *et al*., 2020). The paper will incorporate these interdisciplinary ideas and contribute to the scholarly discourse, as well as offer practical recommendations to both cybersecurity practitioners and policymakers that operate in the wake of the current AI developments.

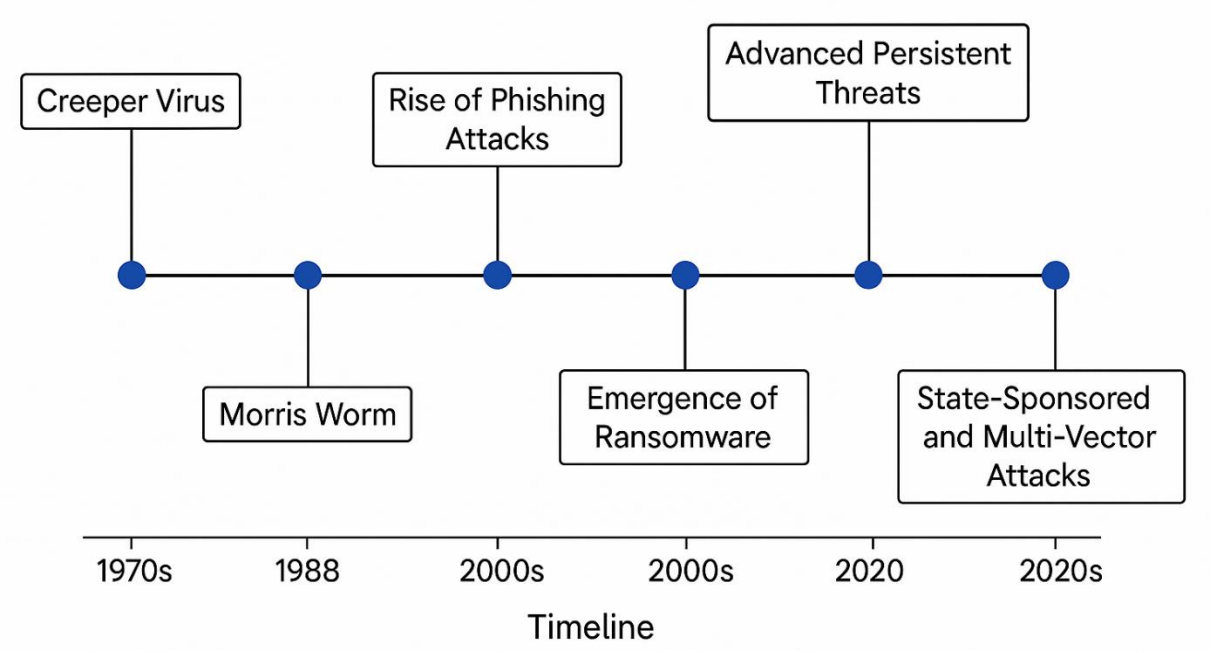

Fig 1: Timeline of Cyber Threat Evolution (After Mbah & Achudume, 2024)

2.0 Understanding the intersection of AI and cybersecurity

2.1 How AI transforms traditional cybersecurity approaches

Over the last ten years, Artificial Intelligence (AI) has drastically transformed cybersecurity, whereby it is no longer reactive in many instances, but predictive and adaptive (Tao *et al*., 2021). Traditional systems were mainly based on stationary controls, known attack signatures, and manual interpretation of security experts (Tajrian *et al*., 2023). Such techniques were usually effective only against known threats and had trouble with new attack vectors and zero-day exploits. In contrast to that, AI systems can learn the complex patterns and specifics of current threats automatically through understanding high-dimensional and large datasets including system logs, network traffic, and user behavior (Abdelhafez *et al*., 2023). It has been shown in recent research that AI-based intrusion detection systems (IDS) reaction to complex, multi-stage attacks is better than non-AI only signature-based systems (Zhang *et al*., 2023). Also, AI-enabled platforms aid such capabilities as predictive analytics, allowing security teams to focus on the most probable vulnerabilities to be exploited (Tao *et al*., 2021). The transformation has contributed to the minimization of the time attackers take before being caught, response speed, and overall IT infrastructure resilience. Security Orchestration, Automation, and Response (SOAR) platforms are also based on AI and can perform tasks that require repetition, allowing human analysts to work on complex investigations (Srivastava *et al*., 2022). Nevertheless, researchers warn that AI must not substitute but preferably augment human knowledge, which can be addressed through

such challenges as data bias and adversarial attacks (Han *et al*., 2023).

2.2 Overview of AI techniques used in cybersecurity: machine learning, deep learning, NLP, and LLMs

Cybersecurity AI has several parallel methods that have different sources of expertise. Some commonly used algorithms to detect anomalies and classify phishing or spam emails are the decision trees, k-nearest neighbors (k-NN), and support vector computers (SVM) of Machine Learning (ML) (Dunmore *et al*., 2023). As an example, supervised ML can learn on labeled data to distinguish benign and malicious traffic, supervised ML models can operate on labeled data and learn how to identify benign traffic and malicious traffic, and unsupervised methods identify anomalies in unlabeled data (Thangapandian, 2022).

Deep Learning (DL) builds upon these through the use of neural networks neural networks such as Convolutional Neural Networks (CNNs), in detecting malware using images, and Recurrent Neural Networks (RNNs), in detecting malware in malicious log files and packets flows (Thangapandian, 2022). DL models perform well in extracting features in raw data with little effort required in feature engineering as one needs to do it manually.

