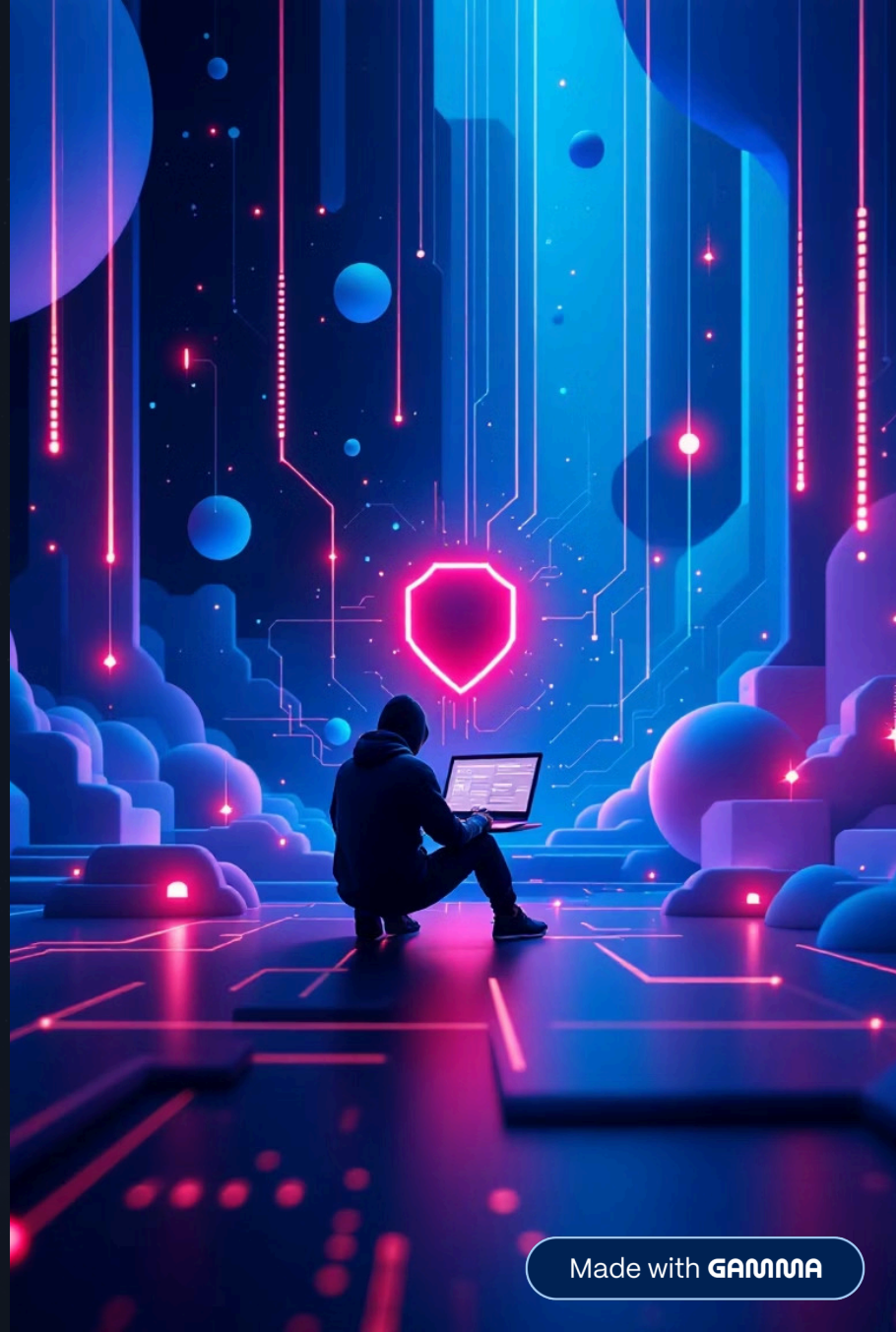
Tema: Segurança de Prompt em Modelos de LLM: Protegendo a Inteligência Conversacional contra Manipulações e Vazamentos

O objetivo é garantir que modelos de linguagem (LLMs) operem de forma segura, prevenindo ataques de *prompt injection*, vazamento de informações sensíveis e manipulação indevida de respostas. O projeto busca definir boas práticas e mecanismos de proteção para uso seguro de IA generativa em ambientes corporativos.

