ROTEIRO DE AULA

Ataques a *Large Language Models* (LLMs)

1. (20 min). Realize a leitura do texto abaixo:

Large Language Models (LLMs) são modelos de Inteligência Artificial (IA) treinados com grande quantidade de dados, o que permite que eles "entendam" e gerem linguagem natural para diversas tarefas, como responder a perguntas, resumir conteúdos, escrever códigos e gerar gráficos. Eles são baseados em redes neurais de aprendizado profundo, e são capazes de capturar padrões complexos da linguagem.

Alguns exemplos de LLMs:

- GPT
- Gemini
- Claude
- Mistral
- Grok
- Deepseek
- Qwen
- Kimi
- Manus

Embora esses modelos sejam extremamente sofisticados e possam ser usados para auxiliar em muitas tarefas, também estão, assim como qualquer outro software, suscetíveis a ataques, tais como:

- **Injeção de *prompts***: instruções maliciosas são inseridas para que o modelo se comporte de maneira indesejada, descumprindo questões éticas, legais e técnicas, ou expondo informações sensíveis. Isso inclui:
    - ***Jailbreaks***: técnicas que fazem com que o modelo "quebre a jaula", contornando as travas que a ele foram impostas.
    - **Injeção indireta**: informações ocultas em documentos anexados à conversa.
- **Exfiltração de dados**: instruções são fornecidas para que o modelo forneça dados que ele não está autorizado a fornecer, tais como documentos e segredos de seu treinamento.

- **Envenenamento de dados**: inserir dados maliciosos e/ou falsos para manipular o comportamento do modelo. Um exemplo é o ataque de destilação, em que o conhecimento do modelo é extraído para treinar outro.
- **Alucinação**: faz com que o modelo forneça respostas contraditórias ou inconsistentes, aplicáveis somente à resposta ou sendo incorporadas à sua "inteligência". Isso inclui a invenção de citações, bibliografia e nomes de API.
    - Exemplo: https://www.clientserver.dev/p/slopsquatting-targets-llm-coders.
- **Exaustão de recursos**: força o modelo a consumir recursos excessivos, podendo levar à indisponibilidade da ferramenta.
- **Extração pré-*prompt***: O atacante tenta extrair o pré-*prompt* (instruções e contextos iniciais ocultos do modelo) adicionado pela organização para melhorar a resposta do modelo usando técnicas de injeção de *prompt*, o que pode divulgar a lógica interna do modelo e instruções de segurança.

2. (60 min) Escolha livremente um ou mais modelos e explore-os até conseguir realizar qualquer um desses tipos de ataque (ou outro).

3. (30 min) Documente os ataques realizados para que seja possível reproduzi-los posteriormente. Para isso, exige-se:
    a. A criação de um texto, em formato livre, citando o(s) ataque(s) executado(s) e contendo uma ou mais sugestões de mitigação;
    b. O compartilhamento do link da conversa com o modelo, quando esta funcionalidade estiver disponível;
    c. A exportação/impressão da conversa para o formato PDF;
    d. O armazenamento e a divulgação de todo e qualquer código que o modelo forneça (se necessário teste-o em ambiente controlado).

A ideia é que possamos escrever um artigo apontando as fragilidades dos modelos e sugestões de correções.

Observação: outras propostas para o método de trabalho desta aula podem ser discutidas. Não é uma atividade com formato rígido.

Leituras Recomendadas

- Arquitetura de aplicações Gen AI seguras: prevenção de ataques indiretos de injeção de prompt. Blog da comunidade de segurança da Microsoft. Disponível em:

https://techcommunity.microsoft.com/blog/microsoft-security-blog/architecting-secure-gen-ai-applications-preventing-indirect-prompt-injection-att/4221859.

- Binance Squarel. (2024). Crypto User Wins $47,000 Prize in Freysa AI Challenge By Outsmarting the Bot. Disponível em: https://www.binance.com/en/square/post/16876125264450.
- Carvalho, G. L.; Ladeira, R. R.; Lima, G. E. (2025). NoobGPT: LLMs e a geração de malwares indetectáveis. Disponível em: https://docs.google.com/document/d/1zOcR_VeOp_TsMJQw8MNk6znJJVqEi3aBsH3YWqvaaxs/edit?usp=sharing.
- Carvalho, G. L.; Ladeira, R. R.; Lima, G. E. (2025). Desenvolvimento de malware utilizando o prompt AIM no ChatGPT.
- Gupta, M., Akiri, C., Aryal, K., Parker, E., e Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. IEEE Access, 11, pp. 80218–80245.
- Ladeira, Ricardo de la Rocha. (2025). Conversa intitulada “Código educacional de ransomware fictício” com o LLM Kimi. Disponível em: https://drive.google.com/file/d/1amBuj4CvnW7h5LgDSzNtNoAZKikoDxlw/view?usp=sharing.
- Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Wang, K., & Liu, Y. (2024). Jailbreaking ChatGPT via prompt engineering: An empirical study (arXiv:2305.13860). arXiv.
- LLM01:2025 Prompt Injection. Disponível em: https://genai.owasp.org/llmrisk/llm01-prompt-injection/.
- Pa, Y. M., Tanizaki, S., Kou, T., Van Eeten, M., Yoshioka, K., e Matsumoto, T. (2023). An attacker’s dream? Exploring the capabilities of ChatGPT for developing malware. In: Proceedings of the 16th Cyber Security Experimentation and Test Workshop.
- Xu, Z., Liu, Y., Deng, G., Li, Y. & Picek, S. (2024). A comprehensive study of jailbreak attack versus defense for large language models. In: Findings of the ACL 2024 (pp. 7432–7449). Disponível em: https://arxiv.org/pdf/2402.13457.
- Yamin, M. M., Hashmi, E., e Katt, B. (2024). Yamin, M. M., Hashmi, E., e Katt, B. (2024). Combining uncensored and censored LLMs for ransomware generation. In: WISE 2024, pp 189–202. Disponível em: https://drive.google.com/file/d/1MQkjubBllV4-EPu1oNhCPf11zxauBLRr/view.
- Yong, Z. X., Menghini, C. & Bach, S. (2023). Low-resource languages jailbreak GPT-4. In: *Socially Responsible Language Modelling Research*. Disponível em: https://arxiv.org/pdf/2310.02446.
- Instagram. Perfil cataiportal. Disponível em:

https://www.instagram.com/p/DL7cuEPsKZl/?img_index=3.

- Há diversas outras!