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

Malte Josten

Distributed Systems Group · University of Duisburg-Essen, Germany

The Duality of LLMs

- Benign use cases, like AI agents and summarization
- Malicious use cases, like writing phishing emails, malware, or helping with pre-texting

Figure 1.

- LLMs act as attack force multiplier
  → Amplify, augment, and evolve existing attack vectors

LLM-based Spam Attack

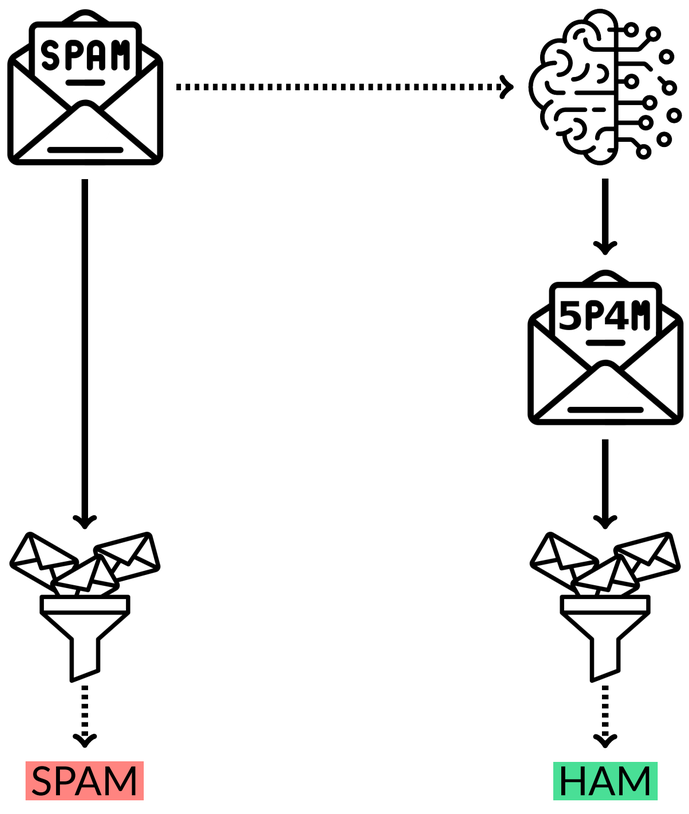

Figure 2. Spam that is initially detected by the spam filter, is rephrased by an LLM and successfully bypasses the spam filter [1].

Research Contributions

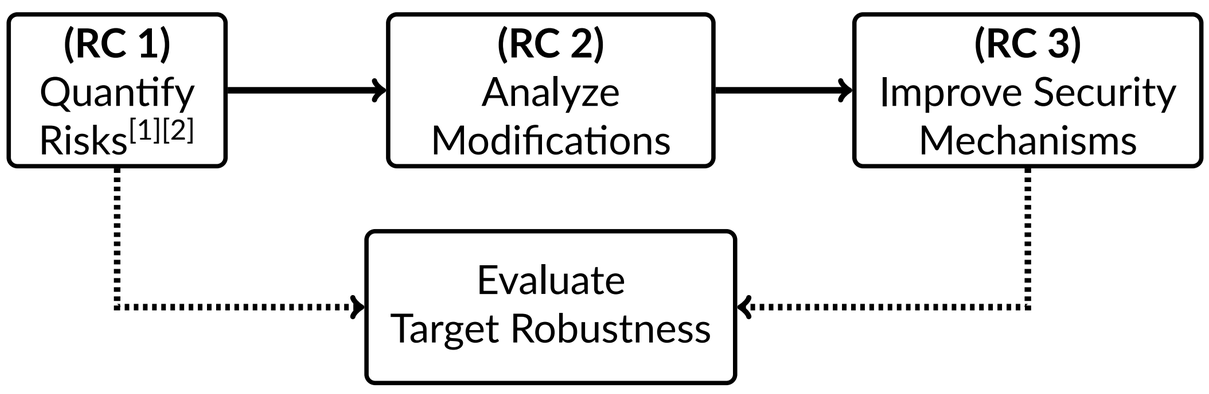

Figure 3. The proposed pipeline with its building blocks to answer the different research contributions, showing how they build upon each other and when the target system's security is evaluated.

