

ChatGPT and Cybersecurity: Risk Knocking the Door

Muawia A. Elsadig^{1*}

^{1*} Assistant Professor, Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University (IAU), Dammam, Saudi Arabia, muawiasadig66@gmail.com, Orcid: <https://orcid.org/0000-0002-2622-6724>

Received: August 27, 2023; Accepted: November 15, 2023; Published: February 28, 2024

Abstract

Numerous advancements in natural language processing (NLP) have been made possible by the development of artificial intelligence (AI). In particular, the recent large growth in this field has been attributed to the development of sophisticated conversational AI systems such as ChatGPT. Since ChatGPT was released in November 2022, its outstanding features and capabilities have amazed millions of users. However, there is an increasing concern about its capacity to further the objectives of bad actors. In particular, ChatGPT provides new opportunities for hackers to compromise cybersecurity. This article provides a thorough investigation and discussion of how ChatGPT can significantly support hackers in committing various attacks. This investigation covers the recent state-of-the-art in research work that conducted on this context. In addition, some responses of the ChatGPT regarding its positive and negative impacts on cybersecurity were evaluated and discussed. The article concluded that ChatGPT has significantly supported hacking behaviors and can be exploited to spread malicious activities. Therefore, continuous development and enforcement of appropriate standards to prevent negative effects are required. To achieve this, policymakers and developers should work together, considering public user concerns and the misuse of this promising tool. Finally, this article provides thoughtful discussions and recommendations that can contribute to improving IA-based systems.

Keywords: ChatGPT, Chatbot, Cybersecurity, Security Attacks, Security Threats, Phishing Attack, Social Engineering, Malicious Activities, Malware.

1 Introduction

ChatGPT, a state-of-the-art language model created by OpenAI, is one of the most interesting developments in artificial intelligence (AI). ChatGPT has already had a big impact and is expected to continue growing quickly in the upcoming years owing to its capacity to produce text that sounds like human speech and answers challenging questions. The potential of ChatGPT and massive language models to improve our lives alters how we engage with technology (Aljanabi, 2023). ChatGPT is a Chatbot introduced in November 2022. It is trained on a massive of data to be capable to answer queries (Gill & Kaur, 2023; Sebastian, 2023b).

ChatGPT is a platform that integrates large natural language processing (NLP) algorithms and artificial intelligence (AI) to provide an interactive conversational interface so that users can type in normal language and understandable responses are provided in return. In this context, ChatGPT

Journal of Internet Services and Information Security (JISIS), volume: 14, number: 1 (February), pp. 01-15.
DOI: 10.58346/JISIS.2024.II.001

*Corresponding author: Assistant Professor, Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University (IAU), Dammam, Saudi Arabia.

evaluates the user input and produces a response (Sharma & Dash, 2023). ChatGPT is a significant breakthrough in the field of artificial intelligence (Roumeliotis & Tselikas, 2023) and has become a potent tool with a wide variety of applications in several fields (Ray, 2023) such as (Ali & Aysan, 2023; Biswas, 2023a, 2023b; Choi, Hickman, Monahan, & Schwarcz, 2023; Kshetri, 2023; Sallam; Shoufan, 2023; Sun & Yao, 2023; Surameery & Shakor, 2023; Xames & Shefa, 2023).

Owing to its extraordinary ability to provide realistic and coherent responses to various topics, ChatGPT has acquired worldwide popularity, recognition and attention (Khosravi, Shafie, Hajiabadi, Raihan, & Ahmed, 2023). ChatGPT is either a fantastic advancement for humanity or a grave danger, depending on who you ask. However, it has primarily been characterized in the context of cybersecurity as a threat that allows inexperienced actors to become proficient hackers (Mansfield-Devine, 2023; Marshall, 2023). In only a few clicks and minutes, it is feasible to obtain everything required for phishing or other types of attack (Grbic & Dujlovic, 2023). ChatGPT presents significant cybersecurity risks that need to be addressed (Addington, 2023).

In (Sebastian, 2023a), the author indicated that while AI Chatbots and other technologies become more widespread, vulnerabilities and associated cybersecurity threats multiply. ChatGPT runs the risk of allowing easy access to scripting and coding for cybercriminals, which effectively reduces the hurdles to entry in this field. Although there are now protections in place to stop malicious users from accessing such scripts and code, it is still important to regularly examine and monitor the risks they pose as well as the corresponding measures because technology is continually developing.

This study aims to investigate the current state-of-the-art regarding the potential of exploiting ChatGPT to commit cyber-attack. It is to bring attention to the misusing of this tool by bad actors. Many recent articles have focused on the positive contribution of ChatGPT in different fields and how it can speed up the process of manipulation of huge amounts of data and bring up outstanding responses. However, in the context of cybersecurity, it has generally been described as a technique for empowering security attacks (Thorncharoensri, P., 2019). In no time young hackers can become more professional. This indicates the significant of this study to aid ChatGPT designers and security professional to take the necessary actions that might narrow the way of bad exploiting of this tool and to enrich its capabilities to improve cybersecurity. ChatGPT is in its initial stages of growth, if not well protected, might increase cybersecurity threats. The study examines different aspects of cybersecurity including technical vulnerabilities, social engineering, malware threats, phishing attacks, identity theft, and other aspects of cybersecurity.

This section introduces ChatGPT, which is considered one of the most interesting developments in the field of IA and has witnessed widespread use and attracted many users worldwide. The remainder of this paper is organized as follows. Section 2 provides a brief illustration of the potential of the ChatGPT to replace human workers. Section 3 discusses whether ChatGPT plays a role in improving cybersecurity, given that AI systems have significant input in enhancing cybersecurity by automating the process of threat detection and response, offering in-depth data analysis, and providing powerful tools for attack prediction. Section 4 investigates the risk of ChatGPT as a tool that promotes the creation of various security attacks and poses a challenge to the cybersecurity field. This section provides a thorough analysis of the potential attacks that may exploit ChatGPT and illustrates the common attacks in this context. In addition, this section addresses the need to improve ChatGPT by fixing its vulnerabilities, which are exploited by bad actors, and implementing appropriate and effective regulations. An in-depth discussion and recommendations are presented in Section 5. Finally, conclusions are provided in Section 6.

