Proposta Formal do Projeto

Universidade Federal do Agreste de Pernambuco - UFAPE
Disciplina: Segurança da informação
Professor: Sérgio Mendonça
Equipe: Álvaro Miguel, Juan Gustavo, Lucas Emanoel, Lucas Willian, Mauro Vinícius, Leonardo Nunes, Vandielson Tenório e Joás Gomes

1. Título do Projeto e Objetivo de Segurança

- **Título:** Segurança de Prompt em Modelos de LLM: Protegendo a Inteligência Conversacional contra Manipulações e Vazamentos.
- **Objetivo de Segurança:** O objetivo central deste projeto é garantir que Large Language Models (LLMs) operem de forma segura, confiável e alinhada aos seus propósitos originais, especialmente em ambientes corporativos. A solução busca definir e implementar boas práticas e mecanismos de proteção para assegurar a **confidencialidade, integridade e disponibilidade** (tríade CIA) da inteligência artificial generativa.
    - **Confidencialidade:** Prevenir ataques de *prompt injection* que resultem no vazamento de informações sensíveis, como dados de clientes, segredos comerciais, propriedade intelectual ou as próprias configurações do sistema (vazamento de prompt).
    - **Integridade:** Impedir a manipulação indevida das respostas do LLM, garantindo que o modelo não seja coagido a gerar desinformação, conteúdo malicioso ou a quebrar suas políticas de segurança (*jailbreaking*).
    - **Disponibilidade:** Mitigar ataques que visam esgotar os recursos do modelo, como a negação de serviço (DoS), que podem degradar a qualidade do serviço para os usuários e gerar custos operacionais elevados.

2. Ameaças e Vulnerabilidades

A crescente integração de LLMs em sistemas críticos introduziu uma nova superfície de ataque na "camada semântica", onde a linguagem natural funciona como uma interface de comando. As principais ameaças que este projeto visa mitigar são:

- **Prompt Injection:** Esta é a vulnerabilidade mais crítica, classificada como o risco número um (LLM01) pelo projeto OWASP Top 10 para Aplicações de LLM. Ocorre quando um atacante insere instruções maliciosas na entrada do usuário, fazendo com que o modelo execute ações não intencionais. A ameaça se manifesta de várias formas:
    - **Injeção Direta:** O atacante insere comandos diretamente no prompt para manipular a resposta imediata.

    - **Injeção Indireta:** As instruções maliciosas são ocultadas em fontes de dados externas (páginas web, documentos, e-mails) que o LLM processa, expandindo a superfície de ataque.
    - **Injeção Armazenada:** O payload malicioso é persistido em uma base de dados e ativado posteriormente por um usuário legítimo.
- **Vazamento de Dados (Data Leakage):** Consequência direta da injeção de prompt, onde o LLM é enganado para revelar informações confidenciais às quais tem acesso. Isso pode incluir desde dados pessoais de clientes (PII) até algoritmos proprietários e estratégias de negócio, resultando em violações de conformidade (LGPD, HIPAA), perdas financeiras e danos à reputação.
- **Quebra de Políticas (*Jailbreaking*):** Refere-se a técnicas que exploram as nuances da linguagem para contornar as salvaguardas de segurança e ética do modelo. O objetivo é forçar o LLM a gerar conteúdo que ele foi treinado para recusar, como discurso de ódio, instruções para atividades ilegais ou desinformação. Técnicas comuns incluem role-playing (como o prompt DAN - *Do Anything Now*), ofuscação de tokens e o uso de sufixos adversários.
- **Falta de Monitoramento:** A ausência de um sistema robusto de auditoria e logs impede a detecção de atividades suspeitas em tempo real. Sem monitoramento, é impossível identificar padrões de ataque, responder a incidentes de forma eficaz ou coletar dados para aprimorar as defesas de forma iterativa.

3. Ambientes Críticos e Clientes

A solução proposta é projetada para ser aplicável a uma vasta gama de organizações, com foco especial em dois segmentos principais:

- **Empresas Corporativas:** Organizações que utilizam LLMs para otimizar operações internas, como:
    - **Chatbots internos:** Para suporte de RH e TI.
    - **Assistentes de código:** Para acelerar o desenvolvimento de software.
    - **Análise de documentos:** Para resumir relatórios, contratos e e-mails. Nesses cenários, o risco principal é o vazamento de propriedade intelectual, dados estratégicos e informações confidenciais de funcionários.
- **Setores Sensíveis:** Indústrias onde a segurança e a precisão dos dados são de máxima importância.
    - **Saúde:** Para proteger dados de pacientes (PHI), evitar a geração de conselhos médicos incorretos e garantir a conformidade com regulamentações rigorosas.
    - **Financeiro:** Para a segurança de informações de clientes, prevenção de fraudes e garantia da integridade de análises financeiras automatizadas.
    - **Governo:** Para a defesa de dados estratégicos, inteligência e informações classificadas, prevenindo a manipulação de sistemas usados para a tomada de decisão em políticas públicas.

