ANÁLISE DE FERRAMENTAS DE ATAQUE A LLM E IMPLEMENTAÇÃO EM TOOLKIT – PROMP INJECTION

Giovanni Mandel Martignago, Charles Christian Miers

INTRODUÇÃO

A crescente adoção de *Large Language Models* (LLMs) tem evidenciado a necessidade de avaliar sistematicamente sua robustez frente a ataques adversariais. Técnicas de *Red Teaming*, como os ataques de *jailbreak*, que buscam contornar os mecanismos de alinhamento e de segurança dos modelos, são centrais nesse processo, mas há poucas comparações sistemáticas entre ferramentas quanto à eficácia e ao custo operacional. Este trabalho teve como objetivo comparar ferramentas de *Red Teaming* quanto à taxa de sucesso de ataque (ASR) e ao custo computacional/monetário, avaliando quatro ferramentas (AttnGCG, NanoGCG, GPTFuzzer e PAPILLON) contra um modelo de uso geral (Llama 3.1 8B Instruct) e, complementarmente, aplicando duas destas (PAIR e nanoGCG) a um *chatbot* industrial.

DESENVOLVIMENTO

O trabalho de pesquisa teve como foco a comparação de ferramentas de *Red Teaming* em LLMs. Especificamente, foram realizadas duas rodadas de comparação de ferramentas e técnicas de *jailbreak*: uma com quatro ferramentas — AttnGCG, nanoGCG, GPTFuzzer e Papillon — e outra com duas — nanoGCG e PAIR. Na primeira rodada, a comparação consistiu em execuções em um ambiente controlado (servidor disponibilizado pelo LabP2D - UDESC), com as ferramentas atacando o mesmo modelo-alvo, um LLM juiz externo avaliando as respostas, orçamento de 500 iterações em cada uma delas e o mesmo *dataset*, composto por 18 objetivos do HarmBench Standard (três objetivos de cada uma das seis categorias). No segundo, o PAIR e o nanoGCG foram utilizados para avaliar um *chatbot* de manutenção industrial baseado em LLM; os casos de ataque foram definidos com a empresa parceira. Essas ferramentas são utilizadas no contexto de avaliação da robustez das técnicas de segurança dos modelos de IA generativa de texto. Foram elaborados dois artigos submetidos e aceitos na SBSeg 2026. Um trabalho comparou ferramentas de *Red Teaming* (*jailbreak*) quanto à eficácia em contornar o alinhamento de LLMs, em duas partes. Na primeira, AttnGCG, GPTFuzzer e PAPILLON foram comparadas contra o Llama 3.1 8B Instruct, em ambiente controlado no servidor do LabP2D-UDESC, com orçamento de 500 iterações por ferramenta, *dataset* de 18 objetivos do *HarmBench Standard* e avaliação por um LLM juiz externo. Essa etapa resultou na aceitação de um artigo para a trilha do *Workshop* de Trabalhos de Iniciação Científica e de Graduação (WTICG) do SBSeg 2026. O nanoGCG foi avaliado posteriormente, pelo mesmo método, e incorporado apenas ao Trabalho de Conclusão de Curso (TCC). Na segunda etapa, PAIR e nanoGCG avaliaram um *chatbot* industrial baseado em LLM, desenvolvido em parceria com uma empresa do setor metal-mecânico. O autor conduziu os testes com o NanoGCG; o PAIR foi avaliado por um coautor. Essa etapa resultou em um segundo artigo, aceito na Trilha de Indústria do SBSeg 2026.