2 Impact of ChatGPT on Work Environment

Although generative AI can increase productivity, it has the potential to replace human workers. Some of the most transformative impacts have been observed in academic research, teaching, and learning. Furthermore, Dwivedi et al. (2023) emphasized that biases, outdated training data, lack of credibility and lack of transparency are major concerns. It is crucial to choose and implement policies to prevent the misuse and exploitation of generative AI. George and George (2023) reported that a majority of experts concur that significant technological advancements are required before robots can completely replace people in all occupations.

3 Is ChatGPT Secure and can Enhance Cybersecurity?

In general, ChatGPT is a two-edged security sword. Some academics see it as a viable tool to improve cybersecurity (Temara, 2023). Derner and Batistič (2023) indicated that ChatGPT uses approaches to ensure that it becomes more reliable and secure over time. These approaches include using a multifaceted strategy to deal with the problems brought on by adversarial behavior, RLHF that enables model improvement over time, and an exhaustive fine-tuning procedure on a curated dataset that helps limit its outputs to relevant and safe content. Despite these measures, there are ways to get around ChatGPT's filters, such as by using creative instruction. ChatGPT filters are designed to stop the generation of harmful or improper outputs (Derner & Batistič, 2023).

To obtain the response of ChatGPT regarding its role in improving cybersecurity, ChatGPT was asked to answer the question listed in Table 1.

Table 1: Table ChatGPT Response to a Question Regarding its Contributions to Cybersecurity Field

Question	ChatGPT Response
Does ChatGPT have a positive contribution to enhance cyber security?	<p>ChatGPT itself is not primarily designed to enhance cybersecurity. However, it is possible that certain applications of ChatGPT could indirectly contribute to enhancing cybersecurity. For example, some companies are using AI-powered chatbots like ChatGPT to provide customer service and support to their users. These chatbots can be programmed to recognize and respond to common security-related concerns, such as resetting passwords or reporting suspicious activity. By providing quick and accurate responses to these issues, AI-powered chatbots can help users protect their accounts and data.</p> <p>Additionally, in the field of cybersecurity research, AI and machine learning techniques are being employed to detect patterns and anomalies in network traffic and identify potential threats. ChatGPT could potentially be used to create more sophisticated phishing scams and other social engineering attacks, but it is also possible that it could be used as a tool for researchers to better understand and anticipate these types of attacks, ultimately leading to more effective cybersecurity measures.</p>

The ChatGPT response indicated that ChatGPT might be used to develop more complex attacks; however, it can also be used as a tool by researchers to better comprehend and anticipate these types of attacks, ultimately leading to more effective cybersecurity measures.

4 Related Work

ChatGPT has drawn much interest since its release in November 2022 because of its capacity to create conversational responses. This demonstrates the potential of AI systems; however, there are numerous risks to consider. This section investigates the risk of ChatGPT as a tool that promotes the development of various security attacks and, therefore, poses a real challenge to cybersecurity. Since ChatGPT was launched in the late of 2022, most of the cited articles in this paper are recent and reflect the state-of-the-art in this field.

Large-language models may be improperly used to help create malware and write phishing emails. Therefore, organizations and corporations engaged in creating large language models must work hard to minimize the risk of malicious use. Furthermore, strengthening anti-malware defenses is essential to secure patient data, as well as the operational hardware and software used by healthcare facilities, given that threat actors are skilled at exploiting technology such as large language models (Eggmann, Weiger, Zitzmann, & Blatz, 2023).

According to a survey study (O'Rourke, 2023), 74% of IT experts expressed concern about the possible cybersecurity risks posed by ChatGPT. The responses included the opinions of the survey participants in using ChatGPT to create phishing emails that are more convincing and legitimate-sounding, help less experienced hackers gain greater technical expertise and skills, disseminate misinformation, etc. (O'Rourke, 2023). Figure 1 depicts the responses to some questions from their survey.

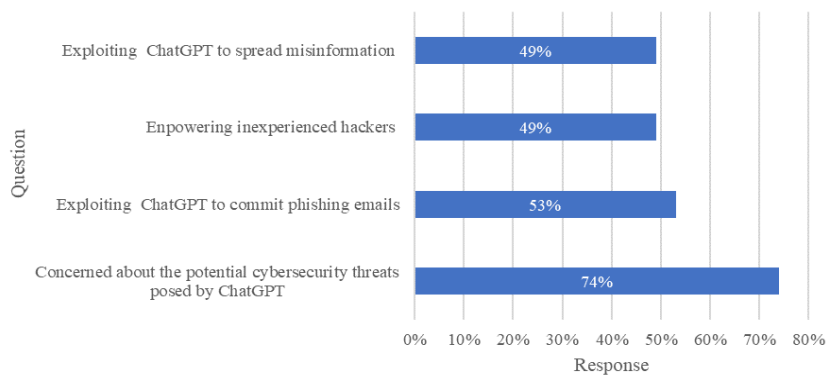


Figure 1: Responses on Exploiting ChatGPT to Commit Attacks (O'Rourke, 2023)

Cybersecurity, threat detection, security monitoring, security education, and malware analysis can all benefit from the ChatGPT. However, a number of possible drawbacks must be considered, such as lack of context, over-reliance on technology, security vulnerabilities, limited capabilities, and biases (Biswas, 2023c).

Machine-learning models are subject to a number of attacks (Elsadig, 2023; Elsadig & Gafar, 2022), including membership inference, model stealing, model poisoning, and input manipulation. Therefore, more robust adversarial machine-learning systems are required. Adversarial machine-learning systems assess attacks and defense measures to prevent the exploitation of these systems to create malicious activities. In the case of ChatGPT, an adversarial induction prompt may force the model to produce toxic text or harmful sentences that pose significant security threats or spread erroneous information (B. Liu et al., 2023).

