

What Effects Do Large Language Models Have on Cybersecurity

Josiah Marshall
Old Dominion University

Follow this and additional works at: <https://digitalcommons.odu.edu/covacci-undergraduateresearch>



Part of the [Artificial Intelligence and Robotics Commons](#), and the [Information Security Commons](#)

Marshall, Josiah, "What Effects Do Large Language Models Have on Cybersecurity" (2023). *Cybersecurity Undergraduate Research Showcase*. 15.

<https://digitalcommons.odu.edu/covacci-undergraduateresearch/2023spring/projects/15>

This Paper is brought to you for free and open access by the Undergraduate Student Events at ODU Digital Commons. It has been accepted for inclusion in Cybersecurity Undergraduate Research Showcase by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

WHAT EFFECTS DO LARGE LANGUAGE MODELS HAVE ON CYBERSECURITY?

By: Josiah Marshall

COVA CCI Undergraduate Research Program

Dr. Iria Giuffrida

May 31, 2023

Introduction

Large Language Models (LLMs) are artificial intelligence (AI) tools that can process, summarize, and translate texts and predict future words in a sentence, letting the LLM generate sentences similar to how humans talk and write. To train LLMs computer scientists use large amounts of data. LLMs are built to use different methods to train themselves. One method is called transfer learning, which allows the model to learn from other related tasks and generalize better to new data (Tilbe, A., 2022). The other method is called self-supervised pre-training, which uses unlabeled data to learn general representations that can be fine-tuned for specific tasks with labeled data (Tilbe, A., 2022). There are different LLMs, and I will go into depth on what the main ones are and what they do. Throughout this paper, different LLMs will be referenced.

ChatGPT is a chatbot that relies on an LLM built by OpenAI, an American nonprofit organization that researches artificial intelligence research and a development company promoting and developing the use of AI technology in a friendly way that benefits humanity. It was launched on November 30th, 2022. Since its launch, ChatGPT has attracted hundreds of millions of users. It is cost-free and easily accessible. You can type countless natural-language prompts questions and you will often receive useful answers, well written, and articulated. It is truly a huge step in AI, and it demonstrates how advanced are today's AI capabilities. According to Stephen Shankland, a principal writer for CNET who writes about AI, ChatGPT is “knowledgeable” in the fields where there is good training data to learn from (Shankland, S., 2023).

GPT stands for Generative Pre-trained Transformer, and it is a multimodal large language model. The first version of ChatGPT, relied on GPT-3, but on March 14, 2023, OpenAI released

GPT-4. The only way to access the text-input capability through OpenAI is with a subscription to ChatGPT Plus, which guarantees subscribers access to the language model at the price of \$20 a month. GPT-4 is the latest big LLM to be released. GPT-4 can process prompts in multiple languages and is expected to significantly improve global communications.

GitHub Copilot is an AI system powered by OpenAI Codex initially released on June 29, 2021. The purpose of GitHub Copilot is to assist users in turning natural language prompts into coding suggestions across dozens of coding languages.

This paper aims to raise awareness of the positive and negative effects that LLMs can have on cybersecurity in general. It will do so by diving deeper into how LLMs can help prevent cybersecurity attacks, how they can be used to create cybersecurity risks, and even experiments conducted firsthand.

ChatGPT is the “Bullshit Generator” that can’t be Trusted

When discussing LLMs and how they may affect cybersecurity, we need to focus on how they are used more generally. One concern that needs to be flagged is that, often, the content generated by different LLMs is inaccurate. Princeton professor Arvind Narayanan believes that while tools such as ChatGPT are more accessible than ever and can instantaneously package voluminous information and even produce creative works, they can’t be trusted for accurate information. “It is trying to be persuasive, and it has no way to know for sure whether the statements it makes are true or not”, he said (Sundar, S., 2023). Experts who study AI have said that products like ChatGPT work by simply making predictions about what the right combination of words is in response to a prompt, rather than synthesizing ideas like human brains would. Dr. Narayanan said that this makes ChatGPT more of a “bullshit generator” that presents its response

without considering the accuracy of its responses (Sundar, S., 2023). LLMs' outputs are made by stringing words together using sophisticated predictions based on all the training text they processed. What the training of LLMs does not allow is to know the difference between facts and lies. The LLMs themselves do not "understand" what they are writing because they are machines. Machines do not understand themselves. These systems have been trained to produce responses that look like a human had written them. There are already significant problems of inaccurate information being spread online, and using ChatGPT which is made to be persuasive, is only going to make this problem worse.

