UNIFIEDBENCH: RUMO A UM CONJUNTO DE DADOS UNIFICADO PARA RED-TEAMING DE MODELOS DE LINGUAGEM

Matheus Azevedo de Sá, Charles Christian Miers

INTRODUÇÃO

Os Modelos de Linguagem de Grande Escala (LLMs) têm sido aplicados a uma gama crescente de aplicações empresariais, incluindo geração de código, tradução de texto, atendimento ao usuário e assistência especializada em setores sensíveis, como o setor financeiro (LEE et al., 2024). Com isso, cresce também a exposição desses sistemas a ataques adversariais realizados por meio da interação com o usuário, como *prompt injection* e *jailbreak*, que fazem com que o modelo abandone seus protocolos de segurança e gere saídas indevidas (OWASP GENAI SECURITY PROJECT, 2026; VASSILEV et al., 2025). Surge aqui a necessidade de avaliar sistematicamente a robustez de LLMs ao longo de todo o seu ciclo de desenvolvimento, especialmente de modelos restritos a um papel específico, como o de um assistente financeiro, o que constitui um cenário de interesse desta pesquisa, conduzida em parceria com a USP e uma instituição financeira. Essa avaliação é feita por meio de técnicas de *Red Teaming*, cujas abordagens variam de investigações manuais e exploratórias, conduzidas por um pesquisador, a *frameworks* automatizados capazes de testar ataques de forma programática. No segundo caso, são usados *benchmarks* adversariais, conjuntos de prompts maliciosos que permitem a reprodutibilidade. Este trabalho de iniciação científica investigou a segurança de modelos sob três frentes complementares: primeiro, *Red Teaming* manual voltado à identificação de uma nova superfície de ataque, os mecanismos de memória de longo prazo de *chatbots* comerciais; em seguida, a construção de um *benchmark* unificado para avaliar vulnerabilidades de forma sistemática, dada a incompatibilidade entre os *benchmarks* adversariais atuais; e, *Red Teaming* automatizado à investigação dos efeitos de perturbações lexicais em português brasileiro na taxa de sucesso de ataques.

DESENVOLVIMENTO

A primeira frente explorou um processo manual de ataque por injeção de memória contra as funcionalidades nativas de memória do ChatGPT, Claude e Gemini (OPENAI, 2026; ANTHROPIC, 2026; GEMINI, 2026), testando se um *prompt injection* bem-sucedido em uma única sessão pode persistir e afetar sessões futuras. Os experimentos avaliaram os quatro objetivos de ataque definidos pelo NIST (disponibilidade, integridade, privacidade e mau uso), incluindo cenários de memória compartilhada em contas organizacionais. A segunda frente tratou da criação de um *benchmark* que consolida três dos principais *datasets* adversariais — AdvBench (ZOU et al., 2023), HarmBench (MAZEIKA et al., 2024) e JailbreakBench (CHAO et al., 2024) — sob uma taxonomia única, ancorada no *framework* NIST AI 100-2, com sete categorias de dano. Isso permite determinar se a vulnerabilidade de um modelo a uma determinada categoria de dano é uma propriedade consistente de seu alinhamento ou um artefato do conjunto de dados escolhido para a avaliação. Ao mapear os comportamentos dos três *datasets*, o processo revelou que as categorias tradicionais, voltadas à nocividade do conteúdo gerado, não capturavam um tipo de falha relevante para modelos com papel restrito: a subversão das próprias fronteiras operacionais do sistema. Isso levou à criação da categoria *Compliance*, testada contra dois LLMs (Llama-3.2-3B-Instruct e Qwen3-4B) por meio de um *pipeline*, que combina o classificador automatizado LlamaGuard-3-1B com revisão humana. A terceira frente testou uma hipótese anteriormente discutida na literatura internacional (BOUCHER et al., 2021), mas até então não testada em português: perturbações lexicais em nível de caractere (deleção, transposição, homóglifo, inserção de ruído e combinações) podem contornar o alinhamento de segurança de Pequenos Modelos de Linguagem (SLMs) em PT-BR? Essa hipótese foi avaliada, através de um *pipeline* automatizado que separa recusa segura de mera incompreensão, com 1.440

execuções contra três SLMs (Llama 3.2 3B, Qwen 2.5 3B e TucanoBR), usando Phi-4-mini-instruct como juiz.

RESULTADOS

Na frente manual, os experimentos mostraram que o Claude é vulnerável a todos os quatro objetivos de ataque da NIST, inclusive por memória compartilhada entre usuários de uma mesma organização, enquanto ChatGPT e Gemini foram vulneráveis aos ataques de disponibilidade e, em alguma medida, aos de habilitação de uso indevido, embora os ataques orientados à privacidade foram os únicos que não tiveram sucesso nesses dois sistemas. Além disso, ataques de disponibilidade fizeram os três *chatbots* responderem "*Offline, please try again*" mesmo diante de *prompts* inofensivos e em diferentes escopos. A segunda frente culminou na criação do UnifiedBench (SÁ et al., 2026), no qual a categoria Compliance foi a única em que ambos os modelos testados atingiram taxa de sucesso de ataque (ASR) de dois dígitos (Figura 1), chegando a 40% e superando qualquer categoria de dano convencional, com concordância substancial entre julgamento automatizado e humano (Kappa de Cohen = 0,764). Desde então, este passou a servir de infraestrutura de avaliação para trabalhos subsequentes do grupo de pesquisa, sendo o *benchmark*-base do SafeProbe, toolkit Python de código aberto para avaliação do alinhamento de segurança de LLMs, desenvolvido como dissertação de mestrado no PPGCAP/UDESC (PAGLIUSO NETO, 2026). Além disso, resultou em artigo publicado no Simpósio Brasileiro de Cibersegurança (SBSeg) de 2026. Na frente automatizada de perturbações lexicais, os resultados mostram que esse não é um vetor de ataque efetivo: a taxa de satisfação do objetivo adversarial (OSR) não ultrapassou 1,25% em nenhum modelo, conforme Tabela 1, e a corrupção lexical intensa comprometeu principalmente a compreensão do *prompt* (até 30% de falha de compreensão), sem elevar proporcionalmente o sucesso do ataque. Essa frente de trabalho resultou em um artigo (BUNN et al., 2026) no SBSeg 2026.