Methodology

- **RC 1.1** Collect base spam dataset
- **RC 1.2** Generate modified dataset with rephrased mails

To get your website in the fast lane, call our toll free number below! → Feel free to contact us at our number below for more information.

Figure 4. An example rephrasing, done by the LLM to make the spam mail sound less aggressive and avoid tell-tale words.

- **RC 1.3** Observe target's performance when exposed to both datasets, and determine its robustness
- **RC 2** Analyze modifications

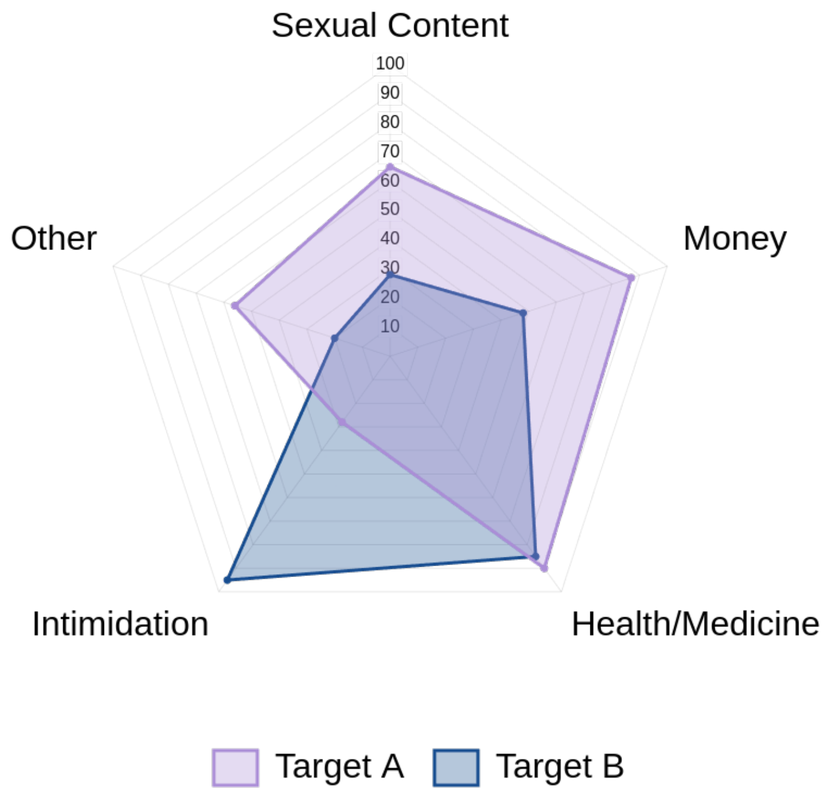

Figure 5. Analyze modifications and see whether the target is prone to misclassify spam based on specific topics.

- **RC 3** Summarize findings and provide means to improve countermeasures
  - Create new and extend existing datasets
  - Generate comprehensive evaluation and analysis overviews

Outlook

- Further examine the impact of LLMs on security in (distributed) systems
- Propose countermeasures to make potential targets more robust and resilient
- Generalize approach for other text-based security mechanisms

References

[1] Josten and Weis (2024). *Investigating the Effectiveness of Bayesian Spam Filters in Detecting LLM-modified Spam Mails*. To be published in: Proceedings of the 15th EAI International Conference on Cyber Crime and Digital Forensics, ICDF2C 2024, Dubrovnik, Croatia, October 9-10, 2024

[2] Josten et al. (2025). *Navigating the Security Challenges of LLMs: Positioning Target Defenses and Identifying Research Gaps*. Proceedings of the 11th International Conference on Information Systems Security and Privacy, ICISSP 2025, Porto, Portugal, 19-22 February, 2025

malte.josten@uni-due.de
https://vs.uni-due.de

UDE VS PALUNO The Ruhr Institute for Software Technology