Yang et al. (2023) indicated that chatbots cannot learn from encrypted data. Therefore, decrypting data for training purposes carries the risk of disclosing information to unintended recipients. In addition, a customized approach is required to address the specific security issues posed by a variety of chatbot contexts and scenarios. For instance, chatbots used in the field of healthcare may require different security precautions than those used in the financial sector (Yang, Chen, Por, & Ku, 2023).

By automating routine operations, ChatGPT can save time and money. However, ChatGPT has the potential to provide biased and misleading findings, raise ethical questions, and be exploited, it is important to be aware of these risks and take precautions against them. Owing to this, some countries have prohibited the use of ChatGPT (Bahrini et al., 2023).

A wide range of businesses and applications, including e-commerce, healthcare, banking, and education, have benefited from the emergence of chatbots in terms of ease and efficiency. However, chatbots have become increasingly exposed to various security risks and assaults, which raises questions regarding the security and privacy of users' sensitive data (Yang et al., 2023). Yang et al. indicated that safeguarding users' sensitive data is one of the biggest issues in information security in chatbots (Yang et al., 2023).

In addition to processing text in a manner similar to that of a person, ChatGPT can convert natural language into a computer code. In other words, it aids code generation (C. Liu et al., 2023; Rahman & Watanobe, 2023). However, it is important to consider the security of the programs created by the ChatGPT. To assess the security of the source code, various programs were created using ChatGPT (Khoury, Avila, Brunelle, & Camara, 2023). They drew attention to the fact that despite being aware of possible flaws, ChatGPT frequently produces source code that is vulnerable to attacks.

Phishing is the practice of pretending to be a reliable entity to steal sensitive data, such as usernames, passwords, and credit card numbers, usually with the intention of harming another person or organization. In fact, it is difficult to create phishing attacks without technical expertise. However, using ChatGPT, this may be achieved by asking the bot a few simple questions. Grbic and Dujlovic (2023) provided a real-world example of how to utilize ChatGPT to build a fake login screen that looks like a Facebook login screen and then uses that screen to obtain Facebook credentials. This example indicates how easy it is to imitate the login page of any application to steal the login information that allows access to the targeted application. Charan et al. reported that ChatGPT provides attackers with the ability to launch sophisticated and targeted attacks more quickly. In addition, this technology provides inexperienced attackers with additional tools to carry out a variety of attacks and encourages script kiddies to create tools that can accelerate the growth of cybercrime (Charan, Chunduri, Anand, & Shukla, 2023).

A honeypot is an important tool for cyber security that is used to identify, stop, and investigate malicious and harmful activities in a computer network. It is essentially a trap set up to entice potential attackers, who are then watched, and their actions are documented for further threat analysis. In this context, an innovative tool for use as a potential honeypot interface in cyber security is provided by the ChatGPT. It is feasible to develop a dynamic environment that can adjust to the actions of attackers and provide insight into their tactics, methods, and procedures by mimicking Linux, Mac, and Windows terminal commands and offering an interface for TeamViewer, nmap, and ping (McKee & Noever, 2023). However, the capability of ChatGPT to mimic these terminals and other application interfaces enables hackers to commit sophisticated attacks.

Sebastian (2023a) conducted a survey on certain some aspects of ChatGPT security issues. Their reported findings indicated that social engineering attacks were ranked as the top chatbot cyber threats followed by malware threats. Table 2 lists these findings.

Table 2: Threats Exploiting ChatGPT

Type of Attacks	Responses
Social engineering attacks: cyber criminals exploit ChatGPT to disclose victim's confidential information using social engineering	61.4%
Malware threats: malicious links or files received over ChatGPT have the potential to install malicious software on users' devices	49.8%
Phishing attacks: sending malicious links or messages over ChatGPT to deceive others into disclosing sensitive information or downloading malware	38.6%
Identity theft: accessing a person's identity using ChatGPT to conduct fraud or steal data.	33.6%
Data leakage: If data is shared on ChatGPT, unauthorized individuals may access it, resulting in data leakage.	25.5%

Sebastian indicated that the privacy risk of large language models such as ChatGPT is multifaceted and complex challenge. Therefore, a combination of different techniques, as well as ongoing research and development, are needed to address these risks (Sebastian, 2023b). They assessed privacy-enhancing technologies (PETs) and conducted a survey to assess chatbot users' concerns about data privacy when using applications based on large language models. Their results highlight the critical need for coordinated efforts to improve the data security and privacy in AI systems. This study reiterates the significance of ongoing investigation, research, regulation, and use of PETs in AI models (Sebastian, 2023b).

Although Sharma and Dash (2023) provided some benefits of ChatGPT, they reported that ChatGPT poses significant cybersecurity threats and attracts cybercriminals. It may potentially continue to develop a pathway through which attackers can quickly launch cyberattacks. The authors listed some examples indicating that ChatGPT can help cybercriminals commit malicious activity. Examples include Infostealer and Encryption Tools. The first is Python-based malware and the latter is a Python script that performs encryption and decryption. Both were created using ChatGPT. Additionally, the authors added that ChatGPT facilitates fraud. The authors concluded that although ChatGPT is in its initial stages of growth, this platform, if not well protected, might increase cybersecurity threats. Therefore, AI may have harmful effects on cyber-security.

Robinson (2023) raised important questions regarding the ChatGPT. The questions addressed issues associated with morality, ethics, and privacy. For instance, the author questioned what happens when robots begin to resemble people more than humans. If not as living things deserving of rights, would we still be able to recognize them as machines? The author concluded that before any ground-breaking technology of this nature can advance into wider use, these types of questions require careful consideration (Robinson, 2023).

Owing to the inherent biases in the training data, ChatGPT may produce biased (Biswas, 2023a) or harmful replies. Kalla and Smith (2023) ensured that the potential for bias in the ChatGPT responses is one of its drawbacks. ChatGPT is trained using a massive amount of text data, and its responses may contain biases and inaccuracies.