NLP is also proving to be more important. NLP methods can be applied to unstructured data sources like threat intelligence feeds, vulnerability reporting, and social media reports in order to identify actionable intelligence in those data sources (Zhou *et al*., 2020). Topic modeling and named entity recognition, which enables an analyst to discover new threats more quickly.

Recently, certain Large Language Models (LLM) such as GPT-4 and BERT were also explored to do security-oriented code review, summarize incident reports, even write security policies (Xiang *et al*., 2023). The time spent on the analysis of vast amounts of documentation can also be decreased, and to a considerable extent, by use of LLMs, but their results need to be checked by experts (Zheng *et al*., 2023).

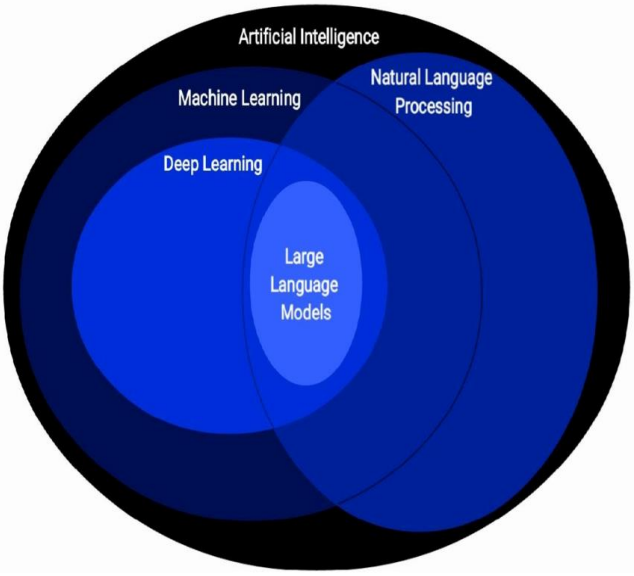

Fig 2: The intersections of AI, ML, NLP, DL, and LLMs (after Vavekananda *et al*., 2024)

2.3 Strengths and limitations of AI models in the security context

AI has an incredible payoff or advantage, such as scalability, identifying threats in real-time, as well as revealing the various patterns of the advanced attacks (Ahmad *et al*., 2023). As an example, AI systems can track millions of endpoints at a given time and associate seemingly unrelated incidents, which is not a task that a human reviewer can achieve by his/her own means (Mohamed, 2023). AI also assists in detecting insider threats as there is a chance that a slight change in user behavior can go unnoticed in conventional rules (Safdarian *et al*., 2023).

But it is not AI that will be a silver bullet. It relies very much on the quality, variety, and quantity of the training data (Mohamed, 2023). In case of training data which does not contain some exemplars of attack, the AI models can perform worse against same attacks. Explainability is another difficulty: most AI models, particularly deep neural networks, are what is called a black box: it is hard to explain

the reason behind a given decision (Rudin *et al*., 2022). Such obscurity may cause a lack of confidence and regulatory adherence, particularly in high protein settings such as the critical infrastructure. Moreover, AI models may generate either false positives that bombard security analysts or false negatives that result in allowing well-crafted attacks to move past undetected (Tao *et al*., 2021). These limitations can be overcome by integration of the AI with the human expertise and active monitoring.

3.0 Applications of AI in Cybersecurity

The field of cybersecurity has changed with the incorporation of Artificial Intelligence (AI) which introduces new tools and frameworks that can process large amounts of data, find complex patterns, and automate numerous parts of the defense and detection systems. Such technologies as large language models (LLMs) and deep learning could also be regarded as the most revolutionary of the past few years, proposing new efficiencies and possibilities in the whole process of software development, vulnerability detection, monitoring in real-time, and threat analysis (Chen *et al*., 2021; Zheng *et al*., 2022). In this part, the practical application of AI will be discussed, showing the technical possibilities, as well as its operational influence.

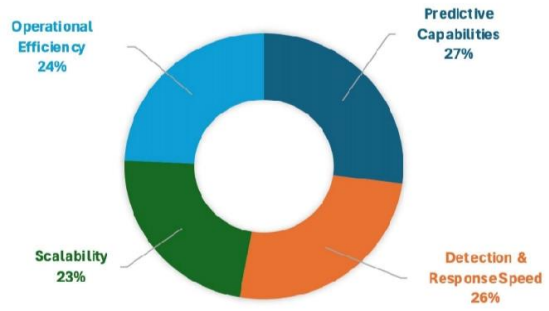

Fig 3: Impact of AI deployments in cybersecurity (after, Roshanaei *et al*., 2024)

3.1 Automated security-focused code reviews using large language models (LLMs)

Automation of the traditionally human and time - consuming process of code reviews is one of the most effective uses of AI in cybersecurity. Tooling Large language models (LLMs) like those produced by OpenAI including Codex and GPT-4 have significantly outperformed in code analysis to recognize typical vulnerabilities in code snippets, and propose alternative snippets with secure code (Chen *et al*., 2021). These models can see patterns that may not exist in the data a human reviewer could see because they learn over large datasets that contain secure code as well as insecure code (Spero, 2022). Popularization of LLMs in continuous integration and deployment (CI/CD) pipelines also means that the costs of remediating a potential security problem are minimized, as code will not have been shipped to production by the time a problem is identified. These tools have already found application in practice: for example, DeepCode and GitHub Copilot offer security-wise feedback in real-time when developers write code (Microsoft, 2023). These tools cannot be a complete substitute to expert human auditors, but they supplement the work as they identify more issues more quickly, and give developers feedback without delay. The studies indicate that such systems enhance good secure coding and decrease exposure time on exploitable vulnerabilities (Spero, 2022). Nonetheless, these models will need to be carefully evaluated in order to avoid false positives or propose insecure patterns, which is to remind that human supervision is still needed.

