Large Language Models as a Cyber Threat: Towards Countering LLM-based Spam Attacks

Malte Josten
supervised by Torben Weis
Department of Distributed Systems
University of Duisburg-Essen, Duisburg, Germany
malte.josten@uni-due.de

***Abstract*—The rapid advancement and accessibility of large language models (LLMs) have significantly enhanced their utility across various disciplines. However, this pervasive presence of LLMs raises substantial security and trust concerns, particularly at the user-level, within the context of cybersecurity. This paper investigates the potential misuse of LLMs in crafting spam emails that are convincingly legitimate, thus bypassing traditional spam filters. We already conducted an experiment utilizing ChatGPT 3.5 Turbo to alter spam emails, resulting in a 70% success rate of these emails being misclassified as legitimate. We present a pipeline demonstrating the ease with which LLMs can be exploited to undermine email security, highlighting the urgency for improved defensive mechanisms. To counteract these threats, we propose a novel methodology designed to enhance the robustness of spam filters against such sophisticated attacks. This methodology comprises three phases: assessing the vulnerability of current systems to LLM-modified spam, detailed examination of the changes imposed by LLMs, and applying insights gained to fortify existing security infrastructures.**

***Index Terms*—natural language processing, cybersecurity, large language models, spam**

I. Introduction

The emergence and increasing sophistication of large language models (LLMs) have made them both easy and relatively inexpensive to use. This accessibility opens up a vast array of potential applications, ranging from benign to malicious. In addition to other studies and statistics that show the lasting impact and relevance of spam and phishing [1]–[4], the aim of this research is to explore the wider implications of using LLMs in both beneficial and malicious contexts. A key focus of our work is to understand how LLM models, such as ChatGPT 3.5 Turbo[1], may be exploited, particularly in the generation of malicious content, such as spam. We found, that by using ChatGPT 3.5 Turbo to modify spam emails, we could increase the rate at which spam was classified as legitimate by 70%, effectievly bypassing the spam filter [5]. This finding as well as [1], [6]–[8] emphasize the ongoing threat posed by spam and phishing, and the potential dangers of LLMs when used for malicious purposes and highlights the need for more robust countermeasures. Therefore, we aim to dive deeper into what makes LLMs useful for malevolent intentions, how they can be utilized, and ultimately, based on the acquired knowledge, propose suggestions and guidelines on how to improve targeted security mechanisms. The implications of our research are significant, especially at the user level, where security, trust, and the ability to resist manipulation are at stake.

II. Research Contribution

There are multiple attack vectors and points-of-interest when dealing with LLM-led cyberattacks. For now, our work only examines machine detectable, processable, and supposedly mitigatable attacks, i.e., we do not consider the human factor. We aim to examine the following research questions:

(RQ 1) 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) [5]. Currently, this holds true for modifying spam emails taken from the *SpamAssassin Public Spam Corpus* [9], with the aid of ChatGPT 3.5 Turbo, and examining the robustness of the default configuration of the spam filter SpamAssassin[2]. The work done by [2]–[4], [6]–[8] 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.

(RQ 2) 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.

(RQ 3) Using the knowledge about the modified data, how can we improve the targeted security mechanisms? Clearly, current security mechanisms are not sufficiently equipped to efficiently handle LLM-modified spam messages.

[1]https://platform.openai.com/docs/models/gpt-3-5-turbo

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

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 retraining a spam filter, are feasible and effective enough to counteract LLM-modified spam, or if more sophisticated adaptations and improvements are required.

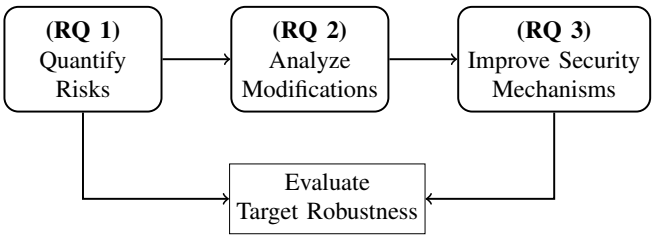

Figure 1. The proposed pipeline with its building blocks to answer the different research questions, showing how they build upon each other and when the target system's security is evaluated.

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.

III. Methodology

To evaluate the security of a target, particularly its robustness against inputs modified by an LLM, we begin by collecting a base dataset, for example, a set of spam emails. This dataset is then transformed or rephrased with the assistance of an LLM. Following this, we observe and analyze how the target behaves when exposed to both the original and the LLM-modified datasets. The final step involves comparing the target's performance across these datasets to determine its robustness against the modifications made by the LLM. Effectively, this yields a metric that enables us to quantify the risk posed by LLM-modified input to various targets, allowing for an empirical comparison.

To address and thoroughly analyze the modifications to the original data set and their effects, we outline a series of methods:

- Evaluating the semantic similarity, by, for example, utilizing the cosine similarity, to assess how closely related the meanings of different text segments (or the entire text) are.
- Topic Analysis to identify which topics are more or less susceptible to detection.

Lastly, we plan to compare various LLMs in terms of each model's adherence to ethical guidelines and employed safeguards to understand which models can be exploited the easiest, thus, posing the biggest threat.

To enhance the robustness of the target against attacks that utilize inputs revised by LLMs, and thereby improve its overall security, we plan to propose a multifaceted strategy, depending on the actual use case. We assume that a newly generated dataset containing the previously misclassified inputs and potentially augmented with synthetic data improves the overall detection rates when used to retrain the target system. Additionally, we want to provide administrators with a comprehensive evaluation and analysis overview. This documentation will equip them with a deeper understanding of their system's vulnerabilities, enabling them to independently devise improvements based on the insights we offer.

IV. Conclusion

This paper elucidates the threat posed by LLM-modified spam emails and describes a pipeline that utilizes these emails with the aim of enhancing existing text-based security mechanisms, such as spam filters. Addressing the pipeline, we propose a method that comprises three main parts: (1) testing the robustness of the target system against LLM-modified spam emails, (2) analyzing the modifications made by the LLM, and (3) using the findings from the analysis to either propose actions for improving the system's robustness or to directly modify the system. More generally, the impact of LLMs on text-based security systems must be further examined. With the help of our research and future work, we set out to analyze and thoroughly understand the effects on cybersecurity, as well as propose countermeasures to make potential targets more robust and resilient.

Acknowledgement

I would like to thank my supervisor Torben Weis for his guidance and my colleagues for their continuous support.

References

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

[2] 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

[3] 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.

[4] 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.

[5] 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

[6] 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

[7] 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.

[8] 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].

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