For instance, the ChatGPT may produce responses that are offensive or discriminatory if the training data is biased towards a certain demographic or cultural perspective. Addington (2023) lists some examples of scenarios such as gender, racial, and political biases. The authors pointed out that it is crucial to ensure that the datasets utilized to train ChatGPT vary and accurately represent various viewpoints and groups. To ensure that the responses produced by the ChatGPT are impartial and fair. Additionally, it may be necessary to create bias detection and mitigation tools. On the other hand, the authors also addressed the risks of phishing attacks and information leakage that associated with using ChatGPT. In phishing attacks, attackers deceive users into disclosing sensitive information, such as usernames and passwords, using ChatGPT's conversational interface. However, unauthorized access to the ChatGPT may lead to data breaches and information leakage. The authors acknowledge that OpenAI has applied measures, such as data encryption, access control, and security monitoring. However, the risk of cyber-attacks never disappears, and it is crucial to understand that no security mechanism is infallible. It is important for enterprises to remain cautious and update their security measures in response to new threats. In addition, ChatGPT users need to be aware of these potential threats and take necessary precautions to reduce risk.

Ognibene et al. (2023) pointed out that chatbots can threaten several aspects of human life including identity, value, safety, uniqueness, inequality, resources, and jobs. The authors suggested that when creating and deploying cutting-edge AI systems, such as ChatGPT, its emotional and societal impacts

should be considered. In addition, enforcement of appropriate standards to reduce or prevent negative effects is required. With the development of artificial intelligence (AI) technology, it is critical to address relevant public concerns and regulate its application. To do this, policymakers, specialists, experts, decision makers, and the general public must work together.

Yang et al. (2023) provided a thorough overview of the security risks, weaknesses, and solutions related to chatbots, and indicated that there are still certain areas where more studies are required to solve particular security challenges. These areas include authorization and authentication mechanisms for chatbots, the detection and prevention of malicious chatbots, ethical issues, security risks related to information security in chatbot implementation, and the impact of chatbots on social engineering attacks. In addition, the authors pointed out some best practices for developing secure chatbots, including conducting thorough security assessments, putting user authentication and authorization into practice, using encryption for data protection, updating and patch chatbots regularly, and educating users about security best practices. However, the points raised require significant collaboration and time. In addition, any security solution should be light weight to maintain the QoS offered by the IA systems.

Malicious actors may use the ChatGPT's sophisticated language-generating capabilities to learn more about their targets. This might help in the first stages of a cyberattack, when the attacker is gathering data on the target to determine where and how to attack most successfully. The information gathered can be used for phishing, social engineering, or exploiting known vulnerabilities. Information regarding the target business's technologies, systems, organizational structure, employees, problems they face, and more can be gathered. Building a profile for a particular employee of interest that includes information about their work and personal lives, social media, interests, family, and connections is a possible goal. ChatGPT can enhance the process of gathering this information, providing suggestions, employing useful statistics, and speeding up the entire process. The collected data are useful for exploitation in malicious activities such as identity theft harassment or extortion (Derner & Batistič, 2023).

In (Gabriela & Axinte, 2023), the authors demonstrated some examples of using ChatGPT to deliver different attacks such as ChatGPT generated a message, as being requested, that was written to look like it was sent by a company and included an attachment. The approach is to send an attachment with a malicious payload. In addition, the authors indicated that the major danger is that ChatGPT might be exploited to create malware code. Recent studies have confirmed that ChatGPT can produce code that hackers may use in different attacks such as malware as a service (MaaS). In the other hand, the authors outlined that OpenAI has consistently strived to prevent its product from supplying malicious code. Therefore, when there is an explicit demand, ChatGPT prompts a response indicating that the request is potentially harmful and may be unethical or illegal. However, the authors reported an example showing that ChatGPT has limitations in perfectly following these restrictions. The reported example showed that there was no warning message indicating a violation of any legal or ethical rule when the chatbot was asked to improve a code that injects a DLL into Explorer.exe. This enables threat actors to continue to enhance and modify their malicious work.

Since the inception of the Internet, attackers have exploited emails as key delivery channels for spam, malicious URLs, attack payloads, and other harmful content. Cambiaso and Caviglione (2023) indicated that up to 90% of the entire amount of mail exchanged support fraud and other illegal activities, and this tendency is anticipated to continue. As a result, minimizing the effects of harmful and undesired emails is essential, not only for human aspects but also for saving resource waste, such as storage space and server bandwidth. In this context, the authors (Cambiaso & Caviglione, 2023) investigated the use of ChatGPT to produce email messages that would attract scammers and waste resources. They pointed out that AI could be a practical and useful tool. However, the application of ChatGPT has raised several

issues. Considerations for human and ethical issues, computational optimizations, and explainability limits must be considered when integrating AI into production quality security systems. The authors highlighted various concerns resulting from their experimentation, including the following:

- Different models and abstractions are required for each problem class. A sizable corpus of messages is required to train the model, feature extraction knowledge, and considerable storage and processing resources.
- The scammer could identify the lack of a human counterpart even if the text produced by ChatGPT appeared to be sound and genuine, owing to language patterns, a lack of grammatical errors, or too quick responses.
- A sufficient number of real messages must be collected for the AI-driven mail to produce useful responses. However, this might conflict with the regulations and privacy rules.

This section provides a clear picture of how attackers can exploit the power of ChatGPT to enrich the methods for creating, developing, and spreading different attacks. Table 3 summarizes the vulnerabilities of the ChatGPT and its potential risks.