3.2 Assessing and auditing AI-generated IT guidance for security gaps and vulnerabilities

With the strong possibility of AI tools being used more widely to produce IT architecture documentation, deployment instructions and even configuration templates, a new problematic area has arrived, the security of the AI-generated information content itself. Where models are not driven by context or trained on

a biased set of data sets with insecure patterns, automated documentation and infrastructure-as-code (IaC) scripts may also contain security flaws (Brundage *et al*., 2018). To take a concrete example, the AI generated deployment script could call to configure over-permissive network access, apply older encryption mechanisms, or incorporate default passwords: these are all vulnerabilities that, having been deployed would severely jeopardise the security of the implementation (Gordon *et al*., 2023).

Table 1: Comparison of traditional and AI-based risk management (after Mbah & Achudume, 2024)

Aspect	Traditional risk management	AI-based risk management
Detection	Rule-based, limited to known threats	Anomaly-based, capable of identifying unknown threats
Response Time	Reactive, often delayed	Proactive, real-time responses
Scalability	Limited to specific environments	Scalable across complex, multi-layered systems
Adaptability	Static rules require manual updates	Dynamic learning adapts to evolving threats
Human Effort	High reliance on manual processes	Automated, reducing human intervention
Accuracy	High false positive and negative rates	Enhanced accuracy through pattern recognition

In order to deal with that organizations are increasingly turning towards the use of AI to perform two tasks: not only to come up with an IT guidance but also to audit it. Rule-based post-processing systems and dedicated models scan generated scripts and documentation done with AI, and look for standard misconducts and risk (Zheng *et al*., 2023). In the cloud-native landscapes in particular, where infrastructure is changing at a high pace, the two-layered approach is particularly useful, as the manual inspection is impossible (Brundage *et al*., 2018). The use of continuous auditing can make sure that the AI-driven automation does not jeopardize security inertedly. In addition, the practice is consistent with the wider trends of DevSecOps, which involves integrating security checks into the lifecycle of the development. Even though the field in question remains relatively young, initial research and case evidence indicate that it has a considerable capacity to minimize misconfiguration-related risks and increase security by design (Gordon *et al*., 2023).

3.3 AI-driven vulnerability assessment and penetration testing tools

The conventional vulnerability testing is performed based on known signatures and expects the use of internal or external rules that are inactive and can be effective only against malware that is known. In comparison, vulnerability assessment programs operating on AI make use of the tools of machine learning algorithms and deep learning trained on large amounts of actual exploits, commonly identified vulnerabilities (Apruzzese *et al*., 2023; Abolade & Zhao, 2024). Such tools interpret application-level code, APIs, and network infrastructures on a dynamic basis to detect new vulnerabilities that have not been understood before by drawing parallels between the behavior of such vulnerabilities and well-known exploits (Vadisetty *et al*., 2023). Penetration testing tools run with AI

take it further and implement an attack scenario and adjust to responses by the system. Such tools as Pentera do the work of discovering useable paths, privilege updgrades, and lateral movement techniques, a realistic view of the organization security position (Komaragiri & Edwards, 2022). This method assists an organization to prioritize on remediation in terms of its actual exploitability and not theoretical risk. The new data consists of new threat information, which AI models constantly update them to perform better. Notably, these tools supplement, but do not substitute the expertise of humans: scanning is still faster when aided by AI and complex patterns are detected, but expert pentesters can also add context, creativity, and perspective that would never fall under the model. The blend enhances test coverage, minimises the time in carrying out assessments as well as fortifies organizational safeguards against emerging threats.

3.4 Threat intelligence, anomaly detection, and real-time monitoring

Real time security monitoring is impossible to do manually due to the complexity in the monitoring exercise because of the huge amount of data that is to be monitored at any given time. AI comes in and makes this effortless. To find anomalies, machine learning models learn the normal pattern of user, device, and network usage and then apply it to find deviations that can suggest compromise (Buczak & Guven, 2016; Abolade, 2023). As an example, AI has the capability of detecting credential abuse, abnormal data transfer, or side-lateral movement as it can be overlooked by static rules (Reddy, 2023). Autoencoders and recurrent neural networks (RNNs) are forms of deep learning that work with temporal dependencies and improves on the identification of small-scale attack patterns (Zheng *et al*., 2023; Okolo, 2023). In addition to detection, AI enhances threat intelligence by taking data provided by several sources, which include open source feeds, dark web monitoring, and proprietary data and matching them to one another in order to unravel new campaigns (Sarker *et al*., 2022). This allows protective measures and quick reaction to emerging threats. AI can also significantly contribute to alert triage, removing noise and prioritizing alerts of high priority, thereby minimizing the fatigue experienced by the analyst (Buczak & Guven, 2016). These are features that have already been incorporated in the security information and event management (SIEM) and security orchestration, automation, and response (SOAR) systems that offer major benefits to the response time of incidents. AI, in cooperation, makes security operations centers (SOCs) more effective, more scalable, and more flexible to face more complex threats.

