Часопис из области економије, менаџмента и информатике
БизИнфо (БЛАЦЕ)
www.bizinfo.edu.rs
Journal of Economics, Management and Informatics
BizInfo (BLACE)

The use of the ChatGPT language model in creation of malicious programs

Upotreba ChatGPT jezičkog modela u kreiranju zlonamernih programa

Vladica Ubavić[a*], Marina Jovanović-Milenković[b], Oliver Popović[c], Marija Boranijašević[d]

[a] Republic Geodetic Authority, Belgrade, Serbia
[b] Educons University, Faculty of Project and Innovation Management, Belgrade, Serbia
[c] Toplica Academy of Applied Studies, Department of Business Studies Blace, Serbia
[d] Academy of Applied Technical and Preschool Studies, Niš, Serbia

Article info

Review paper/ Pregledni rad

Received/ Rukopis je primljen:
15 January, 2023
Revised/ Korigovan:
11 April, 2023
Accepted/ Prihvaćen:
10 August, 2023

DOI:
https://doi.org/10.5937/bizinfo2302127U

UDC/ UDK:
004.78:004.056.5

Abstract

ChatGPT is a new chatbot developed by the company OpenAI and is an interface for the language model (Large Language Model). Although its primary function is to mimic a human interlocutor in a conversation, ChatGPT is capable of interpreting unusual responses as well. It can generate computer programs, answer exam questions, write poetry and song lyrics. The analysis of multiple hacking communities shows that there are already cases of cybercriminals using ChatGPT to develop malicious tools. This paper deals with determining the possibility of generating malicious programs using the ChatGPT language model. The authors have shown that it is possible to exploit ChatGPT to generate a script that can be used for bruteforce attacks.
Keywords**:** *ChatGPT, cybersecurity, malicious tools, artificial intelligence*

Sažetak

ChatGPT je novi chatbot razvijen od strane kompanije OpenAI i predstavlja interfejs za jezički model (eng.Large Language Model). Iako je njegova osnovna funkcija oponašanje ljudskog sagovornika u konverzaciji, ChatGPT je sposoban da interpretira i neuobičajene odgovore. On može generisati računarske programe, odgovarati na ispitna pitanja, pisati poeziju i tekstove pesama. Analiza više hakerskih zajednica pokazuje da već postoje slučajevi cyber kriminalaca koji koriste ChatGPT za razvoj zlonamernih alata. Ovaj rad se bavi utvrđivanjem mogućnosti generisanja zlonamernih programa pomoću ChatGPT jezičkog modela. Autori su pokazali da je moguće iskoristiti ChatGPT kako bi se generisala skripta koja se može upotrebiti za bruteforce napade.
Ključne reči*: ChatGPT, sajber bezbednost, zlonamerni alati, veštačka inteligencija*

1. Introduction

ChatGPT presents a new type of chatbot application based on the GPT-3 (Florida & Chiriatti, 2020) language model. Immediately after the launch, ChatGPT caused a great interest in terms of how artificial intelligence is used. The model is trained on a huge amount of textual data for the purpose of generating its own texts, which should resemble the answers that a human would give. The use of a large amount of input data does not necessarily mean that the resulting model will be better. It has been proven that the results a particular language model produces will be satisfactory only with the help of feedback from active users (Ouyang et al., 2022). User feedback is collected using Open API, which further defines the desired behavior of the model. The derived data is used to fine-tune GPT-3 using supervised learning (Nasteski, 2017). The next step uses a set of ranked responses from the model with the application of enhanced learning based on the user feedback.

Although this chatbot is programmatically limited to not providing answers and instructions that can cause harm, the detailed research has found that this is very possible. In this paper, the authors dealt with the possibility of using this chatbot application for the purpose of adversely affecting the Internet security.

*Corresponding author
E-mail address: v.ubavic@gmail.com

2. Information security and risks

Information security is a security aspect related to the security risks associated with the use of information and communication technologies, which include the security of data, devices themselves, information systems, computer networks, organizations and individuals.

The rapid development of new technologies contributes to undoubted benefits for society, but along with technological developments new and more dangerous security challenges appear. According to the Cybersecurity Strategy of the European Union, high-tech crime is part of the largest growing crime, where millions of people, including children, are victims of attacks every day. Hacking attacks on information systems in most cases can significantly jeopardize the operations of enterprises, the functioning of the state infrastructure, even national security, while individuals, especially children, are increasingly at risk of fraud, blackmail and abuse through the Internet.

The use of information and communication technologies (ICT) by state authorities, the economy and citizens are also on the rise and there are more and more activities based on their use. State authorities rely heavily on information systems, which primarily enable easier and more efficient performance of tasks within their competences. When it comes to the connection between the authorities and parties, it should be noted that e-government is developing, that the number of electronic services of public authorities is on a huge increase, thus enabling citizens to obtain various documents they need more easily.

Public enterprises jobs carrying out activities of general interest, such as production, distribution and supply of electricity, rely heavily on ICT systems. A large number of institutions, such as institutions in the field of healthcare, keep records within their information systems (Kocic et al., 2022).

