Enhancing Phishing Email Detection with Context-Augmented Open-Source Large Language Models

Fabian Nicklas[1], Nicolas Ventulett[2], and Prof. Dr.-Ing. Jan Conrad[3]

[1] University of Applied Sciences Kaiserslautern
fani1001@stud.hs-kl.de
[2] University of Applied Sciences Kaiserslautern
nive1002@stud.hs-kl.de
[3] University of Applied Sciences Kaiserslautern
jan.conrad@hs-kl.de

Abstract. Large Language Models offer a promising approach to improving phishing detection through advanced natural language processing. This paper evaluates the effectiveness of context-augmented open-source LLMs in identifying phishing emails. An approach was developed that combines the methods of Few-Shot Learning and Retrieval-Augmented Generation (RAG) to significantly improve the performance of LLMs in this area. On this basis, it has been shown that the presented approach can significantly improve the recognition rate even for smaller models.

Keywords: Artificial Intelligence, AI, Cybersecurity, Large Language Models

1 Introduction

Phishing attacks are a major threat to cybersecurity, using evolving techniques to trick individuals into revealing sensitive information. With estimated 90% of successful cyber attacks starting with phishing [1], robust detection mechanisms are crucial. Large Language Models (LLMs), such as OpenAI's GPT [2], have revolutionised NLP by using large text corpora to perform tasks beyond text generation. This makes them suitable for detecting phishing emails. This paper presents a promising approach that combines Few-Shot Learning and Retrieval-Augmented Generation (RAG) with open source LLMs to improve the detection of phishing emails. The focus on open source has the advantage that the LLM can be operated in its own isolated network. This can significantly increase the protection of confidential email content.

2 Related Work

Previous studies have investigated simple classification approaches for phishing detection using LLMs [3][4][5][6]. These studies showed that pre-trained models could detect phishing attempts, but focused primarily on commercial models such as GPT. However, there have also been attempts using open source networks. For example, Koide et al. achieved an accuracy of 88.61% using the open source LLM Llama2 [5].

3 Methodology

A balanced dataset was created using legitimate emails from the CSDMC Spam Corpus [7]and phishing emails from the Phishing Pot dataset [8], resulting in 5,800 emails equally divided between phishing and non-phishing. On this basis, several open source LLMs were evaluated, including Phi-3 3.8B (Microsoft Research), OpenChat 7B, Mixtral 8x7B, Mistral 7B, Gemma 7B (Google Deep Mind) and Llama3 (8B and 70B) from Meta AI. As a phishing detection approach, two specific prompts were designed for evaluation. The first prompt followed a persona pattern and instructed the LLM to identify phishing emails. The second prompt included a list of indicators suggesting phishing attempts. The proposed approach extends the context using Few-Shot Learning and RAG. Relevant examples are retrieved from a knowledge base and conditioned with the LLM prior to generation, thereby improving domain-specific performance without additional fine-tuning.

4 Experiments and Results

The performance metrics considered were precision, recall, F1 score and accuracy, with 121,800 classifications performed across different models and settings. The results demonstrate the variability in model performance of different models, which is influenced by architecture and training data. Larger models, such as Llama3 70B, consistently outperformed smaller counterparts. Prompt 2 improved recognition rates for most models, particularly those close to 50% accuracy with Prompt 1. The contextual RAG approach outperformed both prompts for all models, with Llama3 70B achieving the highest accuracy of 95.71%. Notably, the performance of Llama3 8B improved significantly from 51.29% and 66.79% accuracy to 91.47%.

5 Conclusion and Future Work

This study demonstrates that LLMs can effectively distinguish between legitimate and phishing emails. The proposed approach, combining Few-Shot Learning and RAG, significantly improves detection rates, particularly for smaller models. Future research could explore integrating additional datasets and embedding models to further enhance detection accuracy.

References

1. Cloudflare: Bericht zu phishing-bedrohungen (2023) München, Tech. Rep., 2024. [Online]. Available: https://www.cloudflare.com/de-de/lp/ 2023-phishing-report/.
2. J. Achiam, S.A.e.a.: Gpt-4 technical report (2024)
3. D. Nahmias, G. Engelberg, D.K., Shabtai, A.: Prompted contextual vectors for spear-phishing detection (2024)
4. Reynolds, L., McDonell, K.: Prompt programming for large language models: Beyond the few-shot paradigm (2021) CHI EA '21.
5. T. Koide, N. Fukushi, H.N., Chiba, D.: Chatspamdetector: Leveraging large language models for effective phishing email detection (2024)
6. H. Patel, U.R., Iqbal, F.: Large language models spot phishing emails with surprising accuracy: A comparative analysis of performance (2024)
7. N.A.: International conference on neural information processing (2010) Csdmc2010 spam corpus.
8. N.A.: Phishing pot dataset (2024) https://github.com/rf-peixoto/phishing pot.