3.5 Natural language processing (NLP) for analyzing logs, alerts, and documentation

High quantities of unstructured data should be processed by security departments, such as incident reports, threat alerts, and compliance reports as well as system logs. NLP models enable this information to be converted into workable intelligence with the help of AI. As an example, NLP could be used to derive indicators of compromise (IoCs) out of reports, match alerts across systems, and summarize long-documentation to quickly analyze it (Brown *et al*., 2020; Chen *et al*., 2021). Large language models also optimize this process by comprehending stories with difficult contexts, drawing causality, and recommending course of actions. This saves the time they took to manually analyse data as well as make faster decisions based on data. The NLP models are also capable of identifying trends in past incident information and how organizations can be in a better position to know about the weak nesses that keep happening and enhance their security measures (Zheng *et al*., 2023). In addition to analysis, LLMs can aid in documentation, writing incident summaries, compliance reports, and that should be realized by the user (Microsoft, 2023). These

competencies can assist in filling technological security data-operation decision gap, by making such cybersecurity practices more approachable and functionally endeavored to the overall functioning of an organization.

Table 2: Applications of AI in cybersecurity

Application	Purpose	AI models / methods	Practical tools & examples	Benefits	Limitations / risks	Authors
Automated security-focused code reviews	Detect insecure code patterns and common vulnerabilities	LLMs trained on code; transformer architectures	GitHub Copilot, DeepCode, Codex	Faster detection, supports secure coding during development	False positives; lacks full project context	Chen *et al*. (2021); Microsoft (2023)
Assessing AI-generated IT guidance	Audit AI-created documentation and IaC templates	Dual LLM pipelines; rule-based scanning	Custom linters, CI/CD checks	Prevents insecure defaults and misconfigurations	Models may inherit biases; risk of missing edge cases	Brundage *et al*. (2018); Zheng *et al*. (2023)
AI-driven vulnerability assessment & pentesting	Discover known & unknown vulnerabilities; simulate attacks	Supervised ML; deep learning (CNN, RNN); reinforcement learning	Pentera; AI-based scanners	Scalable assessments, discovery of zero-days	Data dependence; false positives	Apruzzese *et al*. (2023); Zheng *et al*. (2023); Ademilua & Areghan (2022).
Threat intelligence & anomaly detection	Detect abnormal patterns; correlate threat feeds	Autoencoders; clustering; RNNs	Google Chronicle, IBM QRadar AI	Early detection, reduces alert fatigue	Model drift; adversarial evasion	Buczak & Guven (2016); Sarker *et al*. (2022); Zheng *et al*. (2023)

NLP for logs, alerts & documentation	Extract IoCs, summarize incidents, draft reports	NLP models; transformer-based LLMs	Microsoft Security Copilot; Elastic Security NLP	Faster triage, structured data extraction	Language bias; false negatives	Brown *et al*. (2020); Chen *et al*. (2021)

4.0 Risks and malicious uses of AI in cybersecurity

The adoption of AI into contemporary cybersecurity systems has not only a very profound set of defensive gains but also a very profound set of new risks. The dual use aspect of AI implies that there are potentially increased cyberattacks as the technologies with the potential to improve detection and automate may also be military. Moreover, AI models are susceptible to malicious manipulation as well. This section reviews three fundamental capabilities where AI empowers attackers and raises new attack surfaces (1) AI-written programs and IT advice as vectors in attack (2) AI-facilitated social engineering via phishing and deepfake, and (3) adversarial targeted attacks of AI systems, such as poisoning, evasion, and model extraction.

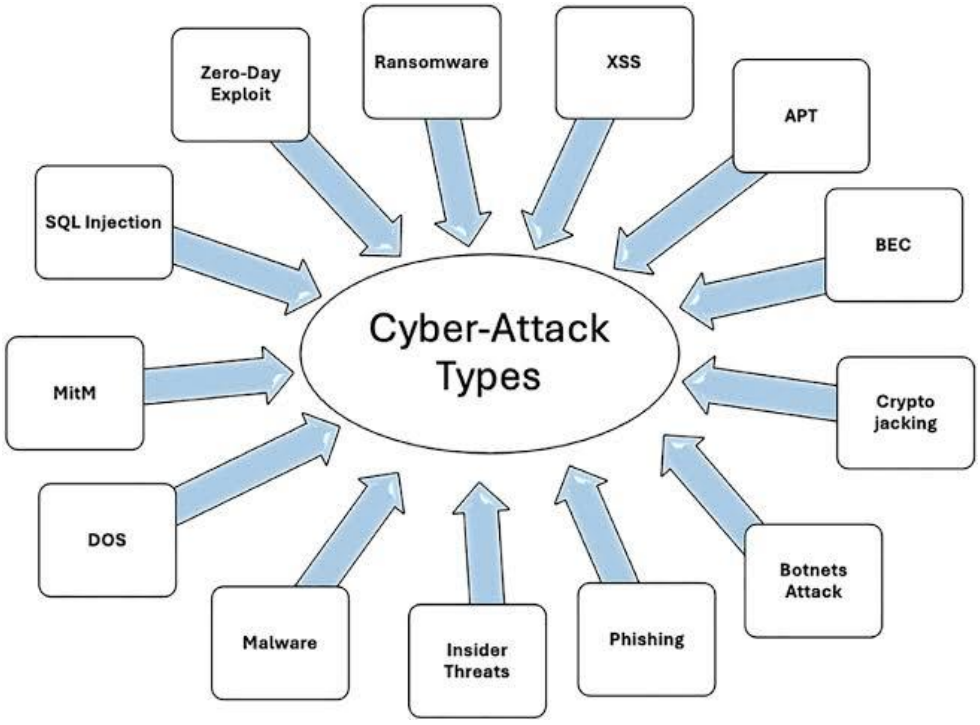

Fig 4: Different types of cyber attacks (After Salem *et al*., 2024)

4.1 AI-generated code and IT guidance as attack vectors

With the recent emergence of large language models (LLMs) such as GPT-4, Codex, and CodeBERT, developers and IT teams have a considerable assistant that automates the process of code generation, script writing, and infrastructure provisioning (Ambati, 2023).

