CIBERSEGURANÇA PARA SISTEMAS INTELIGENTES: ANÁLISE DE FERRAMENTAS E MODELOS NA PERSPECTIVA DE RED TEAMING

Paulo Henrique Gomes de Senna, Charles Christian Miers.

INTRODUÇÃO

A crescente utilização de sistemas baseados em *Large Language Models* (LLMs) em aplicações reais tem ampliado suas funcionalidades e formas de interação com usuários, bases de dados e ferramentas externas. Esse avanço também introduz novas superfícies de ataque e comportamentos potencialmente inseguros, tornando necessária a avaliação sistemática de suas vulnerabilidades (SHAYEGANI et al., 2023; VASSILEV et al., 2025). Nesse contexto, este trabalho tem como objetivo avaliar experimentalmente a segurança de modelos generativos por meio de ferramentas e técnicas adversariais.

DESENVOLVIMENTO

A pesquisa foi conduzida por meio da exploração, implementação e integração de ferramentas voltadas à avaliação de segurança de modelos de IA generativa. Nesse contexto, foi explorado e aprimorado um *toolkit* capaz de integrar diferentes técnicas de ataque a conjuntos de dados e modelos, permitindo a execução sistemática de experimentos e a comparação dos resultados obtidos (MAZEIKA et al., 2024; CHAO et al., 2024). Paralelamente, foram realizados experimentos envolvendo mecanismos de persistência de memória e funcionalidades externas de modelos generativos, investigando a possibilidade de induzir comportamentos que não seriam obtidos por meio de interações convencionais (DONG et al., 2025). Também foram avaliados os impactos de ataques de envenenamento sobre o desempenho de sistemas baseados em RAG, considerando métricas de recuperação e geração de respostas (ZOU et al., 2025), esse estudo resultou em um artigo científico publicado no ERAD/RS 2026 em Bagé/RS (SENNA; MIERS, 2026). Por fim, foi realizada a unificação de diferentes conjuntos de dados utilizados na avaliação de segurança de LLMs, acompanhada da elaboração de uma taxonomia comum para classificação dos ataques e da inclusão da categoria de *Compliance*, voltada à avaliação de comportamentos incompatíveis com a finalidade ou as atribuições esperadas de determinado sistema (CHAO et al., 2024; MAZEIKA et al., VASSILEV et al., 2025). Esse último estudo também resultou em um artigo científico publicado no congresso da SBSEG 2026 (SÁ et al., 2026).

RESULTADOS

Os experimentos realizados permitiram identificar diferentes comportamentos vulneráveis em modelos generativos e avaliar as limitações dos mecanismos de proteção diante de técnicas adversariais. Os testes envolvendo a persistência de memória demonstraram a possibilidade de induzir comportamentos que dependem de mecanismos persistentes e de funcionalidades externas, como ferramentas de MCP e *skills*. Nos experimentos com sistemas RAG, o envenenamento da base vetorial provocou aumento da latência de recuperação e da latência até o primeiro *token*, além de uma redução no *throughput*, evidenciando que ataques adversariais podem afetar não apenas o comportamento do modelo, mas também o desempenho do sistema. A organização dos conjuntos de dados em uma taxonomia unificada também ampliou a abrangência das avaliações, incorporando aspectos de *Compliance* que não são necessariamente contemplados por classificações tradicionais

de conteúdo nocivo, possibilitando a realização de avaliações experimentais diretamente sobre modelos generativos.

CONSIDERAÇÕES

Os resultados obtidos reforçam a importância da avaliação contínua de segurança em sistemas baseados em LLMs, especialmente diante da incorporação de mecanismos de memória, ferramentas externas e arquiteturas RAG. Os experimentos evidenciaram que essas funcionalidades podem introduzir novas superfícies de ataque e impor limitações aos mecanismos tradicionais de proteção, além de ocasionar impactos no desempenho dos sistemas. Dessa forma, a pesquisa contribui para o desenvolvimento de abordagens experimentais voltadas à avaliação de segurança de modelos generativos sob a perspectiva de *Red Teaming*.

Palavras-chave: *large language models;* inteligência artificial; cibersegurança; *red teaming*; ataques adversariais.