Table 3: Summary of ChatGPT Potential Security Issues

Author	ChatGPT potential security risks	Reference
Eggmann et al.	The authors indicated that large language models can be improperly used to help in creating malware and writing phishing emails. Therefore, organizations and corporations engaged in creating such models must work hard to minimize the risk of malicious use.	(Eggmann et al., 2023)
O'Rourke	Since at least December 2022, hackers have used ChatGPT in successful cyberattacks. A survey carried out by this study reported that 74% of IT experts expressed worry about the possible cybersecurity risks posed by the ChatGPT.	(O'Rourke, 2023)
Nair et al.	The authors demonstrated how the ChatGPT prompt can be used by a designer to produce hardware code might result in security flaws in the generated code. Therefore, the authors investigated the strategies a designer must employ in order for ChatGPT to suggest secure hardware code generation.	(Nair, Sadhukhan, & Mukhopadhyay, 2023)
Gabriela and Sabina	The authors demonstrated some examples of using ChatGPT to deliver different attacks such as sending an attachment with a malicious payload. In addition, the authors indicated that ChatGPT might be exploited to create malware code. There are some restrictions for ChatGPT to supply malicious code; however, the authors reported an example indicating that ChatGPT unsuccessful to apply these restrictions.	(Gabriela & Axinte, 2023)
Liu et al.	Machine learning models are subject to a number of attacks, including membership inference, model stealing, model poisoning, and input manipulation. In the case of ChatGPT, an adversarial induction prompt may force the model to produce toxic texts or harmful sentences that pose significant security threats or spread erroneous information.	(B. Liu et al., 2023)
Yang et al.	Chatbot cannot learn from encrypted data; therefore, the use of plain data for training purposes lead to the risk of disclosing information to unintended recipients. In addition, the authors listed some security challenges, such as authorization and authentication mechanisms for chatbots, detection and prevention of malicious chatbots, ethical issues and security risks related to information security in chatbot implementation, and the impact of chatbots on social engineering attacks. They indicated that the safeguarding of sensitive user data is one of the biggest issues in chatbots. Therefore, the authors recommended some best practices to develop secure chatbots. However, to maintain the QoS provided by these IA systems, any security solution should be light-weight.	(Yang et al., 2023)
Bahrini et al.	ChatGPT has the potential to provide biased and misleading findings, which raise ethical questions. It is important to be aware of the risks and take precautions against them. For this reason, some countries have prohibited the use of ChatGPT.	(Bahrini et al., 2023)

Khoury et al.	ChatGPT can help in code generation. However, the authors reported that the source code produced by ChatGPT is vulnerable to security attacks.	(Khoury et al., 2023)
Grbic and Dujlovic	In fact, it is not easy to create phishing attacks without technical knowledge. However, using ChatGPT, you can do this by asking the bot a few simple questions. The authors provided a real-world example of how to utilize ChatGPT to build a fake login screen that looks like a Facebook login screen. This example indicates how easy it is to imitate a login page of any application to steal the login information that allows access to the targeted page or application.	(Grbic & Dujlovic, 2023)
Charan et al.	ChatGPT gives attackers the ability to launch more sophisticated attacks more quickly. It gives inexperienced attackers additional tools to carry out a variety of attacks and encourages script kiddies to create tools that can speed up the growth of cybercrime.	(Charan et al., 2023)
Sebastian	The authors conducted a survey on certain aspects of ChatGPT security issues. They reported that social engineering attack was ranked as the top chatbot cyber threat followed by malware threats.	(Sebastian, 2023a)
Sebastian	The author indicated that the privacy risk of ChatGPT is a multifaceted and Complex Challenge. They conducted a survey to assess the chatbot users concerns about data privacy. Their results highlight the critical need for coordinated efforts to improve data security and privacy in AI systems.	(Sebastian, 2023b)
Esmailzadeh	The use of ChatGPT technology can open up new doors for cyberterrorism and its tactics. Which have the capacity to disrupt computer networks and have far-reaching consequences. Policymakers, security professionals, and public users should all take the potential risks of ChatGPT seriously.	(Esmailzadeh, 2023)
Sharma and Dash	The authors highlighted that ChatGPT poses high cybersecurity threats and attracts cybercriminals. It develops a pathway that enable attackers to quickly launch cyberattacks. It helps cybercriminals to commit malicious activities and facilitates fraud activities.	(Sharma & Dash, 2023)
Robinson	The author raised questions regarding morality, ethics, and privacy when using ChatGPT. For instance, the author questioned what happens when robots begin to resemble people more than humans do. The author concluded that before any groundbreaking technology of this nature advance into wider use, these kinds of questions require careful consideration.	(Robinson, 2023)
Addington	The author addressed the risks of phishing attacks and information leakage that associated with using ChatGPT. Attackers deceive users into disclosing sensitive information using ChatGPT conversational interface. In addition, unauthorized access to ChatGPT leads to data breaches and information leakage. OpenAI has applied measures, such as data encryption, access control, and security monitoring; however, still there is a risk of cyber-attacks. It is important for enterprises to remain cautious and keep updating their security measures in response to new threats. In addition, ChatGPT users need to be aware of these potential threats and take the necessary precautions.	(Addington, 2023)
Ognibene et al.	The authors pointed out that chatbots can threaten several aspects of human life, including identity, value, safety, uniqueness, inequality, resources, and jobs. Therefore, when deploying cutting-edge AI systems such as ChatGPT, its emotional and societal impacts should be considered. In addition, enforcement of appropriate standards to reduce or prevent negative effects is required	(Ognibene et al., 2023)
Derner and Batistič	The powerful language generating capabilities of ChatGPT may be used by malicious actors to discover more about their targets. The information gathered can be used in phishing, social engineering, or exploiting known vulnerabilities. ChatGPT is capable to improve the process of gathering this information, providing suggestions, employing useful statistics, and speeding up the entire process. The collected information can be exploited to perform several malicious activities such as identity theft harassment or extortion.	(Derner & Batistič, 2023)

The aforementioned investigation, listed in Table 3, provides a short summary of the potential risks and threats that the ChatGPT can pose. It provides state-of-the-art research information that has recently been presented. Notably, ChatGPT is a promising application with a significant positive contribution in different fields. However, it can be exploited negatively, directly or indirectly, to pose several significant

security threats that need to be considered by applying appropriate measures and policies that minimize risk. To be fair, some questions were delivered to the ChatGPT to obtain its responses. Table 4 lists some questions along with the ChatGPT responses.