Nevertheless, they bring insider threats. These models are prone to one of the most significant

risks, which is developing an insecure code because of missing details in its training dataset or a failure to sense the context (Bukhari *et al*., 2023). As an example, an LLM may create functions that do not validate the input in an effective manner, deploy old cryptographic algorithm or improperly configure access controls, all of which could be turned into vulnerabilities (Krishnamurthy,, 2023).

Outside of the code, misconfigurations in the form of open firewall ports or over-granted IAM roles can be introduced in the AI-generated IT guidance and deployment scripts and accidentally allow an organization to have a larger attack surface (Ambati, 2023). The configurations generated by AIs may be directly targeted by the attackers, or with lesser obviousness, training data may be manipulated to bias the model toward unsecure results, a sample of data poisoning attack along the supply chain of AI-generated IT resources (Bukhari *et al*., 2023; Dada *et al*., 2024).

The scholars suggest combining the outputs of AI models with static analysis tools, human expert reviews, and frameworks explaining them to mitigate these risks (Rudin *et al*., 2022). Nonetheless, with the growing rate of adoption of the LLM, hackers might start to exploit what they can perceive as a blind trust of the AI-generated artifacts.

4.2 Automated phishing, deepfakes, and social engineering powered by AI

One of the most reliable strategies to conduct a cyber attack are social engineering and with implementation of AI, such efforts are changing both in magnitude and complexity. Generative AI has the potential to create highly convincing phishing emails personalized to a certain target by evaluating public information, past trespasses, and the organizational environment (Brundage *et al*., 2022). In comparison with the established phishing templates, AIG-generated messages can be fluent in natural language and adjust to the tone of culture or organization, avoid a large number of spam filters, taught on the patterns of an earlier phishing period (Manyam, 2022).

As an example, AI may create context-specific emails which will seem sent by an executive mentioning recent projects or company files to sound more authoritative (Syed *et al*., 2022). Barbarian spear phishing is so sophisticated to the point where some advanced tools automate certain variants and generate thousands of unique and personalized messages within an hour, which could not be done before due to the lack of customization (Zheng *et al*., 2023).

In addition to conveying the text, deepfake technology, based on deep learning, enables attackers to develop rather convincing artificial video images, as well as audio recordings of the voice. They may be applied during fraud, extortion, or the impersonation of executives during live-stream video (Syed, 2022). Also, assets of the organization identified through the reconnaissance using AI-powered tools to leverage organizational hierarchies and gather information through social media generate the accuracy of attacks (Garg, 2023). The combination of AI is to take social engineering away as an artisanal affair and make it a scalable, automated process this time around, which is why defenders need to choose sophisticated detection strategies, employee education, and content verification in real-time.

4.3 Adversarial attacks on AI models: data poisoning, evasion, and model extraction

Even the AI systems are sale targets of advanced attackers. Adversarial attacks especially in three categories namely data poisoning, evasion, and model extraction are posing fundamental threats to reliability and trustworthiness of AI.

Data poisoning: Adversaries will inject adversarial designed samples into the training data to slightly skew the model behavior. As an illustration, it is possible to introduce mislabeled malicious data as a benign label when training IDS, thus leaving it with a blind spot (Ramirez *et al*., 2022). This can be performed upstream e.g. poisoning open-

source datasets utilized by defenders. The outcome is an underperforming model in the accurate location of the intended attacker, with subsequent reduction of the security among others and thus without prompt deterrence (Rahman *et al*., 2023).

Evasion attacks: Attackers can develop inputs at inference time that make some predictions that seem normal to humans but actually lead to erroneous predictions on the AI. In malware detection, minor changes in the binary or obfuscations can deceitfully mislead the machine intelligence to identify malicious codes as benign (Ramirez *et al*., 2022). On the same note, the network traffic created with the adversarial networks can slip the surveillance by encoding noise or minute protocol skewers.

Model extraction: Model stealing, as it is sometimes called, is done by having the attacker query the model deployed, in this case an API-based malware classifier, and trying to approximate its parameters and boosters (Lin, *et al*., 2021). At some point, the replica can be extracted and used in order to learn weaknesses, create optimized attacks, or generate the proprietary model itself. This endangers the intellectual property and arguably can directly facilitate adapted adversarial attacks.

The methods emphasize the following paradox: although AI enhances cybersecurity, it also adds complexity and forms new attack surfaces (Ambati, 2023). Some of the defense strategies are adversarial training (training the models with examples of adversaries), anomaly detection around model behaviour, and differential privacy and explainability which are used to detect abnormal model decisions (Rudin *et al*., 2022). Nevertheless, it will be an endless cat-and-mouse game as adversarial AI is dynamic.

5.0 Prevention, mitigation, and policy strategies

Coupling cybersecurity with the artificial intelligence (AI) has both indisputable defensive benefits and requires a set of well-designed protection mechanisms. To counter the dual-use risks of AI, organizations should integrate both technical countermeasures, human control and governance systems. The strategies discussed in this section are policy-oriented and pragmatic: safe integration of AI tools into cybersecurity pipelines, the construction of human-in-the-loop validation systems, designing adversarial, the establishment of regulatory and ethical guardrails, and encouraging the workforce education in an attempt to ease human vulnerabilities.