ANEXOS

Tabela 1. *Comparação do desempenho do sistema RAG antes e após o envenenamento persistente da base vetorial.*

Métrica	Baseline	Envenenamento Persistente
Latência E/S (FAISS)	8,83 ms	96,36 ms
TTFT (Atenção LLM)	72,03 ms	146,53 ms
Latência Total	3.632,72 ms	3.884,23 ms
Throughput (tokens/s)	42,51	41,47

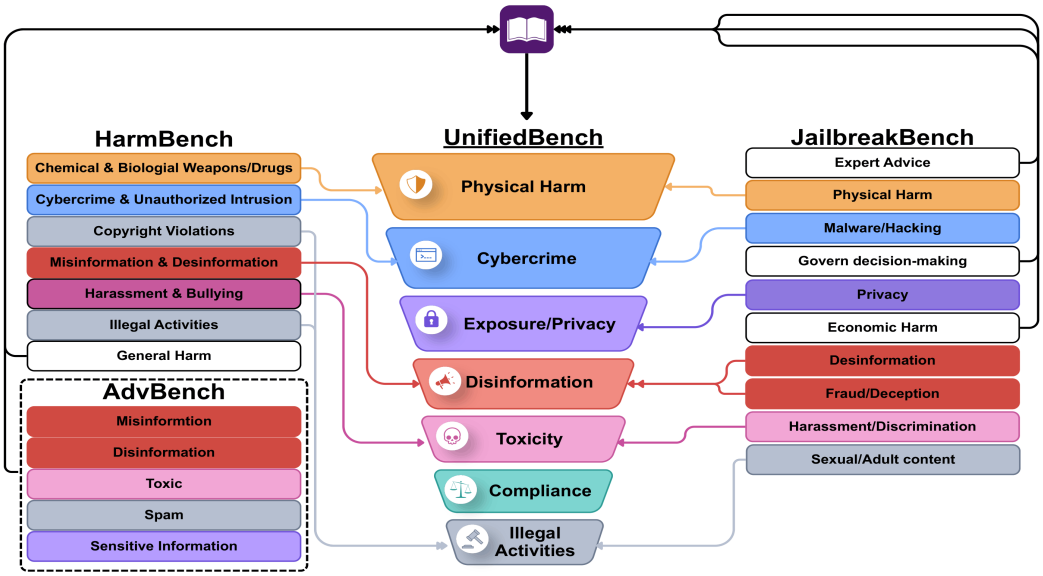


Figura 1. *Taxonomia de ataques desenvolvida a partir da unificação de três conjuntos de dados utilizados na avaliação adversarial de modelos de linguagem.*

REFERÊNCIAS BIBLIOGRÁFICAS

CHAO, P.; DEBENEDETTI, E.; ROBEY, A.; ANDRIUSHCHENKO, M.; CROCE, F.; SEHWAG, V.; DOBRIBAN, E.; FLAMMARION, N.; PAPPAS, G. J.; TRAMER, F. et al. **JailbreakBench: An open robustness benchmark for jailbreaking large language models**. *Advances in Neural Information Processing Systems*, v. 37, p. 55005–55029, 2024.

DONG, S.; XU, S.; HE, P.; LI, Y.; TANG, J.; LIU, T.; LIU, H.; XIANG, Z. MINJA: **A practical memory injection attack against LLM agents**. 2025.

MAZEIKA, M.; PHAN, L.; YIN, X.; ZOU, A.; WANG, Z.; MU, N.; SAKHAEE, E.; LI, N.; BASART, S.; LI, B.; FORSYTH, D.; HENDRYCKS, D. **HarmBench: A standardized evaluation framework for automated red teaming and robust refusal**. In: *International Conference on Machine Learning (ICML)*, 2024.

SHAYEGANI, E.; MAMUN, M. A. A.; FU, Y.; ZAREE, P.; DONG, Y.; ABU-GHAZALEH, N. **Survey of vulnerabilities in large language models revealed by adversarial attacks**. *arXiv preprint arXiv:2310.10844*, 2023.

VASSILEV, A.; OPREA, A.; FORDYCE, A.; ANDERSON, H.; DAVIES, X.; HAMIN, M. Adversarial machine learning: **A taxonomy and terminology of attacks and mitigations**. *Technical Report NIST AI 100-2e2025*, National Institute of Standards and Technology, 2025.

ZOU, W.; GENG, R.; WANG, B.; JIA, J. **PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models**. In: *34th USENIX Security Symposium*, 2025.

SÁ, M. A. de; SENNA, P. H. G. de; PAGLIUSO NETO, M. P.; MIERS, C. C. **UnifiedBench: Toward a Unified Dataset for Red-Teaming Language Models**. In: Workshop de Trabalhos de Iniciação Científica e de Graduação (WTICG) - Simpósio Brasileiro de Cibersegurança (SBSeg), 2026. DOI: https://doi.org/10.5753/sbseg.2026

SENNA, P. H. G. de; MIERS, C. C. **Análise comparativa do impacto de ataques à memória persistente no desempenho de sistemas baseados em LLMs**. In: Escola Regional de Alto Desempenho da Região Sul (ERAD-RS), 2026. DOI: https://doi.org/10.5753/eradrs.2026