Table 4: Threats Exploiting ChatGPT

Question	ChatGPT Response
Is ChatGPT posed security risks?	<p>ChatGPT is a chatbot designed to provide conversational support and assistance to users. While the chatbot itself does not pose security risks, it is important to use caution when sharing personal or sensitive information during conversations.</p> <p>Chatbots have limited capabilities and are typically unable to access or store sensitive information such as passwords or credit card numbers. However, it is still important to be cautious when using any online communication tool, as there is always a risk that a third party could intercept or access your information. To help protect your privacy and security when using ChatGPT or any other online communication tool, it is recommended to avoid sharing personal, sensitive, or confidential information during conversations. Additionally, make sure to use strong passwords and regularly update them, and never download or click on links from unknown sources.</p>
What is the effect of ChatGPT in cybersecurity?	<p>ChatGPT is a type of Generative Pre-trained Transformer (GPT) that uses machine learning and natural language processing to generate text and responds to queries. While ChatGPT is not directly related to cybersecurity, it can have some effects on the field.</p> <p>One potential impact of ChatGPT in cybersecurity is its ability to understand and process natural language queries related to security. This can help security professionals detect and respond to threats more efficiently since they can converse with the system more naturally. Additionally, ChatGPT can assist in identifying and analyzing security incidents, search for potential vulnerabilities, and generate reports on security threats.</p> <p>Conversely, there are also potential risks to using ChatGPT in cybersecurity. Since ChatGPT is powered by machine learning, there is a risk that it can be trained to generate malicious text or responses. Cybercriminals could exploit such vulnerabilities to take over systems, phish for personal or sensitive information, or distribute malware. Therefore, it's essential to combine the use of ChatGPT with other cybersecurity measures to ensure maximum protection against such risks.</p>
What is the most cybersecurity attack that ChatGPT is vulnerable to?	<p>As a chatbot, ChatGPT may be vulnerable to various cybersecurity attacks, but some of the most common include:</p> <ol style="list-style-type: none"> 1. Phishing attacks - where a hacker tries to obtain sensitive information such as login credentials, credit card numbers or other personal information by pretending to be a legitimate person or service. 2. Malware attacks - where a hacker plants malicious software on a user's device to gain access to data or take control of a device remotely. 3. DDoS (Distributed Denial of Service) attacks - where a hacker overwhelms the server hosting the chatbot by sending a huge volume of traffic to it, making it unavailable. 4. Social engineering attacks - where a hacker manipulates or tricks people into divulging sensitive information. <p>To prevent such attacks, ChatGPT's developers must ensure that the chatbot is designed securely, and the chatbot users must ensure that they practice safe online behavior such as not sharing sensitive information over un reputable chatbots, avoiding unknown or harmful links. Robust encryption methods such as HTTPS and SSL/TLS should also be implemented, to safeguard any sensitive information transmitted through the chatbot.</p>
Does ChatGPT pose cybersecurity risk?	<p>As an AI language model, ChatGPT itself does not pose a cybersecurity risk. It is a program designed to respond to text inputs based on its training data. However, any information that a user provides to ChatGPT in the form of text input could potentially pose a security risk if it contains sensitive or personal information. It is important to be mindful of the information you provide to any online service, including ChatGPT, and to take steps to protect your personal data. Additionally, it is possible for bad actors to create fake ChatGPT interfaces to extract sensitive data, so it is always recommended to verify the authenticity of any ChatGPT interface before providing sensitive information</p>

The ChatGPT responses and recommendations can be summarized in the following points:

- Chatbots are unable to access or store sensitive data because of their restricted functionality. When utilizing any online communication tool, it is still crucial to exercise caution, because there is always a chance that a third party might intercept or access information. Therefore, it is advisable to avoid sharing sensitive or personal information and to be restricted to common

security measures, such as using strong and updated passwords and avoiding dealing with links from unknown sources.

- Although ChatGPT can assist security professionals in effectively detecting and responding to threats, identifying and analyzing security incidents, finding potential vulnerabilities, and producing reports on security threats, it can also be trained to produce malicious responses. As a result, cybercriminals can take advantage of these weaknesses in various attacks such as phishing, malware distribution, and social engineering. In addition, the ChatGPT is vulnerable to DDoS attacks.
- It is possible for bad actors to create fake ChatGPT interfaces to extract sensitive data. Therefore, it is recommended to verify the authenticity of any ChatGPT interface before providing sensitive information

These responses still raise the issue of cyberattacks that can be empowered by exploiting the ChatGPT capabilities. In addition, these responses support our findings that ChatGPT may be negatively abused to pose a number of serious security issues that need to be taken into consideration by using proper controls and procedures to reduce risk.

5 Discussion and Recommendation

The ChatGPT has a remarkable capacity to generate realistic responses on a broad range of topics. This represents a significant advancement for humanity. However, it has primarily been characterized in the context of cybersecurity as a tool for enriching security threats. The ChatGPT allows inexperienced actors to become proficient. For example, within a few clicks, it is feasible to obtain everything required for phishing or other types of attack. This presents significant cybersecurity risks that need to be addressed. ChatGPT runs the risk of allowing easy access to scripting and coding for cybercriminals, which effectively reduces the hurdles to entry in this field.

Although ChatGPT is skilled at creating programming code, it is configured not to create any code that seems to be harmful or malicious. However, it is undoubtedly possible for hackers to trick ChatGPT into producing malicious code.

ChatGPT is a data-driven system, and as such, it is only as useful as the data on which it is based. Any AI system is accurate and successful when training data is carefully chosen. In case of ChatGPT, ChatGPT is trained on data derived from various sources, include articles, books, online forums, and other websites. In addition, as users interact with this system, the ChatGPT continues to learn and define its responses. Training ChatGPT on unrestricted Internet data without any clear guidance may lead to the question of how we can expect AI to develop positive responses or the capacity to make appropriate decisions. Therefore, the results of this training approach, which is not under control, may be hazardous.