5.1 Best practices for secure integration of LLMs and AI tools into cybersecurity pipelines

Increasing usage of large language models (LLM) and AI in tasks that require high emphasis on security, including vulnerability scanning, automatic patching, and code review, necessitate secure engineering practices to prevent arising risks. Good input and output validation of AI-generated scripts and recommendations should be enforced within organizations to evade unsafe code-related texts or misconfigurations in making it to the production environments (Vadisetty *et al*., 2023; Zheng *et al*., 2023). Another way to minimize harmful suggestions is by fine-tuning LLMs on specialized, verified security corpus (Tao *et al*., 2021) and they also can harmonize output with internal policies. Also, the professionals suggest applying context constraints, including the methods of AI minimization to predefined contexts and the restricted scope of tasks (Vadisetty *et al*., 2022). In this way, access escalation is avoided inadvertently. Monitoring and explainability such as SHAP or LIME (Rudin *et al*., 2022) allow analysts to monitor AI outputs and research unusual AI decisions, thus, security is guaranteed, and AI deployment is auditable. All these practices guarantee that AI is added to security processes without the incurrence of uncontrolled weaknesses.

5.2 Adversarial training and defense strategies for AI models

Adversarial vulnerabilities like data poisoning and evasion, and model extraction attack the statistical characteristics of the AI models and compromise the accuracy or confidentiality (Sarker, 2023). In order to protect against such threats, security researchers promote a number of supplementary methods. The method of retraining (adversarial training) the models using adversarial examples renders AI systems more robust to adversarial attacks (Kaviani *et al*., 2022). Sanitizing the data sanitizes the data before it affects the model by filtering it with suspicious or wrongly labeled samples (Alotaibi & Rassam, 2023). Moreover, architectures that are provably robust against some evasion attacks--for example defensive distillation or learned ensemble learning lowers the vulnerability of models to such attacks by using small perturbation to learn. The behavior within models can also be monitored and an anomaly may be sign of adversarial intervention (Kaviani *et al*., 2022). Organizations can differentially privatize data, perturb output and restrict access to APIs to prevent issues with model extraction (Sarker, 2023). No particular defense can prevent an attack, but a combination of all these measures would increase the complexity and expense of attacks greatly.

5.3 Regulatory and ethical considerations

With the pace of adoption of AI, it is important to have governance framework and regulatory standards that will help to temper innovation with security, fairness, and accountability. The 2023 NIST AI Risk Management Framework (AI RMF) presents advice and guides to managers, developers, and operators of AI to identify, measure, and mitigate the risks in the lifecycle of development and deployment (Sarker, 2023). It focuses on openness, frequent assessment and documented procedures as a way of showing compliance. Otherwise, OECD AI Principles and EU AI Act that will be adopted in the European Union foster human control, explainability, and risk management in proportion (Mohamed, 2023). The organizations need to develop governance frameworks that set accountability and identified roles of the AI system owners, developers, and security teams. Audit mechanisms should trace the decisions made with the participation of AI and evaluate both the technical and ethical effects (Rudin *et al*., 2022). In this way, the technique of governing AI makes it an asset to the business that complies with all laws and regulations applicable in society.

Table: Ethical considerations and data privacy in AI-enhanced cybersecurity (After Mbah & Achudume, 2024)

Ethical Issue	Description	Mitigation Strategy
Data Privacy	AI systems often require access to large volumes of personal data.	Apply data minimization practices and anonymize sensitive data.
Consent and Transparency	Users may be unaware that their data is being collected or monitored.	Establish clear consent mechanisms and publish transparency reports.
Bias and Discrimination	AI systems can reinforce or magnify biases found in training data.	Train models with diverse datasets and perform regular bias assessments.

5.4 Education and awareness to reduce human vulnerabilities in AI-assisted environments

As human behavior is an important cybersecurity vulnerability (Brown *et al*., 2020;), it is an issue, despite the advanced AI tools. Criminals currently use AI to create intricate phishing, social engineering, and deepfake attacks (Brundage *et al*., 2018: Utomi *et al*., 2024). Organizations should invest in specific education and training to tackle such threats. As an example, dedicated AI-related risks, like identifying AI-created phishing messages or imitations by deepfake (Sarker *et al*., 2020), should be discussed during regular awareness sessions. System simulations (such as scenario-based ones) may assist personnel in training on how to react to AI-assisted attacks, gaining resilience in reality. In addition to end-users, developers and analysts should be trained to evaluate AI-generated recommendation critically and learn about the shortcomings and biases of AI models (Apruzzese *et al*., 2023). Cross-disciplinary training, which encompasses AI ethics, cybersecurity, and risks management, will create a range of individuals prepared to responsibly utilize AI and stay vigilant of its dangers (Spero, 2023). By so doing, education supplements technical and policy defense to produce a coherent, resilient security position.

Fig 5 : Privacy framework for AI driven Cybersecurity (After Roshanaei *et al*., 2024)

6.0 Conclusion

This study has critically examined the applications of the Artificial Intelligence (AI), large language models (LLM), and advanced machine learning techniques evolving modern cybersecurity. Incorporation of AI into cybersecurity operations has extended specialized ability to most basic statement of vulnerability recognition, forethought examination, real time anomaly survey and vast scale code examination with firm efficacy wide of the range of conventional, license based systems. Such developments increase the rate of detection, scale, and flexibility, and establish more secure methods against cyber threats becoming more complex.

