Tema escolhido: Segurança da informação aplicada a IA, protegendo LLMs com RAG no ambiente corporativo.

Dupla: Bruno Hermeto Guimarães, João Comini César de Andrade

Conferências:

- USENIX Security Symposium (Qualis A1): USENIX
  Topic-FlipRAG: Topic-Orientated Adversarial Opinion Manipulation Attacks to Retrieval-Augmented Generation Models: Yuyang Gong, Zhuo Chen, Miaokun Chen, Fengchang Yu, Wei Lu, Xiaofeng Wang, Xiaozhong Liu, Jiawei Liu (https://www.usenix.org/conference/usenixsecurity25/presentation/gong-yuyang)
  Resumo:
  O artigo apresenta um tipo de ataque de manipulação de opinião direcionados a modelos de Geração Aumentada por Recuperação (RAG). O ataque consiste em injetar documentos envenenados na base de dados externa do RAG, esses documentos são sutilmente modificados ou criados para influenciar a resposta do modelo sobre um tópico específico, fazendo adotar uma opinião diferente sem que o usuário perceba. O artigo também mostra que esse ataque é altamente eficaz e furtivo, pois os documentos injetados parecem ser legítimos e relevantes, além de falar da vulnerabilidade crítica na arquitetura RAG e a necessidade de desenvolver defesas mais eficientes contra a manipulação de informações.

- ACM Conference on Computer and Communications Security (CCS) (Qualis A1): ACM
  Riddle Me This! Stealthy Membership Inference for Retrieval-Augmented Generation: Ali Naseh, Yuefeng Peng, Anshuman Suri, Harsh Chaudhari, Alina Oprea e Amir Houmansadr (https://arxiv.org/abs/2502.00306)
  Resumo:
  O artigo introduz um Ataque de Inferência de Pertinência (MIA) contra modelos de Geração Aumentada por Recuperação (RAG).

O método permite que uma pessoa descubra se um documento específico está na base de conhecimento privada do modelo. Para isso, o ataque utiliza perguntas cuidadosamente elaboradas em linguagem que usuários comuns utilizam, no qual a resposta só pode ser encontrada no documento alvo. Os experimentos mostram que o ataque é altamente eficaz, furtivo por não acionarem nenhuma defesa do modelo e eficiente por requerer poucas perguntas, expondo graves riscos de privacidade em sistemas RAG que lidam com dados e documentos confidenciais.

- USENIX Security Symposium (Qualis A1): USENIX
  PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models: Wei Zou, Runpeng Geng, Binghui Wang e Jinyuan Jia (https://www.usenix.org/system/files/usenixsecurity25-zou-poisonedrag.pdf)
  Resumo:
  Esse artigo também fala sobre o eficaz ataque de envenenamento que corrompe a base de conhecimento de sistemas de Geração Aumentada por Recuperação (RAG). Uma pessoa injeta uma pequena quantidade de documentos maliciosos na base de dados, esses documentos são feitos para serem recuperados por perguntas específicas e, em seguida, forçar o LLM a gerar uma resposta incorreta, escolhida por essa pessoa. Com uma grande taxa de sucesso a pesquisa expõe uma vulnerabilidade crítica na segurança dos RAGs, destacando a necessidade urgente de proteger a integridade de suas fontes de informação.

Periódicos:

- IEEE Transactions on Information Forensics and Security (Qualis A1): IEEE
  (Security) Assertions by Large Language Models: Rahul Kande, Hammond Pearce, Benjamin Tan, Brendan Dolan-Gavitt, Shailja Thakur, Ramesh Karri e Jeyavijayan Rajendran (DOI:

10.1109/TIFS.2024.3372809, LINK: https://ieeexplore.ieee.org/document/10458667/authors)

Resumo:

O artigo explora como LLMs podem criar assertivas de segurança para projetos de hardware, empregando linguagem natural para caracterizar propriedades essenciais. Os autores sugerem um framework que analisa a exatidão e a relevância dessas assertivas em comparação com métodos convencionais. Apesar de o estudo se concentrar na segurança do hardware, ele aponta as limitações dos LLMs em tarefas delicadas e enfatiza a necessidade de validação e supervisão do conteúdo produzido. Esses resultados são significativos para ambientes empresariais que utilizam RAG, uma vez que enfatizam a importância de várias camadas de proteção para prevenir instruções maliciosas ou imprecisas.

- COMPUTERS & SECURITY (Qualis A1): Elsevier

  Detecting command injection vulnerabilities in Linux-based embedded firmware with LLM-based taint analysis of library functions: Junjian Ye, Xincheng Fei, Xavier de Carné de Carnavalet, Lianying Zhao, Lifa Wu e Mengyuan Zhang (DOI: https://doi.org/10.1016/j.cose.2024.103971 LINK: https://www.sciencedirect.com/science/article/abs/pii/S0167404824002761#preview-section-abstract)

  Resumo:

  Este artigo introduz o SLFHunter, uma ferramenta que emprega modelos de linguagem ou seja as LLMs para detectar funções de biblioteca suscetíveis a injeção de comandos em ambientes Linux. Essa metodologia une a análise de fluxo de dados a prompts inteligentes, possibilitando a identificação de vulnerabilidades que ferramentas convencionais geralmente não conseguem fazer a detecção. Os autores mostram que a utilização de LLMs amplia a abrangência da análise e identifica falhas críticas com alta precisão, sem comprometer a eficiência

do tempo de execução. O estudo destaca a capacidade dos LLMs de atuar como aliados na segurança automática de sistemas corporativos e embarcados.

- IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (Qualis A1): IEEE
  PrivacyAsst: Safeguarding User Privacy in Tool-Using Large Language Model Agents: Xinyu Zhang, Huiyu Xu, Zhongjie Ba, Zhibo Wang, Yuan Hong, Jian Liu, Zhan Qin, Kui Ren (DOI: 10.1109/TDSC.2024.3372777 LINK: https://ieeexplore.ieee.org/document/10458329/authors#authors)
  Resumo:
  O artigo apresenta o PrivacyAsst, um sistema desenvolvido para preservar a privacidade dos usuários em agentes baseados em LLMs que empregam ferramentas externas. A solução acompanha as interações entre o modelo e os serviços conectados, detectando e impedindo possíveis vazamentos de dados sensíveis. Os autores mostram que o sistema dá um resguardo aos dados e uma ampla cobertura de proteção com impacto mínimo no tempo de resposta. O estudo destaca a relevância de mecanismos de defesa proativos em empresas que implementam IA generativa integrada aos seus sistemas.