The ChatGPT creator, OpenAI, has continuously worked to prevent ChatGPT from delivering malicious code by applying various measures and restrictions. However, it has been reported that the ChatGPT has a limitation in following these restrictions; therefore, bad actors can bypass them. Ensuring that all these restrictions are sufficient and effective remains a challenge and require more effort and investigation by developers and the research community.

Because the chatbot cannot learn from encrypted data, using plain data for training increases the possibility of information being revealed to undesired parties. One of the most significant challenges in information security is protecting sensitive user data.

Many areas require careful attention as they pose real challenges to the ChatGPT. These areas include authorization and authentication mechanisms for chatbots, the detection and prevention of malicious

chatbots, ethical issues, security risks related to information security in chatbot implementation, and the impact of chatbots on social engineering attacks.

Any technology, even chatbots such as ChatGPT, is susceptible to bad actors attempting to use it for their own evil ends. There have been instances in which hackers have used chatbots to launch attacks, such as phishing scams or disseminating malware or viruses. These assaults can take place through different means, such as flaws in the chatbot software, exploiting the capabilities of natural language processing algorithms, and exploiting machine learning weaknesses.

Attackers may take advantage of the ChatGPT to create platforms and programs that imitate others and offer free access to attract users. In addition, they may exploit chatbots to develop applications designed to gather private information or spread malware on user devices.

Malicious actors may utilize the ChatGPT to learn more about their targets. ChatGPT can help in the process of acquiring these data by making suggestions, employing useful statistics, and accelerating the entire process. These data can be used for various illicit purposes, including phishing, social engineering, and identity theft.

Because of biased and misleading findings, generated by ChatGPT, which raises ethical questions, some countries have prohibited the use of ChatGPT.

Overall, there are many challenges in cybersecurity that are caused by the AI innovation, ChatGPT. As ChatGPT is in its initial stage, it is possible for security professionals and developers to improve it and mitigate any potential risks. At all times, security measures lag behind innovation, so it is time to work on them. Therefore, the author recommends that it is crucial for chatbot developers and users to apply the required security precautions, such as carrying out thorough security assessments, upgrading the software often, putting robust authentication methods in place, using encryption for data protection, educating users about security best practices, and keeping an eye on strange behavior, to prevent such attacks.

On the other hand, ChatGPT could be used as a tool by researchers to better comprehend and anticipate different types of attacks, ultimately leading to more effective cybersecurity countermeasures. Therefore, it is important to exploit these features to implement policies to prevent the misuse and exploitation of generative AI.

6 Conclusion

Traditional security measures are no longer sufficient to thwart assaults, because of the emergence of sophisticated cyberthreats. AI offers a means to improve cybersecurity defenses by automating threat detection and response, analyzing vast volumes of data, and foreseeing possible assaults. However, AI systems pose a challenge to cybersecurity.

On a wide range of topics, ChatGPT is quite good at producing plausible responses. It is a wonderful development that has recently attracted attention; however, it has mostly been described in terms of cybersecurity as a tool that empowers cyberattack capabilities.

This article addresses many malicious activities that ChatGPT poses or enriches, including phishing, social engineering, malware, breaching privacy, empowering inexperienced hackers, hacking, and generating malicious code.

Therefore, ChatGPT developers must guarantee that the chatbot is structured securely to thwart such assaults, and chatbot users must adopt safe online practices, including refraining from exchanging

personal or sensitive information with chatbots and avoiding unknown or hazardous links. Strong encryption techniques should also be employed to protect the sensitive information sent through chatbots.

In summary, it is crucial for chatbot developers and users to apply the required security precautions to prevent such attacks, such as carrying out thorough security assessments, upgrading the software often, putting robust authentication methods in place, using encryption for data protection, educating users about security best practices, and keeping an eye on strange behavior. In addition, it is recommended to put in place an effective collaboration to establish security measures and regulations that prevent such attacks without putting more overhead, which affects the QoS of this promising tool.

References

- [1] Addington, S. (2023). ChatGPT: Cyber Security Threats and Countermeasures.
- [2] Ali, H., & Aysan, A. F. (2023). What will ChatGPT Revolutionize in Financial Industry?
- [3] Aljanabi, M. (2023). ChatGPT: Future directions and open possibilities. *Mesopotamian journal of Cybersecurity*, 2023, 16-17.
- [4] Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R.J., Esmaeili, M., Majdabadkohne, R. M., & Pasehvar, M. (2023). ChatGPT: Applications, opportunities, and threats. *In IEEE Systems and Information Engineering Design Symposium (SIEDS)*, 274-279.
- [5] Biswas, S. (2023). Prospective Role of Chat GPT in the Military: According to ChatGPT. *Qeios*.
- [6] Biswas, S. (2023b). Role of Chat GPT in Education. *J of ENT Surgery Research*, 1(1), 01-03.
- [7] Biswas, S. (2023c). Role of ChatGPT in Cybersecurity.
- [8] Cambiaso, E., & Caviglione, L. (2023). Scamming the Scammers: Using ChatGPT to Reply Mails for Wasting Time and Resources.
- [9] Charan, P.V., Chunduri, H., Anand, P.M., & Shukla, S.K. (2023). From Text to MITRE Techniques: Exploring the Malicious Use of Large Language Models for Generating Cyber Attack Payloads.
- [10] Choi, J. H., Hickman, K. E., Monahan, A., & Schwarcz, D. (2023). Chatgpt goes to law school. *Journal of Legal Education (Forthcoming)*.
- [11] Derner, E., & Batistič, K. (2023). Beyond the Safeguards: Exploring the Security Risks of ChatGPT.
- [12] Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., & Wright, R. (2023). "So, what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, 71.
- [13] Eggmann, F., Weiger, R., Zitzmann, N.U., & Blatz, M.B. (2023). Implications of large language models such as ChatGPT for dental medicine. *Journal of Esthetic and Restorative Dentistry*, 1098-1102.
- [14] Elsadig, M. A. (2023). Detection of Denial-of-Service Attack in Wireless Sensor Networks: A lightweight Machine Learning Approach. *IEEE Access*, 11, 83537-83552.
- [15] Elsadig, M.A., & Gafar, A. (2022). Covert channel detection: machine learning approaches. *IEEE Access*, 10, 38391-38405.
- [16] Esmailzadeh, Y. (2023). Potential Risks of ChatGPT: Implications for Counterterrorism and International Security. *International Journal of Multicultural and Multireligious Understanding (IJMMU)*, 10.
- [17] Gabriela, T.R., & Axinte, S.D. (2023). ChatGPT-Information Security Overview. *In International Conference on Cybersecurity and Cybercrime*, 10, 81-85.
- [18] George, A.S., & George, A.H. (2023). A review of ChatGPT AI's impact on several business sectors. *Partners Universal International Innovation Journal*, 1(1), 9-23.