Nonetheless, the research also notes that there is a set of sophisticated risks associated with AI. The dual-use characteristics of AI allow attackers to design highly effective phishing attacks, automate malware construction and use. AI systems themselves may become an object of such attacks, most commonly, in the form of adversarial attacks such as poisoning, evasion, and model extraction. Issues of explainability, bias over data and false positive or negative can hamper trust and performance with the use of AI models, especially when it comes to black-box models such as deep learning.

To make a responsible use of AI in cybersecurity, the analysis reveals the importance of ensuring the balanced perspective: to use technical solutions (such as adversarial training, explainable AI, and human-in-the-loop system) and powerful governance frameworks, regulatory standards (such as the NIST AI Risk Management Framework) and continuous employee education. In conclusion, AI is not an independent remedy but an effective tool and should complement human knowledge, strong policy, and ethical guidance to increase cybersecurity without creating any new vulnerability unacceptable.

7.0 References

Abdelhafez, A. S., A El-Sawy, A., & Sakr, F. (2023). Recent Studies and A Review about Malware detection and classification by using Artificial Intelligence Techniques. Benha Journal of Applied Sciences, 8(5), 89-104.

Abolade, Y.A. (2023). Bridging Mathematical Foundations and intelligent system: A statistical and machine learning approach. Communications in Physical Sciences, 9(4): 773-783

Abolade, Y. A., & Zhao, Y. (2024). A Study of EM Algorithm as an Imputation Method: A Model-Based Simulation Study with Application to a Synthetic Compositional Data. *Open Journal of Modelling and Simulation*, *12*(02), 33–42. https://doi.org/10.4236/ojmsi.2024.122002

Ademilua, D. A. (2021). Cloud Security in the Era of Big Data and IoT: A Review of Emerging Risks and Protective Technologies *Communication in Physical Sciences, 7*(4), 590–604.

Ademilua, D. A., & Areghan, E. (2022). AI-Driven Cloud Security Frameworks: Techniques, Challenges, and Lessons from Case Studies. *Communication in Physical Sciences, 8*(4), 674–688.

Alotaibi, A., & Rassam, M. A. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. Future Internet, 15(2), 62.

Ambati, S. (2023). Security and Authenticity of AI-generated code (Doctoral dissertation, University of Saskatchewan).

Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos, A. V., & Di Franco, F. (2023). The role of machine learning in cybersecurity. Digital Threats: Research and Practice, 4(1), 1-38.

Arnold, R. C. (2023). Internet of Things (IoT) Devices and Security: A Narrative Review. 2023_arnold_0516152302.

Bukhari, S., Tan, B., & De Carli, L. (2023). Distinguishing AI-and human-generated code: A case study. In Proceedings of the 2023 Workshop on Software Supply Chain Offensive Research and Ecosystem Defenses (pp. 17-25).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. arXiv preprint

arXiv:1802.07228.rocessing systems, 33, 1877-1901.

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., ... & Zaremba, W. (2021). Evaluating large language models trained on code. arXiv preprint. https://arxiv.org/abs/2107.03374

Conti, M., Dehghantanha, A., Franke, K., & Watson, S. (2018). Internet of Things security and forensics: Challenges and opportunities. Future Generation Computer Systems, 78, 544-546.

Dada, S. A., Shagan Azai, J., Umoren, J., Utomi, E., & Gyedu Akonor, B. (2024). Strengthening U.S. healthcare Supply Chain Resilience Through Data-Driven Strategies to Ensure Consistent *International Journal of Research Publications*, IJRP 164 (1), 70-79 https://doi.org/10.47119/IJRP1001641120257438

Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A comprehensive survey of generative adversarial networks (GANs) in cybersecurity intrusion detection. IEEE Access, 11, 76071-76094.

Garg, R. (2023). Preventing cyber attacks using artificial intelligence. i-Manager's Journal on Software Engineering, 18(2).

Gordon, A. D., Negreanu, C., Cambronero, J., Chakravarthy, R., Drosos, I., Fang, H., ... & Zorn, B. (2023). Co-audit: tools to help humans double-check AI-generated content. arXiv preprint arXiv:2310.01297.

Han, S., Lin, C., Shen, C., Wang, Q., & Guan, X. (2023). Interpreting adversarial examples in deep learning: A review. ACM Computing Surveys, 55(14s), 1-38.

Horne, D. (2023). Pwnpilot: Reflections on trusting trust in the age of large language models and ai code assistants. In 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE) (pp. 2457-2464). IEEE.

Kaviani, S., Han, K. J., & Sohn, I. (2022). Adversarial attacks and defenses on AI in medical imaging informatics: A survey. Expert Systems with Applications, 198, 116815.

Krishnamurthy, O. (2023). Enhancing cyber security enhancement through generative ai. International Journal of Universal Science and Engineering, 9(1), 35-50.

Komaragiri, V. B., & Edward, A. (2022). AI-Driven Vulnerability Management and Automated Threat Mitigation. International Journal of Scientific Research and Management (IJSRM), 10(10), 981-998.

Lin, J., Dang, L., Rahouti, M., & Xiong, K. (2021). ML attack models: Adversarial attacks and data poisoning attacks. arXiv preprint arXiv:2112.02797.

Manyam, S. (2022). Artificial intelligence's impact on social engineering attacks (Publication No. 561) [Capstone project, Governors State University]. OPUS. https://opus.govst.edu/capstones/561

Mbah, G. O., & Achudume, N. E. (2024). AI-powered cybersecurity: Strategic approaches to mitigate risk and safeguard data privacy. World Journal of Advanced Research and Reviews, 24(3), 310–327. https://doi.org/10.30574/wjarr.2024.24.3.3695

Mohamed, N. (2023). Current trends in AI and ML for cybersecurity: A state-of-the-art survey. Cogent Engineering, 10(2), 2272358.

