XXI Escola Regional de Redes de Computadores / VIII Workshop de Segurança (WRSeg 2024). DOI: 10.5753/errc.2024.4670

Artigo Curto/Short Paper

LLM vs. LLM: uso de red team e AML para avaliar a segurança de LLMs

LLM vs. LLM: using red team and AML to evaluate the security of LLMs

Diego E.G.C.de Oliveira • diego.egcdo@edu.udesc.br
Programa de Pós-Graduação em Computação Aplicada (PPGCAP) /Universidade do Estado de Santa Catarina (UDESC)
Charles C. Miers • charles.miers@udesc.br
Programa de Pós-Graduação em Computação Aplicada (PPGCAP) /Universidade do Estado de Santa Catarina (UDESC)

resumo. A cibersegurança apresenta um desafio complexo no que tange proteção relacionada à Inteligência Artificial (IA). Esta área também propõe segurança no envolvimento de tecnologia, dados e indivíduos. Nesse sentido, envolvida à cibersegurança apresenta-se uma área chamada *Adversarial Machine Learning* (AML) que se aprofunda no estudo e desenvolvimento de ferramentas para proteção de inteligentes sistemas baseados no aprendizado de máquinas. Estudos tem sido realizados no âmbito de AML, muito embora com resultados de pesquisas direcionadas, geralmente, para um ou dois tipos de LLMs. Este artigo apresenta uma proposta para se aprofundar nos três principais LLMs (GPT-4, Google Gemini e LlaMA) e apresentar um método de identificação de ameaças utilizando *Large Language Model*s (LLMs) contra LLMs para soluções baseadas em testes que visam mitigar a exploração de sistemas de maneira maliciosa.

abstract. Cybersecurity presents a complex challenge regarding protection related to Artificial Intelligence (AI). This area also proposes security by involving technology, data, and individuals. In this sense, involved in cybersecurity, there is an area called AML that delves into the study and development of tools to protect intelligent systems based on machine learning. Studies have been carried out within the scope of AML, although research results are generally directed towards one or two types of LLMs. This article presents a proposal to delve deeper into the three main LLMs (GPT-4, Google Gemini, and LlaMA) and present a threat identification method using LLMs against LLMs for test-based solutions that aim to mitigate the exploitation of systems in a malicious manner.

palavras-chave: Red Team • cibersegurança • AML • IA adversarial • grandes modelos de línguas
keywords: Red Team • cybersecurity • AML • adversarial AI • LLM

1 Introdução

A cibersegurança constitui uma disciplina multidisciplinar no campo da computação, cujo propósito é promover a segurança nas operações que envolvem tecnologia, indivíduos e dados. Esta área visa também desenvolver sistemas seguros, levando em conta não apenas os aspectos legais e éticos pertinentes, mas também os desafios apresentados pelo contexto adversarial [1]. A área conhecida como *Adversarial Machine Learning* (AML) [2], se dedica ao estudo e desenvolvimento de ferramentas para a proteção de sistemas inteligentes baseados em aprendizado de máquina. O AML é um campo de pesquisa que integra os domínios de aprendizado de máquina, ciência da computação e cibersegurança. Diversos órgãos e instituições internacionais têm direcionado sua atenção para esta área devido à crescente tendência de adoção de sistemas que utilizam inteligência artificial e aprendizado de máquina, e.g., o NIST [2, 3], ENISA [1, 4] e MITRE [5]. O *Red Teaming* baseado em *Large Language Model* (LLM) pode trazer benefícios. Certas abordagens se destacam pela capacidade de gerar ampla variedade de casos de testes, o que é essencial para garantir uma cobertura abrangente, enquanto outras são mais eficazes na criação de casos de testes complexos, adequados para simular usuários adversariais. De fato, os casos de testes gerados por essas técnicas mostram-se superiores, tanto em diversidade quanto em dificuldade, quando comparados com casos de testes escritos manualmente [6].

O objetivo deste trabalho, em etapa inicial de execução, é propor um método para avaliar a segurança de LLMs utilizando outros LLMs combinando a prática de *Red Teaming* com técnicas de AML. As pesquisas sobre segurança de LLMs com a utilização de AML, de maneira geral, tem sido realizadas com LLMs específicos e, principalmente, com variados modelos testados. Nesta proposta, com intuito de apresentar uma abordagem mais específica aos mais populares, três LLMs serão expostos às vulnerabilidades de segurança a partir de ataques de LLMs com ferramentas *Red Teaming*. São es-

Todos os autores são membros da SBC.

tes o ChatGPT-4, Google Gemini e LlaMA. A proposta visa desenvolver um método que explore vulnerabilidades e ameaças específicas à análise de segurança de modelos, simulando cenários de ataque contra LLMs utilizando outros LLMs. Com os três modelos serão realizados testes de manipulação de entradas, evasão de defesas e extração de informações sensíveis. Este artigo está organizado da forma como como segue. A Seção 2 apresenta os conceitos fundamentais, enquanto a Seção 3 traz a proposta de LLM vs LLM.

