Segurança da Informação aplicada a IA, protegendo LLMs com RAG no ambiente corporativo

Dupla: Bruno Hermeto Guimarães, João Comini César de Andrade

1. INTRODUÇÃO

Embora a arquitetura de Geração Aumentada por Recuperação (RAG) seja usada para tornar os LLMs (Modelo de Linguagem de Grande Escala) mais seguros no ambiente corporativo, ela não é totalmente eficaz. Pelo contrário, ela se torna um novo alvo para ataques eficazes e furtivos como prompt injection e envenenamento de documentos. Quando bem-sucedidos, esses ataques resultam em graves consequências como vazamento de dados e a geração de informações falsas. Por isso, o público-alvo para entender e acabar com essas ameaças são as lideranças executivas e as equipes de TI e segurança da informação, responsáveis por implementar e proteger esses sistemas. No final, os grandes beneficiários de uma implementação segura são a própria empresa, que protege sua reputação e seus dados, e os usuários, que podem confiar na ferramenta para obter informações corretas, seguras e não terão seus dados pessoais vazados.

Acreditamos que, ao adotar uma estratégia de defesa em profundidade ou defesa profunda, é possível minimizar os riscos de segurança associados ao uso de LLMs com RAG em ambientes corporativos. Essa estratégia abrange a aplicação de diversas camadas de segurança, que incluem políticas transparentes, filtragem de documentos, validação das respostas produzidas e isolamento das ferramentas que operam em conjunto para reduzir a superfície de ataque. Estudos recentes mostram que ataques como injeção de prompt e exfiltração de dados geralmente exploram falhas em diversas fases do pipeline, o que ressalta a necessidade de uma arquitetura unificada.

Além disso, as evidências técnicas mostram que a maioria dos ataques ocorre quando o modelo é exposto a comandos maliciosos incorporados em documentos recuperados, o que torna essencial o uso de filtros semânticos e classificadores de risco. A validação da

saída permite detectar padrões suspeitos antes de mostrar ou executar o conteúdo, enquanto o uso de sandboxes garante o controle sobre ações externas. Estratégias como essas já estão sendo testadas em situações reais por empresas como Microsoft e OpenAI, com resultados promissores. Com base nesses indícios, acreditamos que nossa proposta pode oferecer uma solução eficaz, mensurável e viável para aprimorar a segurança sem comprometer a funcionalidade e o desempenho do sistema.

Várias abordagens concorrentes têm sido empregadas na tentativa de reduzir os riscos de segurança em LLMs, porém apresentam limitações significativas. Os chamados guardrails genéricos, que se baseiam em palavras-chave ou classificadores simples, tendem a produzir falsos positivos e não são capazes de identificar ataques mais complexos, como instruções disfarçadas ou ofuscação de comandos. Por outro lado, os detectores de jailbreak dependem de padrões conhecidos e podem ser facilmente contornados por variações sutis. Outro ponto é que os prompts de sistema longos, que buscam estabelecer regras para o modelo, podem ser suprimidos por comandos maliciosos inseridos no contexto recuperado, principalmente em pipelines RAG. Quando o modelo lida com dados sensíveis e ferramentas externas, essas soluções isoladas não oferecem garantias suficientes.

Em ambientes empresariais que exigem confidencialidade, rastreabilidade e conformidade, é imprescindível um controle sólido. Por isso, propomos uma arquitetura que combina diferentes mecanismos de defesa, atuando em conjunto para bloquear ataques antes, durante e depois da geração da resposta. Essa abordagem integrada se mostra mais eficaz e adaptável às complexidades reais do uso de IA em sistemas empresariais, superando as limitações das soluções atuais.

O principal problema a ser resolvido é a questão de segurança em modelos que utilizam a arquitetura RAG, e para resolve-lo é necessário desenvolver e implementar novos mecanismos de defesa para detectar e neutralizar ataques, assim garantindo mais segurança e confiabilidade para garantir a proteção dos dados e do sistema.