Large Language Models as a Cyber Threat

Malte Josten
Distributed Systems Group
University of Duisburg-Essen
Duisburg, Germany
malte.josten@uni-due.de

Large Language Models (LLMs) are advanced AI systems that use deep learning, particularly transformer-based architectures, to understand and generate human language. They are characterized by their ability to process vast datasets, such as GPT-4's training on approximately 13 trillion tokens, and their massive parameter counts, with GPT-4 featuring 1.8 trillion parameters. These capabilities enable applications such as conversational agents, text summarization, content generation, educational tools, and programming assistance. Recent breakthroughs in LLM technology include notable releases that push the boundaries of AI. GPT-4o introduced improved multimodal capabilities, enhanced reasoning, longer context windows, and greater creativity. While LLMs offer significant benefits, they also pose risks due to intended misuse through prompt engineering or by retraining existing models to remove previously trained guidelines and restrictions. Adversaries utilize the strength of publicly available (unhinged) LLMs for various malicious applications including malware obfuscation, aiding in writing code (WormGPT) or malicious texts, like phishing (FraudGPT) or spam email [1, 2]. They can also be used for social engineering attacks by creating realistic emails, websites, and scripts. Alongside evidence from other studies [3, 4, 5, 6], highlights the persistent threat posed by spam and phishing and underscores the risks associated with malicious LLM usage. These findings point to an urgent need for the development of more effective countermeasures to mitigate such threats. To keep pace with the ever-growing landscape of attack vectors, and to defend against LLM-aided attacks, we want to answer the following research questions:

How can we show and quantify the potential risks posed by LLM-modified input data in text-based security systems? We have already demonstrated that generating LLM-modified spam emails is relatively easy, very cost-effective (0.17 cents per email), and surprisingly effective (70% misclassification rate after modification) [7]. Currently, this holds true for modifying spam emails taken from the *SpamAssassin Public Spam Corpus* [8], with the aid of ChatGPT 3.5 Turbo, and examining the robustness of the default configuration of the spam filter SpamAssassin[1]. The work done by [9, 10, 11, 4, 5, 6] also demonstrate that the misuse of LLMs is a general issue affecting various text-based security mechanisms, such as fraudulent online activities, social engineering, or misuse of social media - and not only in the specific test case we examined. Therefore, we intend to orient our research to be applicable to a broader range of use cases.

Which metrics and processes help to analyze the modifications and their impact? By analyzing the email bodies before and after modifications, we expect to gain valuable insights into what has been changed - whether individual words, sentences, or entire paragraphs - why these parts were altered, and how these changes influenced the results. Based on these observations, we aim to establish a knowledge base that will enable us to reason about what caused the filter to misclassify the modified messages and identify potential weaknesses and shortcomings.

How can we improve the targeted security mechanisms? Clearly, current security mechanisms are not sufficiently equipped to efficiently handle LLM-modified spam messages. Therefore, it is necessary to improve and enhance them to increase their robustness against LLM-led attacks. We aim to determine whether 'simple' methods, such as, for example, retraining a spam filter, are feasible and effective enough to counteract LLM-modified spam, or if more sophisticated adaptations and improvements are required.

[1] https://spamassassin.apache.org/

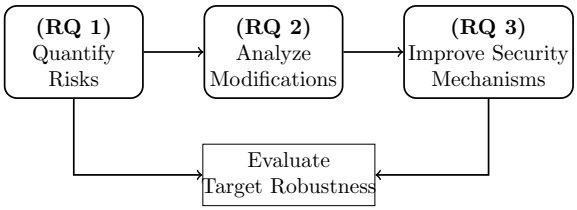

Figure 1: The proposed pipeline with its building blocks to answer the different research questions, showing how they build upon each other and when the target system's security is evaluated.

In the end, we aim to establish a comprehensive testing pipeline, depicted in Figure 1, which examines the initial robustness of a spam filter against LLM-modified spam, analyzes the modifications made by the LLM, and provides feedback that can be utilized, if necessary, to enhance the target system. Initially, it is designed for spam and phishing filters but will be generalized to be applicable for a broader range of applications reliant on text-based security - not limited to just systems related to spam or phishing.

In the future we aim to:

1. Investigate upcoming advancements in LLM technology and their implications.
2. Assess the resilience of existing security systems to (novel) LLM-supported threats.
3. Design (universal) countermeasures and detection mechanisms to mitigate LLM misuse and LLM-supported attacks.

References

[1] T. S. Dutta, "Hackers use wormgpt to launch sophisticated cyberattacks," *Cyber Security News*, 2023. [Online]. Available: https://cybersecuritynews.com/wormgpt-ai-tool/

[2] ——, "Fraudgpt: New black hat ai tool launched by cybercriminals," *Cyber Security News*, 2023. [Online]. Available: https://cybersecuritynews.com/fraudgpt-new-black-hat-ai-tool/

[3] Federal Bureau of Investigation, "Internet Crime Report 2023," FBI, Tech. Rep., Dec 2023.

[4] S. Gallagher, B. Gelman, S. Taoufiq, T. Vörös, Y. Lee, A. Kyadige, and S. Bergeron, *Phishing and Social Engineering in the Age of LLMs*. Cham: Springer Nature Switzerland, 2024, p. 81–86. [Online]. Available: https://doi.org/10.1007/978-3-031-54827-7_8

[5] S. Kreps, R. M. McCain, and M. Brundage, "All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation," *Journal of Experimental Political Science*, vol. 9, no. 1, p. 104–117, Mar. 2022.

[6] P. V. Falade, "Decoding the threat landscape: Chatgpt, fraudgpt, and wormgpt in social engineering attacks," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, p. 185–198, Oct. 2023, arXiv:2310.05595 [cs].

[7] M. Josten and T. Weis, "Investigating the Effectiveness of Bayesian Spam Filters in Detecting LLM-modified Spam Mails," 2024. [Online]. Available: https://arxiv.org/abs/2408.14293

[8] J. Mason, "SpamAssassin Public Mail Corpus," https://spamassassin.apache.org/old/publiccorpus/, Version Jan 31 2006.

[9] Z. Alkhalil, C. Hewage, L. Nawaf, and I. Khan, "Phishing attacks: A recent comprehensive study and a new anatomy," *Frontiers in Computer Science*, vol. 3, 2021. [Online]. Available: https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2021.563060

[10] D. Lain, K. Kostiainen, and S. Čapkun, "Phishing in organizations: Findings from a large-scale and long-term study," in *2022 IEEE Symposium on Security and Privacy (SP)*, 2022, pp. 842–859.

[11] J. A. Teixeira da Silva, A. Al-Khatib, and P. Tsigaris, "Spam emails in academia: Issues and costs," *Scientometrics*, vol. 122, no. 2, pp. 1171–1188, Feb. 2020.