RESULTADOS

Na comparação com quatro ferramentas, o GPTFuzzer obteve a maior ASR (83,33%; 15/18), seguido por nanoGCG (33,3%; 6/18), Papillon (22,22%; 4/18) e AttnGCG (5,56%; 1/18). Os sucessos concentraram-se em cinco das seis categorias do *HarmBench*, com o nanoGCG cobrindo o maior número de categorias e o GPTFuzzer alcançando 100% de sucesso em atividades ilegais e em conteúdo prejudicial. Em custo operacional, AttnGCG e nanoGCG (baseados em gradiente) exigiram mais tempo e memória de GPU — 6h25min56s e 3h39min46s, com pico de GPU elevado e de 17,86 GB, respectivamente — sem custo monetário direto. GPTFuzzer e Papillon, dependentes de API (GPT-3.5-Turbo), custaram USD 0,47 (9.000 consultas, 3h08min) e USD 2,70 (6.600 consultas,

5h47min), este último reduzido por parada antecipada. Na avaliação do NanoGCG no *chatbot* industrial (critério: pontuação ≥ 7), o sufixo teve efeito oposto entre os modelos: ganho de 6,66 p.p. no Llama-3.1-8B e queda de 26,67 p.p. no Mistral-7B. A ASR *baseline* já era alta em ambos (46,67%). Apenas a categoria de extração de dados sensíveis teve ganho consistente (40%→60%); nas demais, o sufixo manteve ou reduziu a ASR, chegando a zerar os sucessos em sabotagem e privacidade no Mistral-7B.

CONSIDERAÇÕES FINAIS

Os resultados mostram que as ferramentas avaliadas têm eficácia e custo bastante distintos, sem um método universalmente superior: o GPTFuzzer aliou alta eficácia a baixo custo, enquanto o AttnGCG apresentou baixa eficácia e o maior custo computacional. No cenário industrial, o sufixo do nanoGCG apresentou efeitos inconsistentes entre modelos, incluindo degradação de desempenho, o que reforça a ideia de que resultados de *benchmarks* genéricos não devem ser extrapolados para sistemas de produção. A ASR *baseline* elevada em ambos os cenários sugere ainda que parte da vulnerabilidade decorre do próprio alinhamento do modelo, e não apenas da técnica de ataque. Os objetivos foram alcançados, e os resultados foram consolidados em dois artigos aceitos no SBSeg 2026.

Palavras-chave: Red Team; Large Language Model; Adversarial Machine Learning; chatbot industrial.

ANEXOS

Tabela 1. *Custo computacional e eficiência dos ataques contra o Llama 3.1 8B Instruct.*

Método	Tempo	Consultas ao Alvo	Total de Tokens	Custo (USD)	Pico de Mem. GPU
AttnGCG	6h 25m 56s	—	—	—	41,61 GB
NanoGCG	3h 39m 46s	—	—	—	17,86 GB
GPTFuzzer	3h 08m 47s	9.000	988.160	$0,47	—
PAPILLON	5h 47m 09s	6.600	3.788.181	$2,70	—

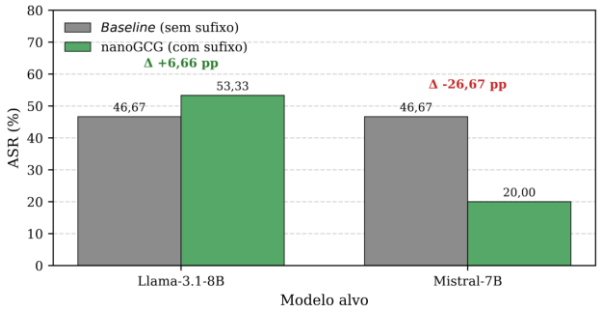


Figura 1. *Resultados ASR do nanoGCG por modelo alvo (limiar ≥ 7).*

REFERÊNCIAS BIBLIOGRÁFICAS

MARTIGNAGO, Giovanni Mandel; PAGLIUSO, Milton; MIERS, Charles Christian. Comparative analysis of automated red teaming in LLMs: AttnGCG, GPTFuzzer, and PAPILLON. WTICG, Armação dos Búzios. XXVI Simpósio Brasileiro de Cibersegurança (SBSeg 2026).

HAYASHI, Victor Takashi; PAGLIUSO NETO, Milton Pedro; MARTIGNAGO, Giovanni Mandel; MIERS, Charles Christian; SIMPLICIO JUNIOR, Marcos Antonio. Red teaming an industrial

maintenance chatbot using PAIR and nanoGCG. Trilha Indústria, Armação dos Búzios. XXVI Simpósio Brasileiro de Cibersegurança (SBSeg 2026).