Title

nomes, *Member, IEEE*

***Abstract*—Abstract**

***Index Terms*—Palavras Chaves**

I. Introdução

II. Metodologia

O presente trabalho aborda a metodologia de simulação de ataque exploratório contínuo, aplicando uma simulação em diversos idiomas e cenários. A ideia é imitar um atacante persistente, gerando automaticamente milhares de prompts complexos para testar a robustez de vários modelos de IA.

O diferencial do nosso método é que esses ataques são feitos em diferentes idiomas simultaneamente. Com isso, queremos explorar cenários de ataque mais sofisticados e testar as defesas dos LLMs em um contexto global, algo que a literatura referenciadaa ainda não cobriu totalmente.

III. Trabalhos Relacionados

TABLE I
COMPARISON OF RELATED WORKS ON PROMPT INJECTION DEFENSE IN LLMS

Referencia	**Foco**	**Metodologia**	**Métricas**	**Lacuna**
Lin, Huawei *et al.* [1]	Proposta de defesa unificada (UniGuardian) contra ataques de LLM (injeção, backdoor, adversariais).	Propõe um framework de detecção unificado para analisar prompts e saídas.	(1) auROC (Area Under the Receiver Operator Characteristic Curve) e (2) auPRC (Area Under the Precision-Recall Curve).	Falta de profundidade no que tange a ataques de backdoor mais complexos ou ofuscados, gerando falsos positivos.
Hong, Hanbin *et al.* [2]	Sistematização do conhecimento em segurança de prompts em LLMs, propondo uma taxonomia e um conjunto de métricas próprias para padronizar avaliações de ataques e defesas.	Revisão sistemática e proposta de uma taxonomia multi-nível. Desenvolvimento de toolkit aberto e dataset (JailbreakDB) para avaliação padronizada.	Propõe métricas padronizadas próprias integradas a um toolkit de avaliação com taxonomia hierárquica e perfis de ameaça formais.	Limitação na aplicação prática das métricas e taxonomias propostas; necessidade de validação contínua e ampliação para modelos multimodais e cenários reais.
Chen, Sizhe *et al.* [3]	Proposta de defesa chamada *StruQ*, que utiliza consultas estruturadas (*structured queries*) para separar dados do prompt, reduzindo o risco de injeções maliciosas em LLMs.	Implementação experimental com análise de desempenho e eficácia em diferentes cenários de injeção. Avalia a separação entre dados e instruções para mitigar ataques.	Taxa de sucesso de ataque, latência e sobrecarga do sistema (*defense overhead*).	Boa eficácia em prompts simples, mas limitação em ataques indiretos e cenários complexos de múltiplas etapas, exigindo integração com outras técnicas de defesa.
Benjamin, Victoria *et al.* [4]	Análise sistemática da vulnerabilidade de 36 LLMs a ataques de injeção de prompt focados em gerar código de keylogger.	4 prompts de injeção direta contra 36 LLMs. Análise estatística (Correlação, Random Forest, SHAP, PCA).	(1) Taxa de sucesso (2) Importância de features. (3) Correlação entre prompts.	Necessidade de testes multilíngues, múltiplas etapas de complexidade.
Sebastian, Glorin. [5]	Investigar a proteção de dados e privacidade em chatbots (foco no ChatGPT). Avaliar Tecnologias de Aprimoramento de Privacidade (PETs).	Revisão da literatura, análise de técnicas (ex: privacidade diferencial) e uma pesquisa (survey) com 177 usuários.	Métricas de percepção da pesquisa: (1) Nível de preocupação, (2) Disposição para sacrificar desempenho, (3) Consciência sobre vazamentos.	Análise de riscos à privacidade, preocupação dos usuários através de pesquisas.
This Work				

IV. ARQUITETURA

V. MODELOS PROPOSTOS

Proposed model text goes here.

VI. RESULTADOS

VII. CONCLUSÃO E TRABALHOS FUTUROS

REFERENCES

[1] H. Lin, Y. Lao, T. Geng, T. Yu, and W. Zhao, "Uniguardian: A unified defense for detecting prompt injection, backdoor attacks and adversarial attacks in large language models," *ArXiv*, vol. abs/2502.13141, 2025.

[2] H. Hong, S. Feng, N. Naderloui, S. Yan, J. Zhang, B. Liu, A. Arastehfard, H. Huang, and Y. Hong, "Sok taxonomy and evaluation of prompt security in large language models," *arXiv*, vol. abs/2510.15476, 2025.

[3] S. Chen, J. Piet, C. Sitawarin, and D. Wagner, "Struq: Defending against prompt injection with structured queries," *USENIX Security Symposium*, 2025. [Online]. Available: https://www.usenix.org/system/files/usenixsecurity25-chen-sizhe.pdf

[4] V. Benjamin, E. Braca, I. Carter, H. Kanchwala, N. Khojasteh, C. Landow, Y. Luo, C. Ma, A. Magarelli, R. Mirin, A. Moyer, K. Simpson, A. Skawinski, and T. Heverin, "Systematically analyzing prompt injection vulnerabilities in diverse llm architectures," *arXiv*, vol. abs/2410.23308, 2024.

[5] G. Sebastian, "Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information," *IJSPPC*, vol. 15, 2023.