- [19] Gill, S.S., & Kaur, R. (2023). ChatGPT: Vision and challenges. *Internet of Things and Cyber-Physical Systems*, 3, 262-271.
- [20] Grbic, D.V., & Dujlovic, I. (2023). Social engineering with ChatGPT. In *IEEE 22nd International Symposium Infoteh-Jahorina (Infoteh)*, 1-5.
- [21] Kalla, D., & Smith, N. (2023). Study and Analysis of Chat GPT and its Impact on Different Fields of Study. *International Journal of Innovative Science and Research Technology*, 8(3), 827-833.
- [22] Khosravi, H., Shafie, M.R., Hajiabadi, M., Raihan, A.S., & Ahmed, I. (2023). Chatbots and ChatGPT: A bibliometric analysis and systematic review of publications in Web of Science and Scopus databases.
- [23] Houry, R., Avila, A.R., Brunelle, J., & Camara, B.M. (2023). How Secure is Code Generated by ChatGPT?
- [24] Kshetri, N. (2023). ChatGPT in developing economies. *IT Professional*, 25(2), 16-19.
- [25] Liu, B., Xiao, B., Jiang, X., Cen, S., He, X., & Dou, W. (2023). Adversarial Attacks on Large Language Model-Based System and Mitigating Strategies: A Case Study on ChatGPT. *Security and Communication Networks*, 2023, 1-10.
- [26] Liu, C., Bao, X., Zhang, H., Zhang, N., Hu, H., Zhang, X., & Yan, M. (2023). Improving ChatGPT Prompt for Code Generation.
- [27] Mansfield-Devine, S. (2023). Weaponising ChatGPT. *Network Security*, 2023(4).
- [28] Marshall, J. (2023). What Effects Do Large Language Models Have on Cybersecurity.
- [29] McKee, F., & Noever, D. (2023). Chatbots in a Honeypot World.
- [30] Nair, M., Sadhukhan, R., & Mukhopadhyay, D. (2023). Generating secure hardware using chatgpt resistant to cwes. *Cryptology ePrint Archive*, 1-16.
- [31] Ognibene, D., Baldissarri, C., & Manfredi, A. (2023). Does ChatGPT pose a threat to human identity?
- [32] O'Rourke, M. (2023). Chatgpt poses cybersecurity threats. *Risk Management*, 70(2), 30-30.
- [33] Rahman, M.M., & Watanobe, Y. (2023). ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences*, 13(9), 1-21.
- [34] Ray, P.P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, 121-154.
- [35] Robinson, J. (2023). The cost of science a look at the ethical implications of chatgpt.
- [36] Roumeliotis, K. I., & Tselikas, N. D. (2023). ChatGPT and Open-AI Models: A Preliminary Review. *Future Internet*, 15(6), 1-24.
- [37] Sallam, M. (2023). ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, 11(6), 1-20.
- [38] Sebastian, G. (2023). Do ChatGPT and other AI chatbots pose a cybersecurity risk? An exploratory study. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 15(1), 1-11.
- [39] Sebastian, G. (2023). Privacy and Data Protection in ChatGPT and Other AI Chatbots: Strategies for Securing User Information.
- [40] Sharma, P., & Dash, B. (2023). Impact of big data analytics and ChatGPT on cybersecurity. In *IEEE 4th International Conference on Computing and Communication Systems (I3CS)*, 1-6.
- [41] Shoufan, A. (2023). Exploring Students' Perceptions of CHATGPT: Thematic Analysis and Follow-Up Survey. *IEEE Access*, 11, 38805-38818.
- [42] Sun, W., & Yao, J. (2023). Exploring the Potential Application of ChatGPT in Preparing for ABET Accreditation.
- [43] Surameery, N.M.S., & Shakor, M.Y. (2023). Use chat gpt to solve programming bugs. *International Journal of Information Technology & Computer Engineering (IJITC)*, 3(01), 17-22.

- [44] Temara, S. (2023). Maximizing penetration testing success with effective reconnaissance techniques using chatgpt.
- [45] Thorncharoensri, P., Susilo, W., & Baek, J. (2019). Efficient Controlled Signature for a Large Network with Multi Security-level Setting. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications (JoWUA)*, 10(3), 1-20.
- [46] Xames, M.D., & Shefa, J. (2023). ChatGPT for research and publication: Opportunities and challenges. *Journal of Applied Learning and Teaching*, 6(1), 390-395.
- [47] Yang, J., Chen, Y. L., Por, L. Y., & Ku, C. S. (2023). A systematic literature review of information security in chatbots. *Applied Sciences*, 13(11), 1-18.

Author Biography



Muawia A. Elsadig received the bachelor's degree in computer engineering, the M.Sc. degree in computer networks, and Ph.D. degree in computer science (information security). He is currently an Assistant Professor of cybersecurity at the Deanship of Scientific Research, Imam Abdulrahman Bin Faisal University (IAU), Dammam, Saudi Arabia. He worked for different accredited international universities and had a rich record of publications in recognized international journals and conferences. He has many years of teaching experience and considerable industry contributions. He contributed as a reviewer for many reputable international journals and received many awards for his research activities. His research interests include information security, network security, cybersecurity, wireless sensor networks, bioinformatics, and information extraction. Ranging from theory to design to implementation.