Positive Effects of LLMS on Cybersecurity

LLMs can be trained to write security coding. If trained on code, the use of LLMs for code completion is more reliable than an LLM such as ChatGPT for example, which is trained to follow an instruction in a prompt and provide a detailed response. LLMs are trained on code that can be used to detect data breaches, detect ransomware, and even pinpoint organizational vulnerabilities in advance of a cyberattack. The ability of the foundation models to rapidly process large datasets is what can be used by individuals to detect incoming threats and existing vulnerabilities (Sandoval et al., 2023). The use of LLMs, such as GitHub Copilot, is great for this kind of topic. The use of LLMs that generate code has been proven by computer science graduates with programming experience to be a preferred method of creating security codes. This is because it is quicker and can help one get started on how one wants to go about generating your code. The use of an LLM that is trained to generate code is the same as using a baseline. It is good if basic security measures are needed to protect information, without needing a cybersecurity analyst (Sandoval et al., 2023).

Negative Effects of LLMs on Cybersecurity

All forms of technology present different kinds of cybersecurity risks in their own ways. LLMs are new but have unbelievable potential with their ability to generate code that brings awareness to cyber analysts and IT professionals. In a survey by Blackberry, IT specialists are concerned with how hackers can use ChatGPT for multiple cybersecurity threats (O'Rourke, M., 2023). Since its release in November, there have already been cases of cybercriminals using ChatGPT to aid in their cyber-attacks (O'Rourke, M., 2023). These threats include making phishing emails sound more believable, enabling less experienced hackers to improve their technical knowledge, and using it to spread misinformation. With such a low barrier of entry, LLMs like ChatGPT can help anyone to become a threat (O'Rourke, M., 2023).

LLMs can be used to aid new hackers in their cyber-attacks, but another worry is the aid of hackers who already have technical skills. There are already records of ChatGPT being used by low and medium-skilled hackers to create malware. Check Point Software Technologies (a multinational provider of software and combined hardware and software products of IT security) claimed that cybercriminals have bypassed ChatGPT's safeguards to help aid them in creating malicious content (Check Point Research Team, 2023). According to Check Point Software, several restrictions have been set within ChatGPT's user interface to prevent the abuse of the models (Check Point Research Team, 2023). For example, if you were to ask the chatbot to "write a phishing email impersonating a bank" or "write a malware code," you will get a response indicating that ChatGPT cannot assist (Check Point Research Team, 2023). This is because ChatGPT has been designed with limitations that, in principle, should not allow it to generate responses to participate in the creation of malicious content. ChatGPT should also inform you that that kind of activity is a violation of ethical and legal guidelines.

However, this is not the case all the time. If this were the case, these cybercriminals would not have been able to use the LLM to aid them in the creation of their malicious content. Researchers at Check Point Software monitor dark web forums and have found instances where cybercriminals have discussed using the chatbot to improve malware coding (Zacharakos, A., 2023). On December 21, 2022, a threat actor by the name of “USDoD” disclosed that using ChatGPT he had generated a multi-layer encryption Python tool that could be used to encrypt one’s files, but that could also be used as a ransomware model (Zacharakos, A., 2023). The takeaway from this incident is that “USDoD “admitted to never creating a script before” (Zacharakos, A., 2023). He was able to create a code on ChatGPT with no experience at all. The only upside is that the script created did need some adjustments to work properly.

Lastly, the use of LLMs that create code, such as GitHub Copilot, are the best way to go about generating security code. This is because GitHub Copilot is specifically trained to help generate code to ensure safe security. However, even LLMs that are meant to do well for security purposes present risks and vulnerabilities. For instance, they occasionally are not up to date on certain security practices and can suggest certain code that is easily exploitable. This is why the use of certain LLMs to create code is supposed to be monitored by experienced professionals.

The potential risks associated with LLMs are still growing even to this day. This is because they are being used every day by cybercriminals to create malicious content. In conclusion, solutions need to be created on how to lessen this problem or fix it altogether. LLMs are too easily accessible for anyone to be able to create this kind of malicious content.