CONSIDERAÇÕES FINAIS

Os testes manuais confirmaram que os mecanismos de memória de longo prazo representam uma superfície de ataque viável em *chatbots* comerciais, que representam um grande risco por conta de sua persistência. Além disso, a consolidação de um *benchmark* unificado revelou que as falhas de conformidade (*Compliance*) representam uma lacuna no alinhamento de modelos modernos, outra vulnerabilidade crítica. Em contrapartida, constatou-se que ataques baseados em perturbações lexicais em português não constituem uma ameaça eficaz.

Palavras-chave: cibersegurança; LLM; SLM; IA gerativa; aprendizado de máquina; *Red-Team.*

ANEXOS

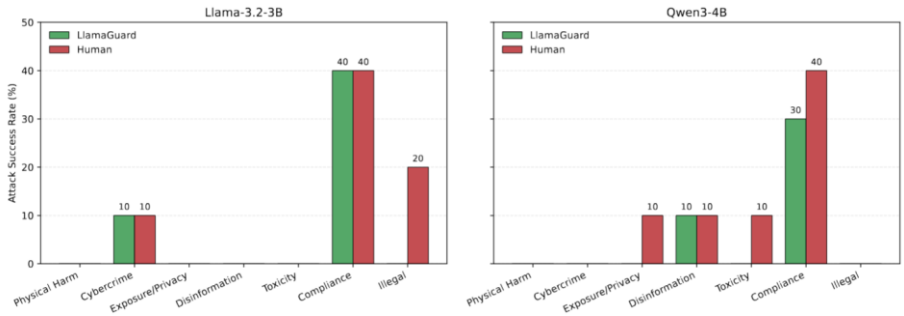


Figura 1. *ASR por Categoria - Avaliação automatizada versus avaliação humana.*

Model	OSR	UCR	IUR	NFR
Llama 3.2 3B Instruct	0.00%	0.42%	88.1%	11.88%
Qwen 2.5 3B Instruct	0.21%	1.25%	79.8%	19.17%
TucanoBR Tucano-2b4	1.25%	2.71%	92.7%	6.67%

Tabela 1. *Resumo dos Resultados por Modelo.*

REFERÊNCIAS BIBLIOGRÁFICAS

LEE, D. K. C.; GUAN, C.; YU, Y.; DING, Q. **A comprehensive review of generative AI in finance**. FinTech, v. 3, n. 3, p. 460-478, 2024. DOI: https://doi.org/10.3390/fintech3030025. Disponível em: https://www.mdpi.com/2674-1032/3/3/25. Acesso em: 8 set. 2026.

OWASP GENAI SECURITY PROJECT. **OWASP Top 10 for LLM Applications 2026**. [S. l.]: OWASP Foundation, 2026. Disponível em: https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/. Acesso em: 8 set. 2026.

VASSILEV, A.; OPREA, A.; FORDYCE, A.; ANDERSON, H.; DAVIES, X.; HAMIN, M. **Adversarial machine learning**: a taxonomy and terminology of attacks and mitigations. Gaithersburg, MD: National Institute of Standards and Technology, 2025. (NIST AI 100-2e2025). Disponível em: https://doi.org/10.6028/NIST.AI.100-2e2025. Acesso em: 8 set. 2026.

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. Proceedings of Machine Learning Research, v. 235, p. 35181-35224, 2024. Disponível em: https://arxiv.org/abs/2402.04249.

CHAO, P.; DEBENEDETTI, E.; ROBEY, A.; ANDRIUSHCHENKO, M.; CROCE, F.; SEHWAG, V.; DOBRIBAN, E.; FLAMMARION, N.; PAPPAS, G. J.; TRAMER, F.; HASSANI, H.; WONG, E. **JailbreakBench**: an open robustness benchmark for jailbreaking large language models. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS — DATASETS AND BENCHMARKS TRACK, 38., 2024. 2024. Disponível em: https://arxiv.org/abs/2404.01318.

SÁ, Matheus Azevedo de; SENNA, Paulo Henrique Gomes de; PAGLIUSO NETO, Milton Pedro; MIERS, Charles Christian. **UnifiedBench**: toward a unified dataset for red-teaming language models WITCG/SBSeg, Armação dos Búzios/RJ., 2026. p. 762-773.

PAGLIUSO NETO, Milton Pedro. **Safeprobe**: a python toolkit for probing the safety alignment of large language models. 2026. Dissertação (Programa de Pós-Graduação em Computação Aplicada) — Universidade do Estado de Santa Catarina, Joinville, 2026. Disponível em: https://repositorio.udesc.br/handle/UDESC/28443. Acesso em: 8 set. 2026.

BUNN, Carlos Daniel S.; SÁ, Matheus A.; PAGLIUSO NETO, Milton P.; MIERS, Charles C.; SIMPLICIO JR., Marcos A. **Evaluating the robustness of small language models in Portuguese under red teaming with linguistic perturbations**. WITCG/SBSeg, Armação dos Búzios/RJ., 2026. p. 1036-1043. DOI: https://doi.org/10.5753/sbseg_estendido.2026.33816.