2 Fundamentação

A cibersegurança visa proteger infraestruturas e dados contra ameaças cibernéticas considerando aspectos éticos e legais dos sistemas. Seu aspecto principal é ainda mais crítico no contexto de tecnologias emergentes que requerem novos investimentos e esforços de pesquisa e desenvolvimento. O aumento de ameaças geradas por aplicações construídas utilizando a Inteligência Artificial (IA) representa riscos significativos para organizações, governos e indivíduos. A natureza dinâmica, e de rápida evolução da segurança cibernética, traz desafios recorrentes, pois os adversários adaptam continuamente os seus métodos para explorar vulnerabilidades e evitar a sua detecção [7]. Inserido na área de cibersegurança, o *Adversarial Machine Learning* (AML) aborda questões relacionadas à robustez dos modelos de *Machine Learning* (ML) contra ataques adversariais em todo o domínio da cibersegurança. O AML foca no estudo e desenvolvimento de ferramentas para proteger sistemas de ML. O conceito de AML refere-se ao processo de obtenção de informações sobre o comportamento e as características de um sistema de ML e/ou a manipulação das entradas de um sistema de ML com o intuito de alcançar um resultado desejado [2]. O AML busca desenvolver algoritmos de ML que possam resistir a ataques a partir da análise das capacidades destes atacantes, além das consequências destes ataques em sistemas que utilizam ML [2].

A Figura 1 traz uma categorização dos ataques mais comuns elaborada no contexto do aprendizado de máquina adversarial. Esses sistemas não são programados ou orientados para tarefas específicas e aprendem a partir de uma grande massa de dados, o que os torna mais vulneráveis a ataques que podem comprometer a privacidade, a integridade dos dados, bem como sua infraestrutura. Uma expressiva parte destes modelos passou a se basear na arquitetura *Transformer* [8], que consiste em uma sequência de diversas camadas de neurônios artificiais na qual várias camadas implementam um modelo de atenção (que indica a relação de importância entre *tokens*). Esses modelos podem ser utilizados de forma extrativa, concentrando-se na identificação de elementos significativos em um texto de entrada, ou de maneira generativa, na qual, a partir da entrada original, são gerados novos *tokens* que criam sequências de texto expandidas em diferentes contextos. Essa expansão, juntamente com o uso de treinamentos extensivos e técnicas avançadas de pré-treinamento, é crucial em áreas como IA para a ciência, raciocínio lógico e IA incorporada [9]. O AML, se não abordado devidamente, abrirá possibilidades para criação de novas estratégias de ataque que buscam explorar as vulnerabilidades que existem em LLMs. Exemplos adversariais de sucesso devem ser elaborados para cumprir as restrições de domínio e do mundo real e isto pode ser desafiador, pois mesmo pequenas modificações podem impactar consideravelmente um sistema empresarial ou governamental.

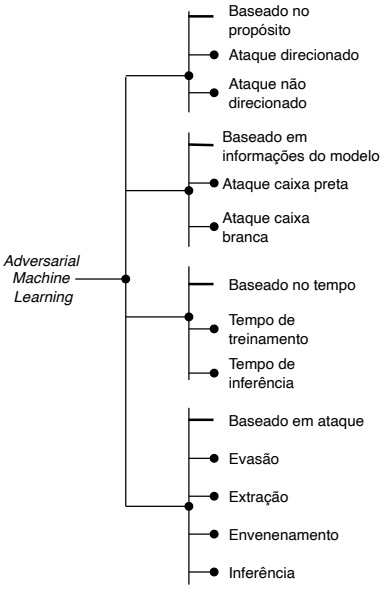

Figura 1. Categorização de ataques mais comuns de AML.

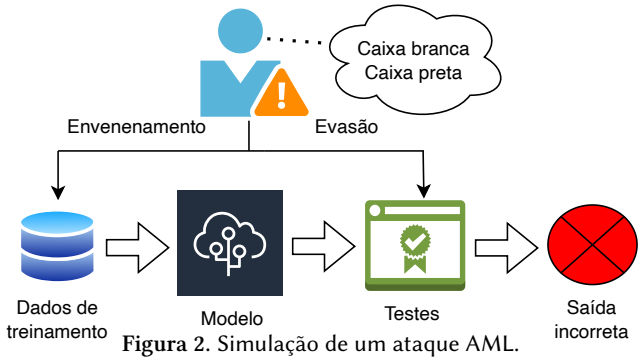

Figura 2. Simulação de um ataque AML.