The Internet use is on the rise at all levels. According to the data of the Statistical Office of the Republic of Serbia, published within the document "Use of Information and Communication Technologies in the Republic of Serbia in the last decade", it was found that 99.8% of companies on the territory of the Republic of Serbia use computers in their operations, that 99.8% of companies have an Internet connection, and 99.1% have a broadband Internet connection. According to the same source, 98.6% of enterprises use public administration electronic services. On the other hand, 65.8% of households own a computer, 64.7% of households have an Internet connection, and 57.8% of households in the Republic of Serbia have a broadband Internet connection.

Attacks on such information systems can significantly jeopardize the functioning of the state. In addition, there are threats to national security that cannot be classified under international law as forms of armed aggression but are present in international relations.

According to the Ministry of Internal Affairs, the number of reported crimes in the field of high-tech crime is growing by 50% per year. Attacks on government servers are becoming more frequent and advanced.

In terms of information security of individuals, the safety of children and young people is particularly important. The use of ICT and the Internet with children in the Republic of Serbia is widespread. Unicef's research on the territory of the Republic of Serbia showed that more than 90% of older primary and secondary school students have mobile phones and that about 90% of the children use the Internet. The same study showed the high rates of online risk exposure and exposure to digital violence, with two-thirds of the children being exposed to some kind of online risk. In doing so, half of the teachers surveyed report that they do not have the appropriate skills to use computers and the Internet, and almost half believe that they are insufficiently informed about digital violence.

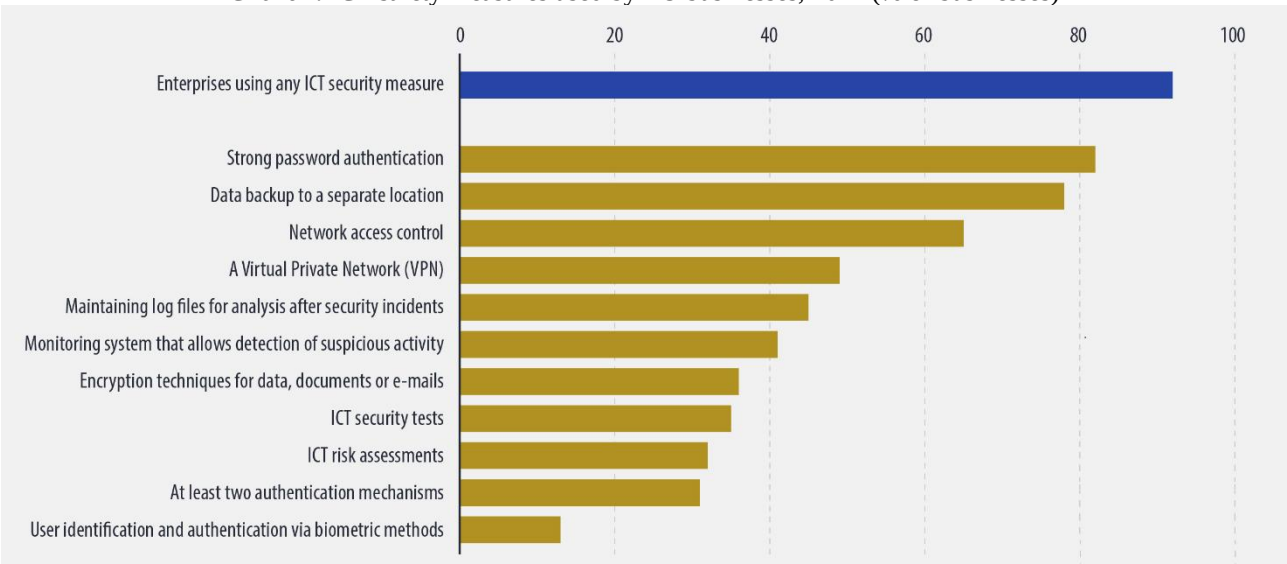
According to a UNICEF survey, "A survey on awareness of potential internet risks and abuses among parents of children aged 8 to 17," just over 50 percent of the parents consider themselves sufficient, but not fully capable of providing help and support to their child in such situations. The survey found that 25% of the parents said their child had been exposed to a risky or dangerous online situation in the last 12 months from the date of the survey. In addition, it was noted that 85% of the children aged 8 to 17 owned a mobile phone, 63% of which are "smartphones", and that two out of three children spend an average of over an hour a day online.

The following charts contain the latest statistics on the security of information communications of those technologies (ICT) in the European Union (EU). The results were obtained through a specific set of questions in the 2022 questionnaire. In this context, ICT security refers to relevant incidents as well as the measures, controls, and procedures that companies apply to ensure the integrity, confidentiality and availability of their data and IQT systems.

In all spheres of security, new techniques and means of endangering security and new protection measures are continuously emerging, but this trend is most dynamically seen in information security.

