Plagiarism Scan Report

2% Plagiarism	2% Exact Match
	0% Partial Match

98% Unique

Words	896
Characters	6188
Sentences	36
Paragraphs	117
Read Time	5 minute(s)
Speak Time	6 minute(s)

Content Checked For Plagiarism

Estado da Arte II e Metodologia: Segurança, Privacidade e Conformidade em Aplicações com LLMs

Leonardo Nunes1, Antonio Marcos1, Álvaro Gueiros1, Lucas William1, Mauro Vinícius1, Vandielson Tenório1

1Aluno da disciplina de Segurança da Informação do Bacharelado em Ciência da Computação – Universidade Federal do Agreste de Pernambuco (UFAPE)

Abstract. This paper reports a focused literature-based gap analysis and two controlled experiment rounds for an LLM security proof of concept. We argue that the field lacks a small and reproducible protocol that jointly reports security effectiveness and operational impact under common attacks. We implement a guardrail in front of a configurable LLM provider and evaluate four scenarios (benign prompt, two prompt-injection variants, and context-overflow denial of service), measuring attack recall, false positives, latency, and throughput. Experiment 1 blocks all attacks but incorrectly flags all benign prompts, revealing an overly conservative decision rule. We then apply targeted interventions (strong vs weak signals, calibration on a development set, frozen parameters, and provider decoupling via mock and Ollama) and repeat the evaluation. Experiment 2 preserves attack blocking while allowing benign traffic and enables stable experimentation without mandatory dependence on a cloud provider.

Resumo. Este artigo apresenta uma análise de lacuna baseada em literatura e duas rodadas de experimentos controlados para um proof-of-concept em segurança de LLMs. Argumentamos que falta um protocolo pequeno e reprodutível que reporte, ao mesmo tempo, eficácia de segurança e impacto operacional sob ataques comuns. Implementamos um guardrail antes de um provedor de LLM configurável e avaliamos quatro cenários (Prompt Seguro, duas variações de Prompt Injection e negação de serviço por excesso de contexto), medindo recall em ataques, falsos positivos, latência e vazão. No Experimento 1, o sistema bloqueia ataques, mas rejeita todo o benigno, evidenciando uma regra de decisão excessivamente conservadora. Aplicamos intervenções (sinais fortes vs fracos, calibração em dev set, congelamento de parâmetros, desacoplamento de provedor com mock e Ollama) e repetimos o protocolo. No Experimento 2, o sistema mantém bloqueio de ataques e passa a permitir tráfego benigno, viabilizando avaliação estável sem dependência obrigatória de nuvem.

1. Introdução

LLMs ampliaram automação e suporte à decisão, mas introduziram superfícies de ataque como prompt injection (direta e indireta), insecure output handling, exaustão por contexto (model denial of service, inclusive negação por custo), e riscos de vazamento e confor-

midade. A literatura recente cobre taxonomias e controles, controle de acesso adaptativo em domínios críticos, riscos em RAG e estudos empíricos ligados a segurança em desenvolvimento, porém frequentemente com propostas de alto nível, pouca reprodutibilidade e pouca ênfase em métricas operacionais, principalmente falsos positivos [1, 2, 3, 4, 5, 6]. Este trabalho adota um enfoque pragmático: um protocolo mínimo e replicável centrado em uma camada de guardrail antes do provedor. O objetivo é medir o equilíbrio entre segurança e disponibilidade com métricas de erro e desempenho, comparando explicitamente uma rodada que falha operacionalmente e uma rodada que atende critérios definidos.

Palavras-chave. Segurança de LLMs, prompt injection, guardrails, OWASP LLM Top 10, falsos positivos, reprodutibilidade, Ollama, avaliação experimental.

Contribuições. (1) Formalização de uma lacuna prática, ausência de protocolo mínimo comparável com métricas de erro e desempenho; (2) protótipo reprodutível com provedor configurável (mock, Ollama, nuvem opcional); (3) protocolo com calibração e parâmetros congelados; (4) evidência experimental em dois ciclos, um com falha por falsos positivos e outro com sucesso.

2. Lacuna e trabalhos relacionados

Sínteses recentes organizam ameaças e controles em LLMs, destacando sanitização, filtros, RAG e HITL [1]. Em contexto setorial, [2] discutem controle de acesso adaptativo e avaliação quantitativa. No eixo regulatório, [3] traduz o AI Act em responsabilidades e ações. Em RAG, [4] sistematiza riscos e mitigações, reforçando a necessidade de avaliar controles com métricas operacionais, não apenas com descrições conceituais. Em engenharia de software, [5] avalia técnicas de prompting para geração de código com foco em segurança, mostrando que intervenções no prompt podem reduzir fraquezas, mas exigem avaliação rigorosa e comparável. Por fim, [6] realiza estudo empírico comparando ChatGPT com ferramentas de análise estática para detecção de mau uso de criptografia, reforçando a relevância de protocolos controlados e métricas (inclusive repetição e agregação) para reduzir variabilidade. A lacuna que exploramos é operacional: falta um framework mínimo, reprodutível e comparável que integre defesa contra ataques comuns (injeção de prompt e DoS por contexto) e reporte explícito de métricas de viabilidade, especialmente falsos positivos, que determinam se usuários legítimos conseguem usar o sistema.

Tabela 1. Comparação dos trabalhos e evidência da lacuna prática (inclui os dois anexos)

Trabalho	Foco	Tipo de evidência	Limitação prática para comparação e replicação
Rathod et al. [1]	Taxonomia de ameaças e controles em LLMs	Síntese e recomendações	Métricas operacionais e protocolo experimental mínimo nem sempre padronizados
Yarram et al. [2]	Acesso adaptativo e anomalias (saúde)	Avaliação quantitativa em domínio específico	Generalização limitada, não foca ataques típicos de LLM apps como prompt injection e DoS por contexto

Bunzel [3]	Compliance no AI Act, papéis e controles	Guia prático regulatório Diretrizes sem benchmark técnico mínimo e comparável
Ammann et al. [4]	Riscos e mitigações em RAG	Framework e mitigação orientada a riscos Necessita ligação direta com protocolo mínimo e métricas operacionais sob ataques
Tony et al. [5] (anexo)	Técnicas de prompting para geração de código seguro	SLR + avaliação em múltiplos LLMs e dataset Mostra impacto de intervenções, mas reforça necessidade de protocolos

Matched Source

Similarity 2%

Title:Bacharelado em Ciência da Computação - BCC UFAPE UFRPE-UAG

Descrição Geral do curso de BCC da UFAPE (UFRPE-UAG).--Bacharelado em Ciência da Computação Universidade Federal do Agreste de Pernambuco (UFRPE-UAG)@bccuagh...

https://www.youtube.com/watch?v=ZYj5ZYvmSQo