Universidade de Pernambuco Programa de Pós-Graduação em Engenharia da Computação (PPGEC)

Proposta de Dissertação de Mestrado

Área: Computação Inteligente

Título: Intercept Prompt Injection Attack- IPIA :Segurança e integridade em Grandes Modelos de Linguagem

Orientador – Fernando Buarque de Lima Neto (fbln@ecomp.poli.br)
Co-orientador – André Caetano (andre.firmo@tjpe.jus.br)

Contexto

A transformação digital do Poder Judiciário brasileiro, impulsionada pelo programa Justiça 4.0, encontrou nos Grandes Modelos de Linguagem (LLMs) um vetor sem precedentes para a otimização da produção textual e suporte à decisão jurídica. No âmbito do Tribunal de Justiça de Pernambuco (TJPE), a plataforma MAIA (Mecanismo Artificial Inteligente de Apoio à Justiça) consolida-se como uma ferramenta estratégica para dar vazão ao volume processual, automatizando minutas, resumos de peças e análises documentais. Contudo, essa rápida adoção ocorre em um cenário onde a infraestrutura de segurança cibernética tradicional se mostra insuficiente. A vulnerabilidade intrínseca das arquiteturas de Redes Neurais Artificiais (RNAs) a manipulações baseadas em linguagem natural introduz riscos sistêmicos severos, uma vez que dados de terceiros — frequentemente maliciosos — passam a ser processados de forma direta por algoritmos com alto grau de autonomia.[1][2]

Problema

Embora os LLMs apresentem um potencial disruptivo para aumentar a eficiência jurídica, sua integração em sistemas de missão crítica no setor judiciário carece de mecanismos robustos de governança e defesas lógicas contra ataques de injeção de prompt (Prompt Injection Attacks). Atualmente, a recepção de petições iniciais, contestações e documentos externos na plataforma MAIA expõe o ecossistema do TJPE a vetores de ataque contemporâneos (como injeções indiretas e envenenamento de ferramentas via protocolos como MCP). A ausência de uma camada de mediação de segurança especializada (IPIA) impede a detecção em tempo real de comandos maliciosos camuflados no fluxo textual legal, criando uma lacuna que ameaça a integridade das decisões automatizadas, a confidencialidade dos dados processuais protegidos por segredo de justiça e a confiabilidade reputacional do próprio tribunal.[3]

Pergunta Principal

Como conceber, implementar e avaliar um serviço integrado de identificação, monitoramento e ação preventiva contra ataques de prompt injection na plataforma MAIA do TJPE, que neutralize de forma sustentável as ameaças emergentes à segurança e integridade dos LLMs sem degradar a qualidade e a performance da assistência textual jurídica?

Perguntas Secundárias

Como contribuir, conceber e estabelecer uma prova de conceito de segurança agêntica no ambiente judiciário, integrando princípios de:

a) Segurança de Engenharia de Prompt e Alinhamento;
b) Governança Tecnológica do Setor Público (Judiciário);
c) Ciência Computacional e IA Explicável (XAI);
d) Validação Criminológica Digital e Forense de IA;

Objetivos

Desenvolver um framework e um serviço denominado IPIA (Intercept Prompt Injection Attack), integrando algoritmos de Processamento de Linguagem Natural (PLN) e classificadores de segurança de Aprendizado de Máquina (ML) para a detecção precoce de anomalias textuais na plataforma MAIA do TJPE. A solução atuará de maneira transparente, interceptando o fluxo de dados de entrada (inputs de usuários e documentos de terceiros) e saídas (outputs dos agentes autônomos). Uma prova de conceito funcional deverá utilizar dados reais de fluxos processuais e interações da plataforma MAIA do TJPE (descaracterizados em conformidade com a LGPD) a fim de realizar um estudo comparativo quantitativo e qualitativo das taxas de falso positivo, falso negativo e latência das defesas implementadas frente a uma biblioteca simulada de ataques contemporâneos.[4][5]

Produtos esperados

a) Revisão Estruturada e Taxonomia Contemporânea.
b) Arquitetura e Modelo Lógico do IPIA.
c) Dataset de Benchmarking Dinâmico Jurídico.
d) Protocolo de Resposta a Incidentes Agênticos.

Referências

[1] Jinqi Lai et al (2023). Large Language Models in Law: A Survey
[2] Jayr Pereira et al. (2024). INACIA: Integrating Large Language Models in Brazilian Audit Courts: Opportunities and Challenges.
[3] Yi Liu et al. (2023). Prompt Injection attack against LLM-integrated Applications
[4] Andy Zhou et al (2024). Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks
[5] Tarek Ali et al. (2023) HuntGPT: Integrating Machine Learning-Based Anomaly Detection and Explainable AI with Large Language Models (LLMs)