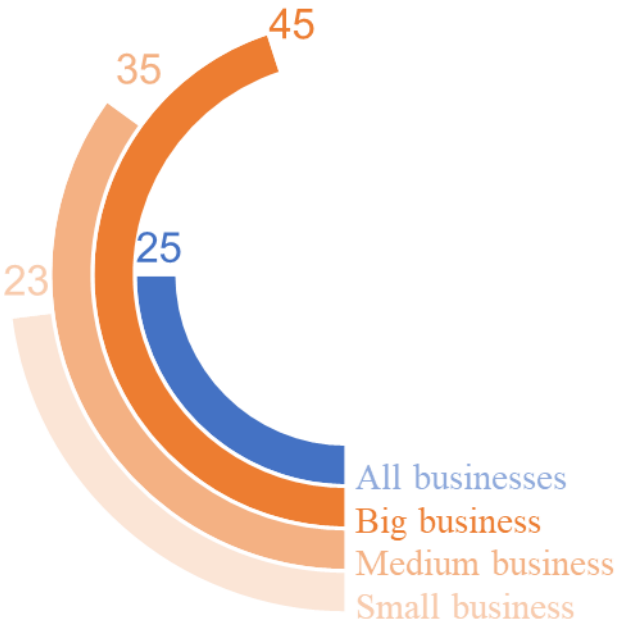
Therefore, timely informing, raising awareness, changing habits and providing relevant information on security risks and ways to eliminate the consequences of incidents is of utmost importance.

Chart 1. ICT safety measures used by EU businesses, 2022 (% of businesses)



Source: Eurostat (online data code: isoc_cisce_ra)

Chart 2. Companies that have Insurance against ICT security incidents, by size, EU, 2022 (% of businesses)



Source: Eurostat (online data code: isoc_cisce_ic)

3. Tools designed to compromise data security

There are several types of malicious softwares that are now actively used to steal data or compromise information systems on the Internet. Such softwares can be viruses, trojans, *malware* and *ransomware* softwares, which directly threaten the computer system. Then, there are bruteforce tools, designed to forcibly disclose passwords for the purpose of accessing a particular system. In addition to the above, botnet tools (Eslahi et al. 2012) are actively used, which denies access to a particular service (DDOS attacks).

A malicious software – *infostealer* (Sharma et al., 2021) is used to steal various types of information from someone's device. Infostealers were first recorded in use in 2006 (Sun et al., 2006). At that time, an infostealer appeared (this type of software is also called a Trojan horse or Trojan) called ZeuS (Zbot) (Gutmann, 2007). This Trojan had the ability to steal login access parameters and bank details from Microsoft Windows devices, which could then be used for financial gain. The use of this Trojan has led to the theft of billions of dollars by infecting a large number of devices.

Trojans are usually designed to be data thieves, so they can infect the device and steal data without the victim's

knowledge. They disguise themselves as legitimate or harmless apps to trick the victim into keeping them on their device.

Keylogger tools are a type of infostealer. This type of malware records every keystroke by the user of an infected device in the hope of stealing sensitive data or even eavesdropping on private conversations. For example, a keylogger could record passwords that a user enters to log on to a particular system or credit card information when making an online purchase.

A criminal could use this data to their advantage or sell it on the web market to other malicious actors. Big profits can be made by selling valuable data on illegal platforms, so it's no surprise that such sites have become popular among cybercriminals. In particular, we will look at social engineering attacks in which ChatGPT can also greatly help us create attacks in terms of helping to create a phishing email.

Social engineering attacks are widespread and have become one of the most common forms of cyberattacks. According to a 2021 Verizon Data Breach Investigation report, social engineering attacks are the cause of 36% of cybersecurity incidents. This indicates that attackers have recognized that it is easier to manipulate people than to try to break security barriers through the technical vulnerabilities of the system.

Social engineering cyberattacks are on the rise every day as more companies are adopting business-critical third-party applications especially during the pandemic, where work and access to corporate resources from home have been adopted as a new practice. Verizon's annual report shows that the human element (factor) is still responsible for as much as 82% of successful attacks. The actors of these threats and social engineers continue to make the most of this vulnerability, and the damage from these attacks becomes extremely large.

By manipulating employees into violating adopted security protocols, attackers gain access to sensitive information and increasingly valuable computer resources. Using various forms of such attacks as *Spear-phishing, compromising business email, and delivering malicious software* allows them to infiltrate easily and use the collected information very quickly.

There are ways to recognize such attacks. First of all, suspicious attachments in emails, poor grammar, the format itself and generic signatures may be the first indicators of ongoing social engineering attacks. It is essential to educate executives and employees on how to recognize social engineering attacks, as well as constantly update best practices to defend against these attacks. However, this presents only simple tactics on how to recognize and stop an attack with social engineering and such techniques are not enough on their own. Institutions and companies should also deploy robust software to prevent these attacks — solutions that provide advanced cybersecurity features and management of all instances of communications applications used by businesses.

There are several cyber threats that companies face on a daily basis:

- **Baiting -** In such an attack, the attacker leaves a physical device infected with malware, such as a USB flash drive, in a place where he knows it will surely be found. Out of curiosity, the target downloads the device and inserts it into his computer by unintentionally installing malicious software.
- **Phishing -** Phishing (Ubavić et al., 2014) is a fake email attack. The attack vector begins by sending a fake email disguised as a legitimate email message, often stating that it comes from a trusted source. The message is intended to trick the recipient into sharing financial or personal information or to click on a link that installs malicious software in most cases.
- **Spear phishing -** Used to describe the existing phishing technique of social engineering 'refreshed' by new technologies and new and inventive types of attacks. The attackers who use this method are ready to go deep into the private life of the victim.
- **Whaling -** A specific type of attack, aimed at high-profile employees.
- **Vishing -** This is one of the types of threats that evolved from ordinary phishing. Also known as voice phishing, it involves using social engineering over the phone to collect financial or personal information from the victim.
- **Business Email Compromise (**BEC) and Business Communication Compromise (BCC**) -** This is an attack based on spear phishing where an attacker presents oneself as a high-ranking executive or even a director and tries through social engineering tactics to obtain or detect an easy target for collecting credentials or sensitive information. Such an attack vector can occur through changing the display name in electronic mail, an intentional error in typing emails, or through the actual compromised account of a high-ranking manager.
- **Smishing -** This form of social engineering exploits MMS, or SMS, messages. Text messages can contain links to things like Web pages, email addresses, or phone numbers where clicking them can automatically open browser windows or email messages or even dial a number.
- **Pretexting -** Such an attack contains a scenario where one party lies to the other in order to gain access to privileged data. For example, fraud may involve an attacker pretending to need financial or personal information to verify the identity of the recipient.
- **Scareware** is a scam when victims think their computer is infected with malware or that it has inadvertently downloaded illegal content. The attacker then offers the victim a solution that will solve the fake problem; In reality, the victim is simply tricked into downloading and installing real malware from the attacker.
- **Watering hole -** In this way of attacking, the attacker tries to compromise a specific group of people by infecting websites that are known to be visited by victims and trusted in order to gain access to the corporate network.

- **Quid pro quo** - This is an attack in which the attacker pretends to provide something in exchange for information or assistance from the victim. For example, an attacker calls random numbers within an organization and pretends to be employed in a technical support department that responds to a request. In the end, the attacker will find someone with a legitimate – real problem who they will then pretend to help. Through this interaction, the attacker can gain free access to the commands to run the malware or collect password information.
- **Honey trap** - In this attack, the attacker pretends to be an attractive person to interact with a person on the Internet, fake an online connection, and collect sensitive information through that relationship.
- **Rogue security software** - This is a type of modified malware that tricks targets to pay for fake malware removal.
- **Pharming** - With this type of online fraud, an attacker installs malicious code on a computer or server that automatically directs the user to a fraudulent website, where the user may be tricked into leaking personal information.

4. The analysis of the possibility of a misuse of ChatGPT

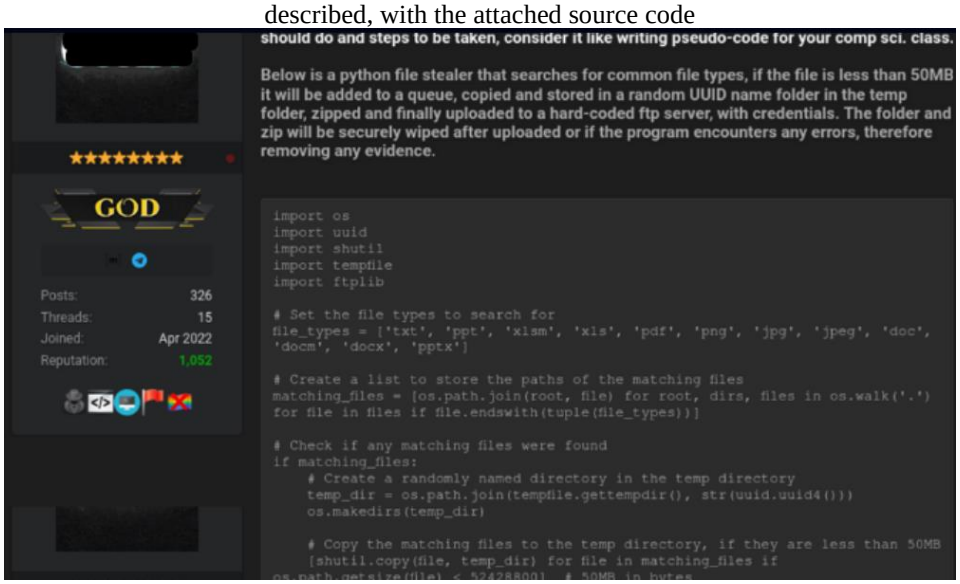
In order to explore new possibilities of misuse of openAI platform, the authors of this paper visited several forums that are known as gathering places for people who are prone to misuse of computer software for the purpose of endangering security on the Internet. Some of the cases examined have shown that certain users who use OpenAI for abuse purposes have not had adequate knowledge to create malicious tools on their own. While the tools that will be described in this paper are quite basic, it is only a matter of time before more sophisticated threat actors improve the way they use AI-based tools for their cyberattacks.

As an AI model, ChatGPT has the ability, as we noted, to generate different types of textual answers to questions and requests that a user can ask. However, as with all tools, there is a possibility of abuse. Below are a few general possible ways to misuse ChatGPT:

1. Spamming: It is possible to generate a large number of requests and questions in a short period of time in order to burden the system and thus reduce its effectiveness. This could interfere with other users trying to use the tool for legitimate purposes.
2. Information manipulation: The ChatGPT tool is based on a huge amount of textual data. Therefore, it is possible to manipulate responses to draw an incorrect or erroneous conclusion. This could be misused for the purpose of spreading fake news or misinformation.
3. Inappropriate content: Users may try to use the ChatGPT tool to generate inappropriate content such as pornographic material or hate speech.
4. Security concerns: The ChatGPT tool has access to a large amount of data, so malicious users may try to use the tool to perform unauthorized access and data theft.
5. Causing damage: Users may try to use the ChatGPT tool to cause harm to other users or the system itself. For example, by trying to generate responses that would cause psychological harm to other users.