4. Mecanismos de Segurança, Requisitos e Tecnologias

O projeto propõe um framework de segurança em camadas, implementando múltiplos mecanismos de defesa para criar uma proteção robusta e resiliente.

- **Mecanismos de Segurança Chave:**
    - **Filtragem e Sanitização de Entrada:** Implementação de uma primeira linha de defesa que analisa todos os prompts antes de serem processados pelo LLM. Isso inclui a remoção de instruções maliciosas conhecidas e a validação do formato da entrada.
    - **Isolamento de Contexto (*Context Isolation*):** Separação estrutural entre as instruções do sistema (confiáveis) e a entrada do usuário (não confiável). Isso é feito através de delimitadores claros e do uso de APIs baseadas em papéis (system, user), dificultando que a entrada do usuário seja interpretada como um comando.
    - **Guardrails de Saída (*Output Guardrails*):** Verificação automática de todas as respostas geradas pelo LLM antes de serem entregues. Este mecanismo escaneia a saída em busca de vazamentos de dados sensíveis e bloqueia conteúdo que viole as políticas de segurança.
    - **Controle de Acesso:** Aplicação do princípio do menor privilégio, restringindo as permissões do LLM e de seus plugins. As interações com sistemas externos (APIs, bancos de dados) devem ser autenticadas e limitadas ao contexto da tarefa.
    - **Auditoria e Logs:** Rastreamento e armazenamento de todas as interações (prompts e respostas) para permitir a detecção de anomalias, a análise de tentativas de ataque e a resposta a incidentes.
- **Requisitos e Tecnologias:**
    - **Requisitos Funcionais:**
        - **Detecção automática de prompts suspeitos:** Utilizando uma combinação de filtros baseados em assinatura e um modelo classificador.
        - **Dashboard de segurança:** Uma interface para visualização em tempo real dos logs, alertas e estado geral da segurança do sistema.
        - **Conformidade Regulatória:** O sistema deve ser projetado para ajudar as organizações a cumprir com a LGPD e a norma ISO 27001, especialmente nos controles relacionados à gestão de riscos e proteção de dados.
    - **Tecnologias Propostas:**
        - **Linguagem e APIs:** Python e a API da OpenAI (ou APIs de outros provedores de LLM).
        - **Frameworks de Validação:** Ferramentas de *red teaming* automatizado como Garak, PyRIT ou `promptfoo` para testes contínuos de vulnerabilidade.
        - **Ferramentas de Observabilidade:** Plataformas para monitoramento e análise de logs de interações com a IA.
    - **Desafios a Serem Endereçados:**
        - **Equilíbrio entre Segurança e Performance:** Filtros excessivamente rigorosos podem aumentar a latência e gerar falsos positivos, prejudicando a experiência do usuário.
        - **Detecção de Ataques Sutis:** Ataques semânticos e contextuais são difíceis de detectar com regras estáticas, exigindo defesas mais inteligentes.

        - **Atualização Contínua:** A segurança de prompts é uma "corrida armamentista" que exige a atualização constante das defesas para combater novas técnicas de ataque.

5. Impacto e Futuro do Projeto

Este projeto não se limita a uma solução técnica isolada; ele se posiciona como uma contribuição fundamental para o ecossistema de IA segura e responsável.

- **Área Emergente:** A segurança de prompts é um dos campos de mais rápido crescimento dentro da segurança da informação, e este projeto aborda uma necessidade crítica e atual do mercado.
- **Pesquisa Acadêmica:** O trabalho desenvolvido tem potencial para se estender em pesquisas sobre IA responsável, contribuindo para a criação de benchmarks e metodologias de avaliação de segurança mais robustas.
- **Governança de IA:** A solução contribui diretamente para a implementação de frameworks de governança, como o NIST AI RMF, e ajuda as organizações a se prepararem para futuras regulamentações, como o AI Act da União Europeia.
- **Framework Open Source:** O projeto pode evoluir para uma solução de código aberto, permitindo que a comunidade colabore e se beneficie de um framework de segurança compartilhado e em constante aprimoramento.