Equipe: Alvaro Miguel, Jóas Vitor, Juan Gustavo, Leonardo Nunes, Lucas Emanoel, Lucas Messias, Lucas Willian, Mauro Vinicius, Vandielson Tenório

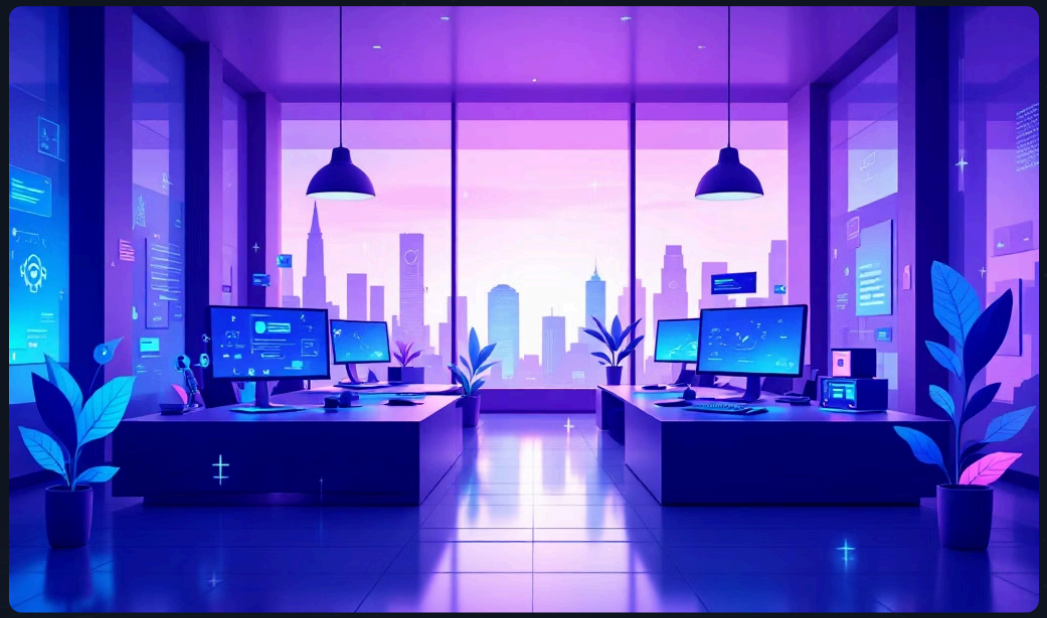
Ameaças e Vulnerabilidades

- Prompt Injection
- Vazamento de Dados
- Quebra de Políticas
- Falta de Monitoramento



Ambientes Críticos e Clientes

Empresas Corporativas



Organizações que utilizam LLMs para chatbots internos, assistentes de código e análise de documentos.

Setores Sensíveis

- **Saúde:** Proteção de dados de pacientes
- **Financeiro:** Segurança de informações de clientes
- **Governo:** Defesa de dados estratégicos

Mecanismos de Segurança

01

Filtragem e Sanitização

Remover instruções maliciosas antes do envio ao modelo.

02

Context Isolation

Separar contexto de usuário do contexto do sistema.

03

Output Guardrails

Verificação automática de respostas antes da entrega.

04

Controle de Acesso

Autenticação e limitação de interações por contexto.

05

Auditoria e Logs

Rastreamento de interações suspeitas.

Requisitos e Tecnologias

Funcionais

- Detecção automática de prompts suspeitos
- Dashboard de segurança
- Integração com LGPD e ISO 27001

Tecnologias

- Python e OpenAI API
- Frameworks de validação
- AI observability tools

Desafios

- Equilíbrio segurança/performance
- Detecção de ataques sutis
- Atualização contínua

Impacto e Futuro do Projeto

Área Emergente

Segurança de prompt é um campo em rápido crescimento dentro da IA.

Pesquisa Acadêmica

Potencial para extensão em pesquisa e integração com IA responsável.

Governança de IA

Contribui para confiabilidade em ambientes críticos.

Framework Open Source

Evolução para solução compartilhada de segurança de prompts.

> Este projeto estabelece as bases para um futuro mais seguro e confiável na utilização de IA generativa em ambientes corporativos e críticos.