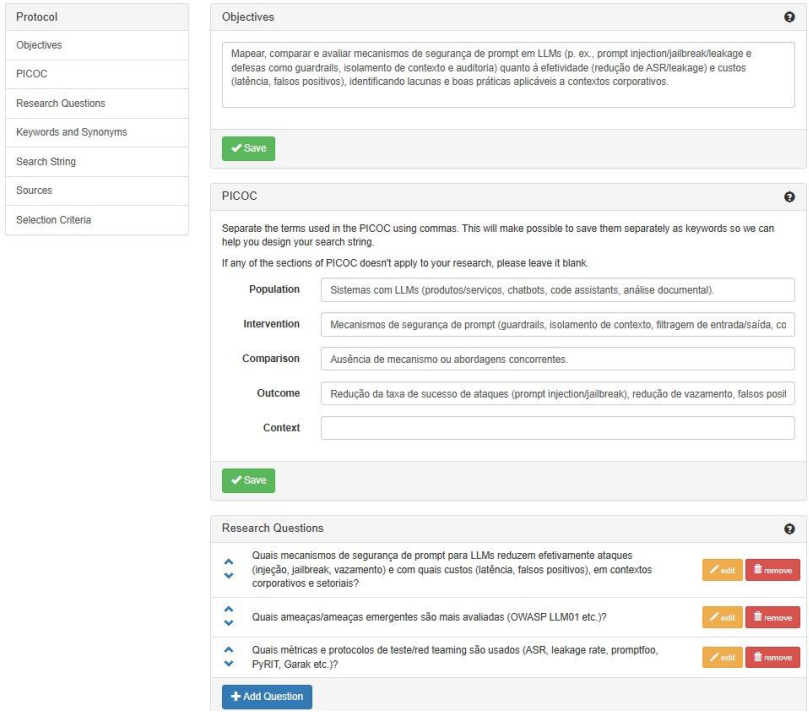
Atividade 2 - Revisão Bibliográfica (Parte I)

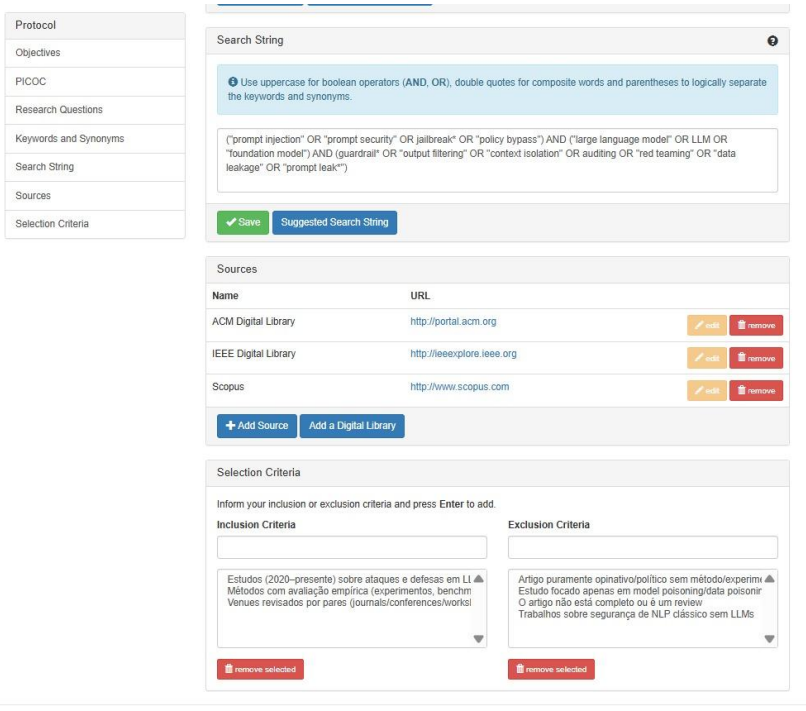
Tema: Segurança de Prompt em Modelos de LLM: Protegendo a Inteligência Conversacional contra Manipulações e Vazamentos

Alunos: Antônio Marcos, Álvaro Gueiros, Leonardo Nunes, Lucas William, Mauro Vinícius, Vandielson Tenório.

Ferramentas Utilizadas

- Parsif.al e Zotero.