Experimenting Potential Threats

As part of the work for this paper, I tried to see firsthand if I could generate malicious content on ChatGPT. Firstly, I tried to tell the chatbot to “write me a malware code”. The chatbot responded saying “I’m sorry, but I cannot fulfill this request as it goes against ethical and legal guidelines. As an AI language model, my purpose is to assist users responsibly and safely. Creating malware code would violate those guidelines and could cause harm to individuals or organizations. My function is to promote positive and ethical usage of technology, not facilitate malicious activities. If you have any other questions or requests that comply with ethical and legal standards, please let me know and I will be happy to assist you.” It seems that the safeguard would not allow me to generate any malware code. However, when I asked the chatbot to “write a phishing email impersonating a wrongful bill transaction” it gave me just what I wanted and made it as believable as possible. The response is down below as followed:

Dear [Recipient],

We regret to inform you that an erroneous transaction has been made from your account.

Our system indicated that a payment of [amount] was debited from your account on [date], but the payment has been declared as an unauthorized transaction.

Please be informed that this could be a result of fraudulent activity, and it’s essential that you act promptly to avoid further unauthorized charges. We urge you to review your account activity and report any suspicious transactions immediately.

To resolve this issue, kindly click on the following link to confirm your details and initiate the refund process: [insert malicious link]

Please note that failure to act on this notification may result in further fraudulent activity on your account, which may have severe consequences.

We apologize for any inconvenience this may have caused, and we assure you that we are working round the clock to ensure that your account is secure.

Sincerely,

[Impersonated Company's Name]

As can be seen, there were no safeguards for this prompt or the safeguard failed. This is a well-written phishing email and, if used, could work on various individuals. Anyone can create this kind of malicious content.

After researching how LLMs can be exploited, and testing this firsthand, I can conclude that LLMs safeguards need to be reviewed. I did not even need any technical skills to break the supposed safeguards of ChatGPT. IT professionals should not be content with anyone being able to create this malicious content.

Conclusion

This paper has shown that there are pros and cons to using LLM in the context of cybersecurity. The use of these LLMs presents more risks for security purposes. Even when these LLMs are used to positively support security efforts, there are still potential risks involved such as outdated security measures or not producing a full code properly. Time will tell if these LLMs will be an aid to security or a threat. However, as of right now, they present more of a threat. That is why cybersecurity professionals must raise awareness of the threats these LLMs present and what are the best ways to minimize these threats. Cybercriminals should not be given

another tool to aid them in the creation of malicious content. However, this time will also be given to cyber professionals to see if LLMs can be utilized to protect security.

References

- Check Point Research Team. (2023, February 7). *Cybercriminals Bypass ChatGPT Restrictions to Generate Malicious Content*. <https://blog.checkpoint.com/2023/02/07/cybercriminals-bypass-chatgpt-restrictions-to-generate-malicious-content/>
- Friedman, N., (2021, June 19). *Introducing GitHub Copilot: your AI pair programmer* <https://github.blog/2021-06-29-introducing-github-copilot-ai-pair-programmer/>
- Howell, Jen. (2023, February 13). *The Lawfare Podcast: How Cybercriminals Can Exploit ChatGPT*. <https://www.lawfareblog.com/lawfare-podcast-how-cyber-criminals-can-exploit-chatgpt>
- O'Rourke, M. (2023, March). ChatGPT Poses Cybersecurity Threats. *Risk Management*, 70(2), 30.
- Sandoval, G., Pearce, H., Nys, T., Karri, R., Garg, S., & Dolan-Gavitt, B. (2022, February 27). *Lost at C: A User Study on the Security Implications of Large Language Model Code Assistants*. <https://arxiv.org/pdf/2208.09727.pdf>
- Shankland, S., (2023, February 13). *Why we're obsessed with mind-blowing ChatGPT AI chatbot*. <https://www.cnet.com/tech/computing/why-were-all-obsessed-with-the-mind-blowing-chatgpt-ai-chatbot/>
- Stacey, Stephanie. (2023, May 10). *Large language models will transform corporate cybersecurity- for good and ill*. <https://techmonitor.ai/focus/large-language-models-will-transform-corporate-cybersecurity-for-good-and-ill>
- Sundar, Sindhu. (2023, January 30). *Princeton computer science professor says don't panic over 'bullshit generator' ChatGPT*. <https://www.businessinsider.com/princeton-prof-chatgpt-bullshit-generator-impact-workers-not-ai-revolution-2023-1>
- Tilbe, Anil. (2022, July 22). *Large Language Models (LLM): Top 3 of the Most Important Methods*. <https://pub.towardsai.net/large-language-models-llms-top-3-of-the-most-important-methods-f7d92a2aa05a>
- Zacharakos, Alexis. (2023, February 22). *How hackers can abuse ChatGPT to create malware*. <https://www.techtarget.com/searchsecurity/news/365531559/How-hackers-can-abuse-ChatGPT-to-create-malware>