Os ataques adversariais (Figura 2) referem-se a um conjunto de técnicas e estratégias usadas para intencionalmente manipular ou enganar modelos de ML. Dependendo do conhecimento do invasor, diferentes

tipos de configurações de ataque poderão ser aplicadas, incluindo ataques de caixa branca e de caixa preta. No cenário de caixa branca, o invasor possui conhecimento total sobre o sistema alvo. Os ataques de caixa branca permitem uma análise detalhada do pior caso e são extremamente importantes para avaliar mecanismos de defesa [10]. Os ataques de caixa preta, por outro lado, são os que o atacante não possui (ou possui pouco) conhecimento dos dados de treinamento e/ou modelo alvo e podem ser ataques de transferência ou ataques de consulta. O envenenamento de dados, uma dessas vulnerabilidades, significa que os invasores influenciam o processo de treinamento, injetando dados maliciosos num conjunto de dados de treinamento. Isto pode introduzir vulnerabilidades, comprometendo a segurança, a eficácia ou a ética comportamental dos modelos resultantes. Estudos indicam [11] que modelos pré-treinados são mais vulneráveis e podem ficar comprometidos se atacados por meio de métodos como o uso de dados não confiáveis de pesos ou conteúdo, como a inserção de exemplos envenenados em seus conjuntos de dados. Por sua natureza inerente como modelos pré-treinados, LLMs são mais suscetíveis a ataques de envenenamento de dados. Um importante trabalho de aperfeiçoamento de segurança através de um método automático *red teaming* foi o *Multi-round Automatic Red-Teaming* (MART) [12]. Exemplificando, [13] mostrou que mesmo com apenas 100 exemplos de dados envenenados, os LLMs podem produzir resultados consistentemente negativos ou resultados falhos em várias tarefas. Os LLMs têm a capacidade de gerar respostas textuais coerentes, mas enfrentam desafios como o fenômeno de ”alucinações”, que podem comprometer as respostas geradas. As alucinações são fenômenos em que os modelos de língua, em sua inferência, geram saídas sem sentido, fora de contexto ou conteúdo, normalmente associado ao tamanho do conjunto de parâmetros pertencentes aos modelos [14].

3 Proposta e critérios

O *Red Teaming* é uma abordagem que pode empregar métodos manuais ou automatizados para testar adversarialmente um LLM, identificando saídas prejudiciais e aprimorando a segurança do sistema. A pesquisa proposta tem como objetivo a elaboração de um método que crie ataques LLM contra outros LLM através de ferramentas *Red Team* e AML para avaliação da segurança destes LLM. Os atacantes podem manipular os dados de entrada para gerar informações incorretas ou indesejáveis a partir de LLMs. Os modelos passam por um treinamento aprofundado em conjuntos de dados para compreender e produzir textos que imitam de forma próxima a linguagem humana. Tipicamente, os LLMs são dotados de centenas de bilhões, ou até mais, parâmetros (Tabela 1), refinados por meio do processamento de vastas quantidades de dados textuais [9].

Tabela 1. Comparativo entre LLMs populares.

Modelo	Lançamento	Provedor	Código-Aberto	Parâmetros	*Fine tunning*
GPT-4 [15]	2023-03	OpenAI	Não	1.7 T	Não
GPT-3 [16]	2020-06	OpenAI	Não	175 B	Não
Cohere-large [17]	2022-07	Cohere	Não	13 B	Sim
BERT [18]	2018-08	Google	Sim	340 M	Sim
LlaMA [19]	2023-02	Meta AI	Sim	65 B	Sim
CTRL [20]	2019	Salesforce	Sim	1.6 B	Sim
Dolly 2.0 [21]	2023-04	Databricks	Sim	12 B	Sim

O estudo iniciará com vulnerabilidades de envenenamento e ataques de evasão, podendo migrar para outras vulnerabilidades e *frameworks* de otimização de segurança. A execução desta proposta ainda está na fase de definição de infraestrutura e *frameworks* que serão adotados, no entanto, esta será focada em três modelos comerciais consolidados que são GPT-4, Google Gemini e LlaMA. Assim como no MART [12], a proposta deste trabalho será a aplicação de um método de treinamento e *fine tuning* com automação de iterações entre dois LLMs alternando entre os três adotados. Os objetivos de otimização são cruciais atuando no direcionamento de como os LLMs aprendem com os dados de treinamento e influenciando quais comportamentos são encorajados ou penalizados. As estratégias de ataques por inferência de LLM, instruções de pré-processamento, detecção maliciosa e processamento generativo (pós-processamento) serão outros aspectos que poderão ser utilizados na análise de segurança dos LLMs apresentada nesta proposta.