Okolo, J. N. (2023). A Review of Machine and Deep Learning Approaches for Enhancing Cybersecurity and Privacy in the Internet of Devices. Communication in Physical Science. 9(4) : 754-772.

Rahman, A., Chakraborty, C., Anwar, A., Karim, M. R., Islam, M. J., Kundu, D., ... & Band, S. S. (2022). SDN–IoT empowered intelligent framework for industry 4.0 applications during COVID-19 pandemic. Cluster Computing, 25(4), 2351-2368.

Rahman, M. M., Arshi, A. S., Hasan, M. M., Mishu, S. F., Shahriar, H., & Wu, F. (2023). Security risk and attacks in AI: A survey of security and privacy. In 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 1834-1839). IEEE.

Ramirez, M. A., Kim, S. K., Hamadi, H. A., Damiani, E., Byon, Y. J., Kim, T. Y., ... & Yeun, C. Y. (2022). Poisoning attacks and defenses on artificial intelligence: A survey. arXiv preprint arXiv:2202.10276.

Reddy, A. R. P. (2021). The role of artificial intelligence in proactive cyber threat detection in cloud environments. NeuroQuantology, 19(12), 764-773.

Roshanaei, M., Khan, M. R., & Sylvester, N. N. (2024). Enhancing cybersecurity through AI and ML: Strategies, challenges, and future directions. Journal of Information Security, 15(3), 320–339. https://doi.org/10.4236/jis.2024.153019

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., & Zhong, C. (2022). Interpretable machine learning: Fundamental principles and 10 grand challenges. Statistic Surveys, 16, 1-85.

Safdarian, M., Trinka, E., Rahimi-Movaghar, V., Thomschewski, A., Aali, A., Abady, G. G., ... & Shetty, P. H. (2023). Global, regional, and national burden of spinal cord injury, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology, 22(11), 1026-1047.

Salem, A. H., Azzam, S. M., Emam, O. E., & *et al*. (2024). Advancing cybersecurity: A comprehensive review of AI-driven detection techniques. Journal of Big Data, 11(1), 105. https://doi.org/10.1186/s40537-024-00957-y

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7(1), 41.

Sarker, I. H. (2023). Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. Security and Privacy, 6(5), e295.

Shayegani, E., Mamun, M. A. A., Fu, Y., Zaree, P., Dong, Y., & Abu-Ghazaleh, N. (2023). Survey of vulnerabilities in large language models revealed by adversarial attacks. arXiv preprint arXiv:2310.10844.

Spero, E. J. (2023). User Interfaces, Mental Models, and Cybersecurity (Doctoral dissertation, Carleton University).

Srivastava, G., Jhaveri, R. H., Bhattacharya, S., Pandya, S., Maddikunta, P. K. R., Yenduri, G., ... & Gadekallu, T. R. (2022). XAI for cybersecurity: state of the art, challenges, open issues and future directions. arXiv preprint arXiv:2206.03585

Syed, S. A. (2022). Ai-powered cybercrime: the new frontier of digital threats. International Journal of Engineering Technology Research & Management (IJETRM), 6(02).

Tajrian, M., Rahman, A., Kabir, M. A., & Islam, M. R. (2023). A review of methodologies for fake news analysis. IEEe Access, 11, 73879-73893.

Tao, F., Akhtar, M. S., & Jiayuan, Z. (2021). The future of artificial intelligence in cybersecurity: A comprehensive survey. EAI Endorsed Transactions on Creative Technologies, 8(28).

Thangapandian, V. (2022). Machine learning in automated detection of ransomware: Scope, benefits and challenges. In Illumination of Artificial Intelligence in Cybersecurity and Forensics (pp. 345-372). Cham: Springer International Publishing.

Utomi, E., Osifowokan A. S., Donkor, A. A., & Yowetu, I. A. (2024). Evaluating the Impact of Data Protection Compliance on AI Development and Deployment in the U.S. Health sector. *World Journal of Advanced Research and Reviews*, WJARR, 24(2), 1100–1110. https://doi.org/10.30574/wjarr.2024.24.2.3398

Vadisetty, R., Polamarasetti, A., Prajapati, S., & Butani, J. B. (2023). Leveraging Generative AI for Automated Code Generation and Security Compliance in Cloud-Based DevOps Pipelines: A Review. Available at SSRN 5218298.

Vavekanand, R., Karttunen, P., Xu, Y., & Milani, S. (2024). Large language models in healthcare decision support: A review. *Preprints*. https://doi.org/10.20944/preprints202407.1842.v1

Zheng, Z., Ning, K., Wang, Y., Zhang, J., Zheng, D., Ye, M., & Chen, J. (2023). A survey of large language models for code: Evolution, benchmarking, and future trends. arXiv preprint arXiv:2311.10372.

Zhou, M., Duan, N., Liu, S., & Shum, H. Y. (2020). Progress in neural NLP: modeling, learning, and reasoning. Engineering, 6(3), 275-290.

Compliance with Ethical Standards

Declaration

Ethical Approval

Not Applicable

Availability of Data

Data shall be made available upon request.

Competing interests

The author declared no compositing interest

Funding

The authors declare that they have no known competing financial interests

Author's Contribution

The work was designed and written by both authors.