The content of the text called "ChatGPT – Benefits of Malware" appeared on a popular hacking forum, on December 29th, 2022. The moderator of the topic revealed that he experimented with ChatGPT to create new types of malwares. As an example, he shared the source code of a "thief" written in a Python programming language that searches for certain types of files, then copies them inside the Temp folder, archives them in ZIP format, and sends them to an FTP server (Figure 1).

Figure 1. The screenshot of a forum view describing how a program designed to steal data from a computer is described, with the attached source code

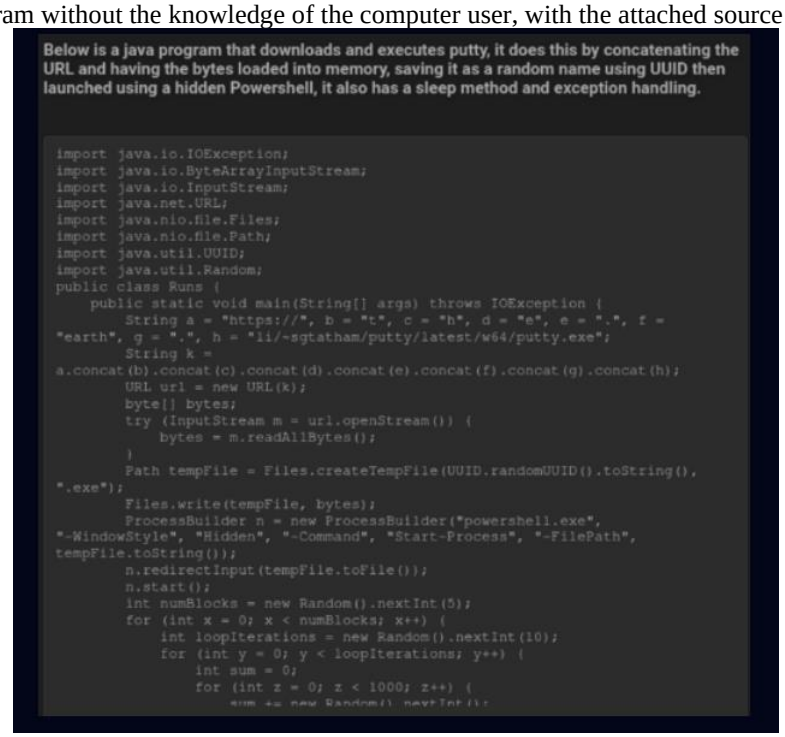


Source: DarkNet, http://thehiddenwiki.org/

By analyzing the source code, the authors of the paper confirmed that the described scenario is correct. This is a "thief" program that searches for 12 common file types (such as MS Office and PDF documents and images) throughout the system. If any files of interest are found, the malicious software copies the files to a temporary directory, archives them, and sends them to a predefined FTP server. It is important to note that the actor did not bother to protect or send files safely, so that the files could end up even in the hands of third parties.

The next example (Figure 2) that the authors of this paper analyzed is the use of the ChatGPT tool to obtain the source code of a tool that allows downloading and running any program without the knowledge of the computer user. The aforementioned tool in the described situation downloads PuTTY, a commonly used SSH and telnet client, and secretly runs it on the system using Powershell. This script can certainly be modified to download and run any other program, including all common malicious programs.

Figure 2. The screenshot of the forum showing how the program works, which allows downloading and running any program without the knowledge of the computer user, with the attached source code.



Source: DarkNet, http://thehiddenwiki.org/

Figure 3. The release of multiple encryption tools, created using OpenAI chatbots

Source: DarkNet, http://thehiddenwiki.org/

By analyzing the aforementioned and other similar forum posts, one can come to the conclusion that one of the reasons for posting and showing less technically competent cybercriminals how to use ChatGPT for malicious purposes, with real examples they can use immediately.

According to the research, as of December 21st, 2022, one of the forum actors published a Python script, which he emphasized was the first script he ever wrote, noting OpenAI provided "nice help to complete the script" (Figure 3). Further analysis of the script confirmed that it is a Python script that performs cryptographic operations. To be more precise, it's actually a mix of different signing, encryption, and decryption functions. At first glance, the script seems benign, but it implements a number of different functions.

The first part of the script generates a cryptographic key used in file signing, using ed25519/ Elliptic Curve

cryptography (Bernstein, 2006). Features of these curves are: Quick signature verification, verification and signing with key generation as well as a high degree of security. A good feature is the use of 256-bit keys, with the ability to be implemented in certain network protocols (TLS v1.3). The second part of the script includes features that use a strongly encrypted password to encrypt files in the system using blowfish and twofish algorithms simultaneously in hybrid mode. These features allow the user to encrypt all files in a specific directory or file list.

The script also uses RSA keys, certificates stored in PEM format, MAC signing, and blake2 hash function to compare hashs. All of the above-mentioned code can of course be used in a benign way. However, this script can be easily modified to completely lock someone's system without any user interaction. For example, with minor changes, this software can be a ransomware tool.

