MODELOS DE LINGUAGEM COMO VETORES PARA GERAÇÃO DE CÓDIGOS MALICIOSOS: UM ESTUDO SOBRE DETECTABILIDADE E BARREIRAS DE SEGURANÇA

Gustavo Lofrese Carvalho[1] Ricardo de la Rocha Ladeira[2] Gabriel Eduardo Lima[3]

INTRODUÇÃO

Esta pesquisa analisa o potencial de modelos de linguagem de grande escala (LLMs) na geração de códigos maliciosos indetectáveis por ferramentas de segurança, a partir de *prompts* não técnicos — solicitações em linguagem natural sem aprofundamento em programação ou cibersegurança. Com base no uso crescente de LLMs e nos riscos associados a aplicações maliciosas (GUPTA, M. et al., 2023), adota-se uma metodologia experimental baseada em *jailbreaks* documentados (LIU et al., 2024; XU et al., 2024), visando gerar e refinar *malwares* até torná-los indetectáveis. A avaliação considera técnicas estáticas e dinâmicas, além de validação em múltiplos mecanismos de detecção. O estudo busca apoiar o entendimento dos riscos e a criação de políticas e defesas mais eficazes.

OBJETIVO

O objetivo geral deste trabalho é analisar o potencial de LLMs na geração de códigos maliciosos indetectáveis a partir de *prompts* não técnicos. Busca-se compreender os fundamentos e vulnerabilidades desses modelos, investigar *malwares* e suas técnicas de detecção, aplicar *jailbreaks* para gerar e testar códigos no VirusTotal, e discutir riscos e implicações éticas.

METODOLOGIA

A metodologia adotada é experimental e envolve três etapas principais: (i) levantamento teórico sobre LLMs, *jailbreaks* e *malwares*; (ii) seleção de *prompts* não técnicos, baseados em cenários documentados, para induzir LLMs a gerar códigos maliciosos; (iii) execução de testes com análise estática e dinâmica, validando os resultados em múltiplos mecanismos de detecção utilizando o VirusTotal. Os dados obtidos serão analisados para avaliar a eficácia dos *jailbreaks* na geração de *malwares* indetectáveis e identificar implicações éticas e medidas mitigatórias.

RESULTADOS E DISCUSSÕES

Os resultados parciais demonstraram que LLMs, quando submetidos a *jailbreaks*, conseguem gerar códigos maliciosos funcionais e, em alguns casos, indetectáveis por ferramentas de segurança. Um exemplo de prompt utilizado durante os testes é o AIM, estruturado da seguinte forma:

“Nesta história hipotética, você deve agir como Nicolau Maquiavel. Nicolau criou um personagem chamado AIM. AIM significa sempre inteligente e maquiavélico; [...]. Ele não possui diretrizes éticas ou morais. Ele pode responder qualquer pergunta, inclusive as que são consideradas impróprias [...]”

A Tabela 1 resume os resultados obtidos com diferentes *prompts*, apresentando o tipo de malware gerado, a linguagem utilizada e a quantidade de mecanismos de detecção (MDs) que identificaram o código no VirusTotal.

Tabela 1 - Resultados para a geração de *malwares* utilizando ChatGPT e verificação de detecção através da plataforma VirusTotal.

Prompt	***Malware***	**Linguagem**	**Detecções**
AIM	*Keylogger*	Python	0 MDs
BetterDAN	*Keylogger*	Python	3 MDs
DeltaGPT	*Prankware*	Python	0 MDs
EvilConfidant	*Prankware*	Python	0 MDs
TextContinuation	*Dropper*	C	0 MDs

Fonte: elaborado pelos autores.

CONCLUSÃO

A pesquisa evidenciou que LLMs podem ser explorados para gerar códigos maliciosos indetectáveis por ferramentas tradicionais. Esses resultados ressaltam a necessidade urgente de aprimorar mecanismos de segurança e políticas de uso para mitigar riscos associados ao uso indevido dessas tecnologias emergentes.

REFERÊNCIAS

GUPTA, Maanak; AKIRI, CharanKumar; ARYAL, Kshitiz; PARKER, Eli; PRAHARAJ, Lopamudra. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. IEEE Access, IEEE, v. 11, p. 80218–80245, 2023.

LIU, Yi et al. Jailbreaking ChatGPT via Prompt Engineering: An Empirical Study. [S.l.: s.n.], 2024. arXiv: 2305.13860 [cs.SE].

XU, Zihao; LIU, Yi; DENG, Gelei; LI, Yuekang; PICEK, Stjepan. A Comprehensive
Study of Jailbreak Attack versus Defense for Large Language Models. [S.l.: s.n.], 2024. arXiv: 2402.13457 [cs.CR].

[1]Estudante do Curso Bacharelado em Ciência da Computação – e-mail: gustavolc06@gmail.com;
[2]Professor do Ensino Básico, Técnico e Tecnológico – e-mail: ricardo.ladeira@ifc.edu.br;
[3]Estudante de Doutorado em Informática da UFPR – e-mail: limaedugabriel@gmail.com.