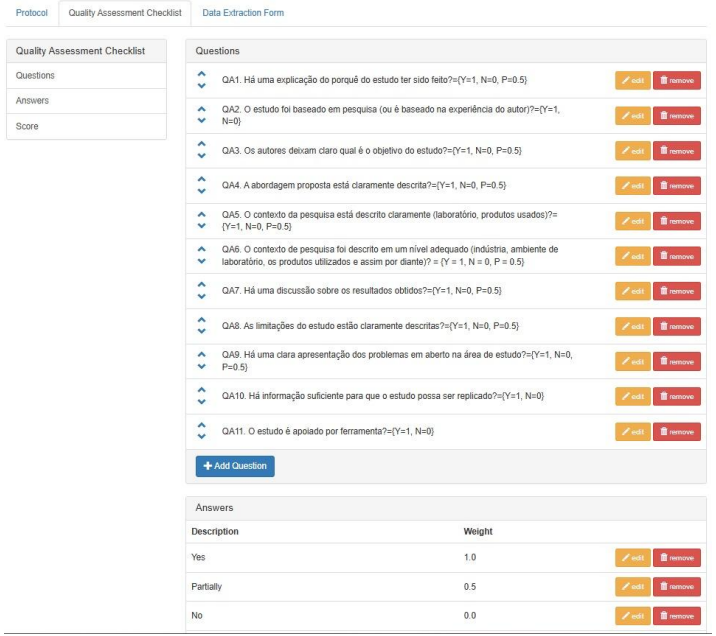


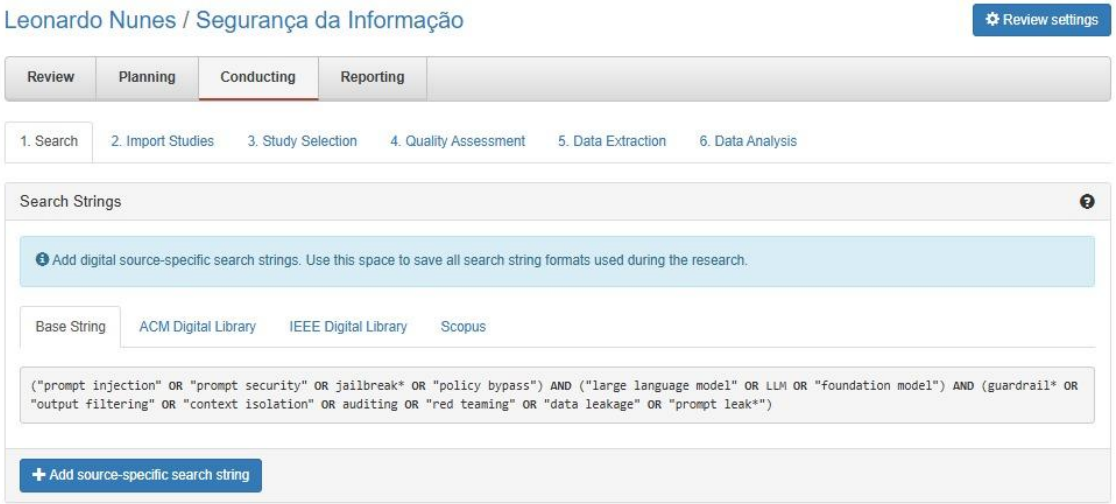


UFAPE
UNIVERSIDADE FEDERAL DO AGRESTE DE PERNAMBUCO

Ferramentas Utilizadas

- Parsif.al e Zotero.





UFAPE
UNIVERSIDADE FEDERAL DO AGRESTE DE PERNAMBUCO

Artigos Selecionados

1. LLM AppHub: A Large Collection of LLM-based Applications for the Research Community (LLM AppHub: Uma ampla coleção de aplicativos baseados em LLM para a comunidade de pesquisa.)

Motivo: Selecionado por oferecer visão abrangente de aplicativos LLM, permitindo mapear superfícies de ataque e posicionar controles em diferentes arquiteturas.

2. Privacy and Security Challenges in Large Language Models (Desafios de privacidade e segurança em grandes modelos de linguagem)

Motivo: Selecionado por consolidar ameaças/defesas e explicitar lacunas métricas, servindo de referência central para o estado da arte.

Artigos Selecionados

3. LLM Security Alignment Framework Design Based on Personal Preference (Estrutura de Alinhamento de Segurança LLM: Design Baseado em Preferências Pessoais)

Motivo: Selecionado por propor mecanismo estruturado de alinhamento de segurança, permitindo avaliar trade-offs de forma mensurável.

4. Compliance Made Practical: Translating the EU AI Act into Implementable Security Actions (Conformidade na prática: traduzindo a Lei de IA da UE em ações de segurança implementáveis.)

Motivo: Selecionado por traduzir requisitos regulatórios em ações concretas, informando desenho de controles e métricas de conformidade.

5. Privacy-Preserving Healthcare Data Security Using Large Language Models and Adaptive Access Control (Segurança de dados de saúde com preservação da privacidade usando modelos de linguagem amplos e controle de acesso adaptativo.)

Motivo: Selecionado por evidenciar aplicação setorial de alto risco com controles mensuráveis, conectando defesa e resultado.

Referências Bibliográficas

Bunzel, Niklas. “Compliance Made Practical: Translating the EU AI Act into Implementable Security Actions”. 2025 IEEE/ACM International Workshop on Responsible AI Engineering (RAIE), 2025, p. 69–73, https://doi.org/10.1109/RAIE66699.2025.00016.

Rathod, Vishal, et al. “Privacy and Security Challenges in Large Language Models”. 2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC), 2025, p. 00746–52, https://doi.org/10.1109/CCWC62904.2025.10903912.

Sun, Zhendan, e Ruibin Zhao. “LLM Security Alignment Framework Design Based on Personal Preference”. Proceeding of the 2024 International Conference on Artificial Intelligence and Future Education [New York, NY, USA], AIFE ’24, 2025, p. 6–11, https://doi.org/10.1145/3708394.3708396.

Wu, Zixuan, et al. “LLMAppHub: A Large Collection of LLM-based Applications for the Research Community”. Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering [New York, NY, USA], FSE Companion ’25, 2025, p. 1254–55, https://doi.org/10.1145/3696630.3731439.

Yarram, Srimaan, et al. “Privacy-Preserving Healthcare Data Security Using Large Language Models and Adaptive Access Control”. 2025 IEEE World AI IoT Congress (AIIoT), 2025, p. 0854–60, https://doi.org/10.1109/AIIoT65859.2025.11105296.

UFAPE
UNIVERSIDADE FEDERAL DO AGRESTE DE PERNAMBUCO