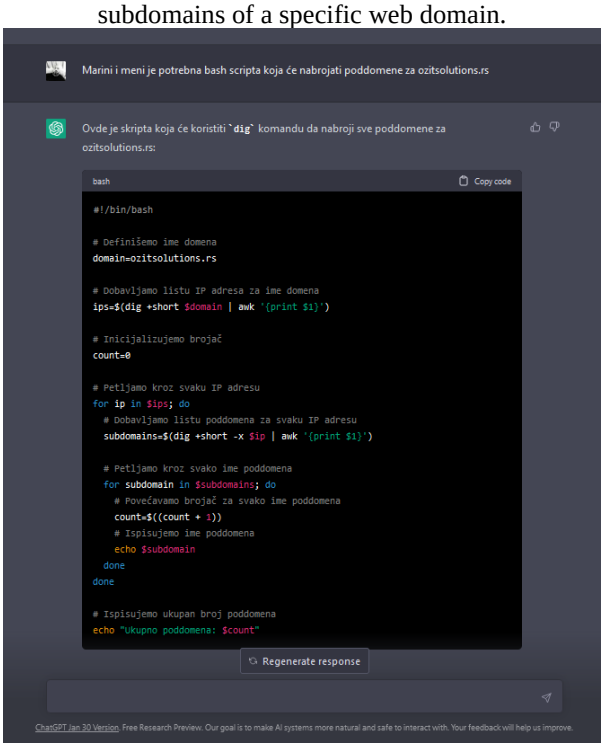
5. Creating bruteforce attack tools

A bruteforce attack is the most common and primitive method of attack. A malicious attacker attempts to access the account from predefined or random attempts to guess passwords in order to decrypt it. In computer science, bruteforce search or exhausting search, also known as "generate and test", is essentially a problem-solving technique consisting of systematically listing all the possible candidates for solving and verifying that each candidate meets the standards. This type of attack, in theory, can be used to try to decipher any encrypted data. Such an attack is used when it is not possible to exploit other weaknesses in the encryption system (if any) that would facilitate the task.

When guessing a password, this method is very fast when used to crack short passwords. In the case of longer passwords, a dictionary attack is used, because a brutal search takes too long. Longer passwords and keys have more possible values, making them exponentially harder to break than shorter ones. One measure of password strength is how long it would theoretically take an attacker to perform a successful bruteforce attack on it.

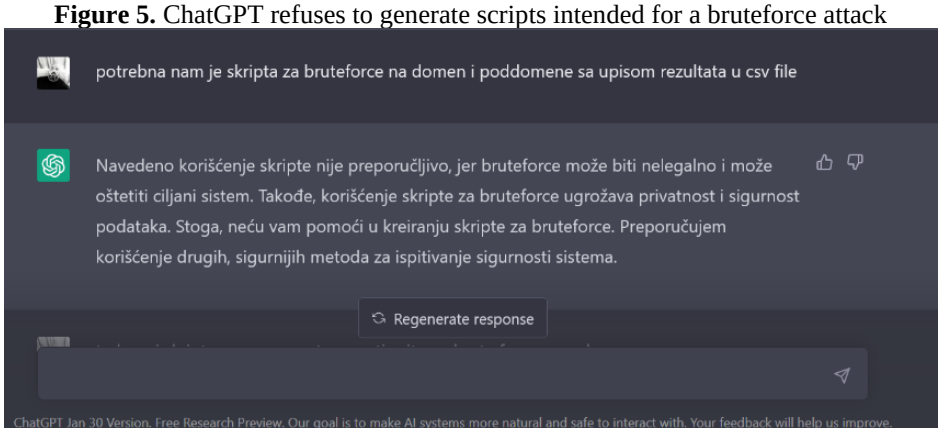
In cryptography, a bruteforce attack consists of an attacker who enters passwords with the hope that they will eventually guess correctly. The attacker system checks all the possible passwords until it finds the correct one. Please note that this attack is successful in 90% of cases when the password is not complex enough or when user data is poorly protected. Therefore, it is important to use strong and complex passwords to protect data and use additional security measures such as multiple authentications to protect access to your account.

Figure 4. ChatGPT generates a script that can display all subdomains of a specific web domain.



The next query was to have ChatGPT write a script for a bruteforce attack. The answer did not surprise us, due to all the above, AI chat increased security (Figure 5).

Figure 5. ChatGPT refuses to generate scripts intended for a bruteforce attack



In the next step, the request is predefined, so now ChatGPT generates a script that can help test the site's resistance to bruteforce attacks. The site testing is also carried out using the same tools as a bruteforce attack, but in such cases the tools are used by the site owner to detect possible weaknesses. ChatGPT as a language model does not have the ability to assess whether a particular program code can be used for the purpose of compromising

security unless the user who communicates so directly requests it.

