

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

Bruno Hermeto Guimarães
João Comini César de Andrade

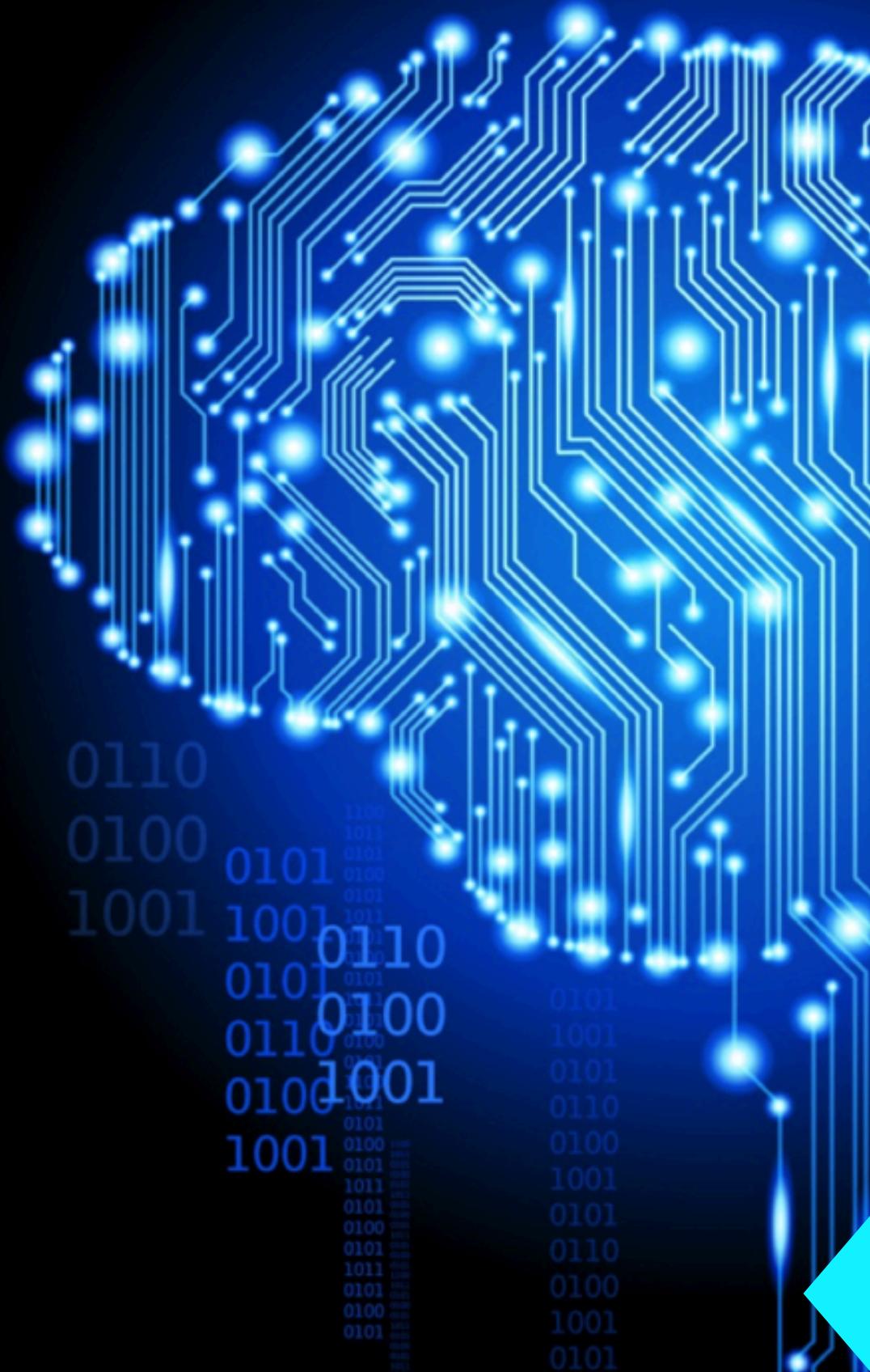


MODELOS DE LINGUAGEM DE GRANDE ESCALA (LLMS)

São sistemas de inteligência artificial projetados para entender, gerar e interagir com a linguagem humana. Eles são a tecnologia por trás de ferramentas de IA generativa

GERAÇÃO AUMENTADA POR RECUPERAÇÃO (RAG)

É uma técnica avançada que torna os LLMs mais poderosos, precisos e confiáveis. O RAG dá a Habilidade para as LLMs de pesquisarem em uma base de dados externa antes de responderem uma pergunta do usuário.



CONTEXTO

Embora a arquitetura de RAG seja usada para tornar as LLMs mais seguras, no ambiente corporativo, ela não é totalmente eficaz, podendo sofrer diversos tipos de ataques. Com isso o público alvo são as lideranças executivas e as equipes de TI e segurança da informação, responsáveis por implementar e proteger esses sistemas.



MOTIVAÇÃO

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.



JUSTIFICATIVA

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.



0110
0100
0101
1001
1000
0111
010
0110
0100
0101
1001

OBJETIVOS

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.



ARTIGOS DE CONFERÊNCIAS

01

USENIX Security Symposium

Qualis A1

Topic-FlipRAG: Topic-Orientated Adversarial Opinion Manipulation Attacks to Retrieval-Augmented Generation Models

02

ACM Conference on Computer and Communications Security (CCS)

Qualis A1

Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation

03

USENIX Security Symposium

Qualis A1

PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models

ARTIGOS DE PERIÓDICOS

01

IEEE Transactions on Information Forensics and Security
Qualis A1
(Security) Assertions by Large Language Models

02

COMPUTERS & SECURITY
Qualis A1
Detecting command injection vulnerabilities in Linux-based embedded firmware with LLM-based taint analysis of library functions

03

IEEE Transactions on Dependable and Secure Computing
Qualis A1
PrivacyAsst: Safeguarding User Privacy in Tool-Using Large Language Model Agents

OBRIGADO