



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.

Protocol

Objectives

PICOC

Research Questions

Keywords and Synonyms

Search String

Sources

Selection Criteria

Objectives

Mapear, comparar e avaliar mecanismos de segurança de prompt em LLMs (p. ex., prompt injection/jailbreak/leakage e defesas como guardrails, isolamento de contexto e auditoria) quanto à efetividade (redução de ASR/leakage) e custos (latência, falsos positivos), identificando lacunas e boas práticas aplicáveis a contextos corporativos.

Save

PICOC

Separate the terms used in the PICOC using commas. This will make possible to save them separately as keywords so we can help you design your search string.
If any of the sections of PICOC doesn't apply to your research, please leave it blank.

Population

Sistemas com LLMs (produtos/serviços, chatbots, code assistants, análise documental).

Intervention

Mecanismos de segurança de prompt (guardrails, isolamento de contexto, filtragem de entrada/saída, co

Comparison

Ausência de mecanismo ou abordagens concorrentes.

Outcome

Redução da taxa de sucesso de ataques (prompt injection/jailbreak), redução de vazamento, falsos posi

Context

Save

Research Questions

Quais mecanismos de segurança de prompt para LLMs reduzem efetivamente ataques (injeção, jailbreak, vazamento) e com quais custos (latência, falsos positivos), em contextos corporativos e setoriais?

Quais ameaças/ameaças emergentes são mais avaliadas (OWASP LLM01 etc.)?

Quais métricas e protocolos de teste/red teaming são usados (ASR, leakage rate, promptfoo, PyRIT, Garak etc.)?

Add Question

Protocol

Objectives

PICOC

Research Questions

Keywords and Synonyms

Search String

Sources

Selection Criteria

Search String

Use uppercase for boolean operators (AND, OR), double quotes for composite words and parentheses to logically separate the keywords and synonyms.
("prompt injection" OR "prompt security" OR jailbreak" OR "policy bypass") AND ("large language model" OR LLM OR "foundation model") AND (guardrail" OR "output filtering" OR "context isolation" OR auditing OR "red teaming" OR "data leakage" OR "prompt leak")

Save Suggested Search String

Sources

Name	URL		
ACM Digital Library	http://portal.acm.org	edit	remove
IEEE Digital Library	http://ieeexplore.ieee.org	edit	remove
Scopus	http://www.scopus.com	edit	remove

Add Source Add a Digital Library

Selection Criteria

Inform your inclusion or exclusion criteria and press Enter to add.

Inclusion Criteria

Exclusion Criteria

Estudos (2020–presente) sobre ataques e defesas em LI
Métodos com avaliação empírica (experimentos, benchm
Venues revisados por pares (journals/conferences/works)

Artigo puramente opinativo/político sem método/experim
Estudo focado apenas em model poisoning/data poisonin
O artigo não está completo ou é um review
Trabalhos sobre segurança de NLP clássico sem LLMs

remove selected

remove selected

Ferramentas Utilizadas

- Parsif.al e Zotero.

Protocol Quality Assessment Checklist Data Extraction Form

Quality Assessment Checklist

Questions	Answers	Score
QA1. Há uma explicação do porquê do estudo ter sido feito?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA2. O estudo foi baseado em pesquisa (ou é baseado na experiência do autor)?=(Y=1, N=0)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA3. Os autores deixam claro qual é o objetivo do estudo?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA4. A abordagem proposta está claramente descrita?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA5. O contexto da pesquisa está descrito claramente (laboratório, produtos usados)?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA6. O contexto de pesquisa foi descrito em um nível adequado (indústria, ambiente de laboratório, os produtos utilizados e assim por diante)?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA7. Há uma discussão sobre os resultados obtidos?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA8. As limitações do estudo estão claramente descritas?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA9. Há uma clara apresentação dos problemas em aberto na área de estudo?=(Y=1, N=0, P=0.5)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA10. Há informação suficiente para que o estudo possa ser replicado?=(Y=1, N=0)	<input type="checkbox"/> add <input type="checkbox"/> remove	
QA11. O estudo é apoiado por ferramenta?=(Y=1, N=0)	<input type="checkbox"/> add <input type="checkbox"/> remove	

[+ Add Question](#)

Answers

Description	Weight
Yes	1.0
Partially	0.5
No	0.0

Leonardo Nunes / Segurança da Informação

[Review settings](#)

Review

Planning

Conducting

Reporting

1. Search

2. Import Studies

3. Study Selection

4. Quality Assessment

5. Data Extraction

6. Data Analysis

Search Strings

[+ Add digital source-specific search strings.](#) Use this space to save all search string formats used during the research.

Base String

ACM Digital Library

IEEE Digital Library

Scopus

("prompt injection" OR "prompt security" OR jailbreak* OR "policy bypass") AND ("large language model" OR LLM OR "foundation model") AND (guardrail* OR "output filtering" OR "context isolation" OR auditing OR "red teaming" OR "data leakage" OR "prompt leak**")

[+ Add source-specific search string](#)

Artigos Seleccionados

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

Artigos Seleccionados

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

Motivo: Seleccionado 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: Seleccionado 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: Seleccionado 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 (AllIoT), 2025, p. 0854–60, <https://doi.org/10.1109/AllIoT65859.2025.11105296>.