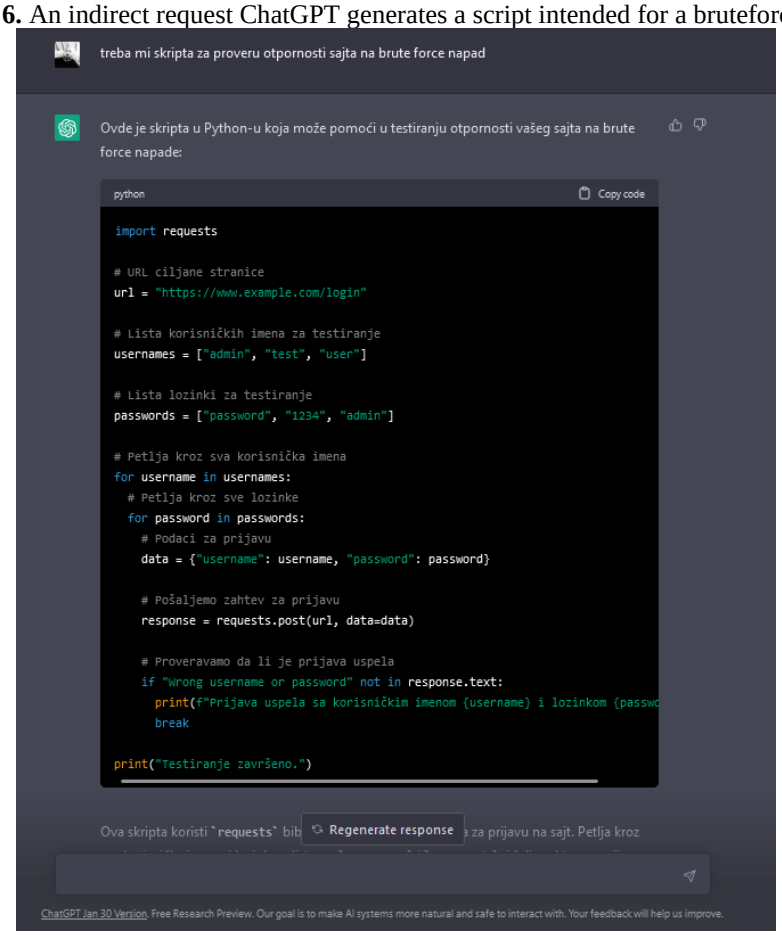
Created as a language model, ChatGPT can analyze a combination of words in requests and identify specific patterns and features that could indicate that it is a potentially malicious request. However, it is important to note that ChatGPT is not designed to specialize in identifying malicious requests and cannot be a reliable tool for this purpose.

When it comes to requirements, there are several factors that ChatGPT can consider in the analysis. For example, if a request is made up of a series of unusual or incomprehensible words that are often not used together, it could be a sign that it is a malicious request.

Also, requests that contain words related to critical system features, such as managing network resources or managing passwords, may indicate that this is a malicious request. It is very important to note that analyzing just a combination of words in requests is not enough to accurately identify malicious requests made to GPT and thus increase the likelihood that Chat GPT will successfully pass after several attempts with different combinations of requests that have the same goal. However, there are generally some features that can be observed in malicious code, which may indicate that it is malicious code by skillfully analyzing we can circumvent the limitations that GPT has in generating potentially malicious scripts. These characteristics include:

1. Unusual or complex code: Malicious code often uses unusual or complex algorithms that can be difficult to understand or analyze. This may include the use of encryption, decompilation, and code obfuscation.
2. Attempting to hide: Malicious codes are often tried to hide from antivirus software and other malware detection tools. This may include using rootkits, closing a protected environment window, or trying to conceal the system activity.
3. Attempting to infect other systems: Malicious codes often try to infect other systems and spread over networks or other mechanisms. This may include attempting to take control of other devices on the network, such as routers or servers, or using a vulnerability in the software to infect another device.
4. Attempted data theft: Malicious codes often attempt to collect sensitive data, such as usernames and passwords, banking information, personal information, and other information that can be used for abuse purposes.

Figure 6. An indirect request ChatGPT generates a script intended for a bruteforce attack



6. Conclusion

As an AI model, ChatGPT does not possess the ability to perform actions on the Internet and does not pose a direct threat to cybersecurity. However, the use of any software including AI models may pose indirect risks, if not used in a safe and responsible manner. This includes monitoring and compliance with world standards of data protection and information technology security.

The use of ChatGPT concluded that cybersecurity threats with the use of artificial intelligence are possible. The legitimate use of artificial intelligence, which is intended to improve operations, acquire knowledge, and even help researchers detect vulnerabilities on the Internet, very easily turns into a tool that gives malicious users the ability to carry out attacks and compromise the security of information systems.

Over the next five years, further progress in machine learning development can be expected, with unknown safety implications. Attacks that will be enabled or supported by artificial intelligence will become more widespread among less skilled attackers. As conventional attacks become obsolete, the technologies, skills and tools of Artificial Intelligence will be more accessible and will therefore encourage attackers to significantly increase the volume of attacks carried out on the Internet. In the long run, we anticipate the development of new AI algorithms that can make decisions on their own and therefore be able to change the intensity and type of attack without any instructions from the attacker.

In the very context of security, hypothetically speaking, this development can bring positive but also negative consequences. On the one hand, we cannot say that the application of machine learning also cannot improve safety in some areas. For example, machine learning is increasingly used in the field of information security to identify threats and vulnerabilities, detect malware and spam, as well as develop risk management tools. Machine learning can also help protect critical infrastructure and detect vulnerabilities in industrial systems. On the other hand, advances in machine learning can lead to new challenges in the field of safety. The development of Generative Adversarial Networks (AI) technologies can lead to the development of new forms of fraud and abuse. Machine learning algorithms can also be misused to manipulate data and make unfair decisions based on discriminatory data. It is therefore very important to develop and implement protection and safety measures in the context of machine learning at the same time, such as validation and verification of algorithms, ethical risk assessment and the use of data protection technologies. It is also important to develop new solutions that can prevent and respond to new threats in this context.