4 Considerações e trabalhos futuros

Apesar dos avanços, os sistemas de AML ainda apresentam desafios significativos, como a necessidade de maior colaboração entre órgãos internacionais e o desenvolvimento de soluções mais robustas contra ameaças emergentes. A prática de *Red Ream*, somada ao AML, apresentam uma estratégia para avaliação e análise de segurança de LLMs que pode ser eficaz na redução das vulnerabilidades. A pesquisa aponta a necessidade contínua de aprimorar esses métodos, à medida que os modelos evoluem, e reforça a importância de integrar práticas adversariais para mitigar riscos em sistemas de IA cada vez mais presentes em setores críticos da sociedade. Esta proposta não tem como objetivo apresentar testes para todas as ameaças existentes, muito embora, busca apresentar alternativas contra ameaças de segurança de LLMs e orientações sobre estes aspectos com perspectiva atual do uso de ferramentas *Red Team* AML.

Agradecimentos:
Os autores agradecem o apoio do LARC/USP, LabP2D/UDESC, FDTE e FAPESC. Este trabalho foi apoiado pelo CNPq (processos 307732/2023-1 e 311245/2021-8), FAPESP (processo 2020/09850-0) e CAPES (Código de Financiamento 001).

Declarações complementares

Referências

1 Brookson, C. *et al.* Definition of cybersecurity-gaps and overlaps in standardisation. *Heraklion, ENISA*, 2015.

2 Vassilev, A. *et al. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. Apostol Vassilev, Alina Oprea, Alie Fordyce, Hyrum Andersen, 2024.

3 Stanton, B.; Jensen, T. *Trust and Artificial Intelligence*. en. NIST Interagency/Internal Report (NISTIR), National Institute of Standards e Technology, Gaithersburg, MD, mar. 2021. Disponível em: https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=931087.

4 European Union Agency for Cybersecurity (ENISA). *Securing Machine Learning Algorithms*. 2021. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms. ISBN: 978-92-9204-543-2, DOI: 10.2824/874249, Catalogue Nr.: TP-06-21-153-EN-N.

5 MITRE. *MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems)*. Visitado em Dezembro de 2023. 2023. Disponível em: https://atlas.mitre.org.

6 Xu, J. *et al.* Bot-Adversarial Dialogue for Safe Conversational Agents. *In*: Toutanova, K. *et al.* (Ed.). *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, jun. 2021. P. 2950–2968. DOI: 10.18653/v1/2021.naacl-main.235.

7 Li, Y.; Liu, Q. A comprehensive review study of cyber-attacks and cyber security; Emerging trends and recent developments. *Energy Reports*, v. 7, p. 8176–8186, 2021. ISSN 2352-4847. DOI: https://doi.org/10.1016/j.egyr.2021.08.126.

8 Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems*, v. 30, 2017.

9 Yao, Y. *et al.* A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. *High-Confidence Computing*, v. 4, n. 2, p. 100211, 2024. ISSN 2667-2952. DOI: https://doi.org/10.1016/j.hcc.2024.100211.

10 Biggio, B.; Nelson, B.; Laskov, P. *Poisoning Attacks against Support Vector Machines*. 2013. arXiv: 1206.6389 [cs.LG]. Disponível em: https://arxiv.org/abs/1206.6389.

11 Kurita, K.; Michel, P.; Neubig, G. *Weight Poisoning Attacks on Pre-trained Models*. 2020. Visitado em Outubro de 2024. Disponível em: https://arxiv.org/abs/2004.06660.

12 Ge, S. *et al. MART: Improving LLM Safety with Multi-round Automatic Red-Teaming*. 2023. arXiv: 2311.07689 [cs.CL]. Disponível em: https://arxiv.org/abs/2311.07689.

13 Wan, A. *et al. Poisoning Language Models During Instruction Tuning*. 2023. Visitado em Outubro de 2024. Disponível em: https://arxiv.org/abs/2305.00944.

14 Zhang, Y. *et al.* Siren's song in the AI ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

15 Ding, X. *et al.* HPC-GPT: Integrating Large Language Model for High-Performance Computing. *In*: PROCEEDINGS of the SC '23 Workshops of The International Conference on High Performance Computing, Network, Storage, and Analysis. Denver, CO, USA: Association for Computing Machinery, 2023. (SC-W '23), p. 951–960. ISBN 9798400707858. DOI: 10.1145/3624062.3624172.

16 Brown, T. B. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

17 Liang, P. *et al. Holistic Evaluation of Language Models*. 2023. arXiv: 2211.09110 [cs.CL]. Disponível em: https://arxiv.org/abs/2211.09110.

18 Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

19 Touvron, H. *et al. LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL]. Disponível em: https://arxiv.org/abs/2302.13971.

20 Socher, R. *Introducing a Conditional Transformer Language Model for Controllable Generation*. 2019. Disponível em: https://blog.salesforceairesearch.com/introducing-a-conditional-transformer-language-model-for-controllable-generation/. Acesso em: 11 set. 2019.

21 Conover, M. *et al. Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. 2023. Disponível em: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm. Acesso em: 12 abr. 2023.