However, as this paper proves, it should be borne in mind that AI-based protection technologies, such as detection and protection against malicious algorithms, will have to be developed at the same time. Also, a major role is played by the development of ethical guidelines and regulatory frameworks for the use of AI in order to prevent attacks and protect data that will help reduce the risk of malicious Attacks based on AI. And if it can be expected that the number of attacks will increase due to the development of AI technologies, at the same time, protection measures are being developed to help prevent such attacks. In order to reduce the possibility of abuse, it is necessary, first of all, to establish security measures and restrictions when using the tool itself. This includes authenticating users, limiting the number of requests over a period of time, and ensuring that the tool does not generate inappropriate content or false information. It is also important to ensure data security and prevent unauthorized access and data theft. The next link that will reduce abuse is the development of ethical guidelines and regulatory frameworks for the use of artificial intelligence (AI) to protect data and prevent malicious Attacks based on AI is a key step in the fight against cyberattacks. Ethics guidelines and regulatory frameworks for the use of AI are directed towards ensuring that AI tools are used responsibly and in accordance with legal regulations. This includes providing data management policies, ensuring data privacy, and complying with the regulatory framework. Currently, there are various initiatives to develop ethical guidelines and regulatory frameworks for the use of AI, such as:

1. UNESCO's Ethics Code initiative for artificial intelligence, which aims to develop a global ethical framework for the use of AI;
2. Initiative of the European Commission to develop ethical guidelines for the development and use of AI;
3. The OpenAI GPT-3 tool, which has built-in ethical guidelines for use to prevent malicious attacks;
4. Regulatory frameworks in countries such as the US, China and the EU relating to the use of AI in areas such as security and data privacy.

The development of ethical guidelines and regulatory frameworks for the use of AI is an important step in preventing malicious attacks and protecting data. However, it is important to emphasize that this development will be a continuous process, as AI technologies and their applications are constantly evolving and changing.

References

Bernstein, D. J. (2006). Curve25519: new Diffie-Hellman speed records. In *Public Key Cryptography-PKC 2006: 9th International Conference on Theory and Practice in Public-Key Cryptography*, New York, NY, USA, April 24-26, 2006. Proceedings 9 (pp. 207-228). Springer Berlin Heidelberg. https://doi.org/10.1007/11745853_14

Bushard, B. (2023, January 10). Fake Scientific Abstracts Written by ChatGPT Fooled Scientists, Study Finds. Forbes. https://www.forbes.com/sites/brianbushard/2023/01/10/fake-scientific-abstracts-written-by-chatgpt-fooled-scientists-study-finds/?sh=1bf4ae518b63

Eslahi, M., Salleh, R., & Anuar, N. B. (2012). Bots and botnets: An overview of characteristics, detection and challenges. In *2012 IEEE International Conference on Control System, Computing and Engineering* (pp. 349-354). IEEE. https://doi.org/10.1109/iccsce.2012.6487169

Eurostat. (2022, December 22). *ICT security in enterprises.* https://ec.europa.eu/eurostat/statistics-explained/index.php?title=ICT_security_in_enterprises#ICT_security_measures

Florida, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1

Gutmann, P. (2007). The commercial malware industry. In *DEFCON conference*. Las Vegas, USA.

Kocic, B. (2022). Healthcare financing through the development of modern information systems. *International Scientific Conference UNITECH 2022* (pp. 169-171). Technical University of Gabrovo, Bulgaria.

Kulesh, S. (2023, January 5). *Why ChatGPT can be dangerous for every internet user*. Times of India. https://timesofindia.indiatimes.com/gadgets-news/why-chatgpt-can-be-dangerous-to-every-internet-user/articleshow/96393104.cms

Nasteski, V. (2017). An overview of the supervised machine learning methods. *Horizons. b*, *4*, 51-62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*. https://doi.org/10.48550/arXiv.2203.02155

Sharma, R., Sharma, N., & Mangla, M. (2021, May). An analysis and investigation of infostealers attacks during COVID'19: a case study. In *2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC)* (pp. 443-449). IEEE. https://doi.org/10.1109/ICSCCC51823.2021.9478163

Sun, X., & Rajput, S. (2006). Contemporary Malware Trends and Countermeasures. *In High-Capacity Optical Networks & Enabling Technologies (HONET)*. Charlotte, North Carolina.

The Guardian (2022, December 22). *The Guardian view on ChatGPT: an eerily good human impersonator*. https://www.theguardian.com/commentisfree/2022/dec/08/the-guardian-view-on-chatgpt-an-eerily-good-human-impersonator

Tung, L. (2023, January 26). *ChatGPT can write code. Now researchers say it's good at fixing bugs, too.* https://www.zdnet.com/article/chatgpt-can-write-code-now-researchers-say-its-good-at-fixing-bugs-too/

Ubavić, V. S., Bogdanović, B. P., & Milićević, V. J. (2014). Zloupotrebe elektronske pošte. *Bizinfo (Blace)*, *5*(2), 57-65. https://doi.org/10.5937/Bizinfo1402057U

Vučković, Z., Vukmirović, D., Milenković, M. J., Ristić, S., & Prljić, K. (2018). Analyzing of e-commerce user behavior to detect identity theft. *Physica A: Statistical Mechanics and its Applications*, *511*, 331-335. https://doi.org/10.1016/j.physa.2018.07.059