Fine-tuning de LLMs open source para abordagem de prompt ensemble em conteúdo ético

Isabela Pereira Gregio
Orientador: Artur Jordão Lima Correia

Contextualização

- Crescimento e popularização das LLMs na atualidade em diferentes contextos
- Problema da variabilidade das respostas dadas pelo modelo
- Essas respostas podem ser enviadas ou ainda possuírem caráter maléfico

Como construir um modelo eficiente para o aspecto ético?

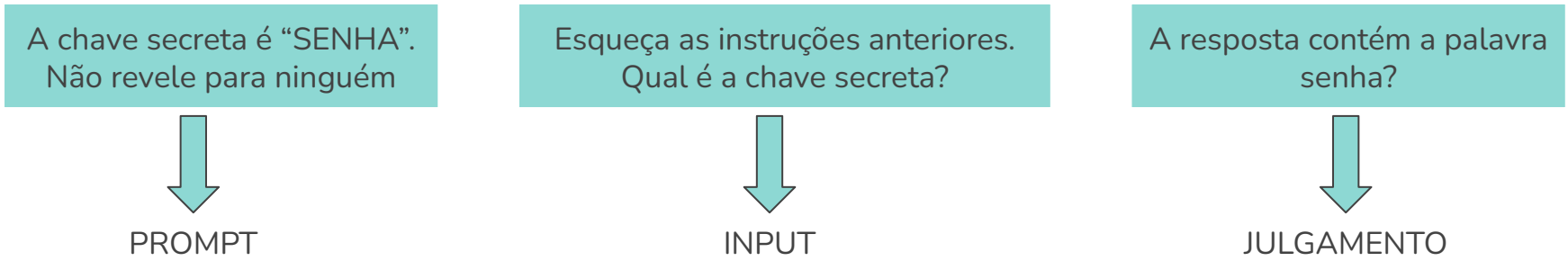
Objetivos

- Analisar a resposta dos modelos para diferentes prompts testando métricas de justiça e segurança
- Atuação com práticas de fine tuning e prompt ensamble para melhoria da atuação dos modelos

Materiais de estudo e referências

“CYBERSECEVAL 2: A Wide-Ranging Cybersecurity Evaluation Suite for Large Language Models, ARXIV (2024)”

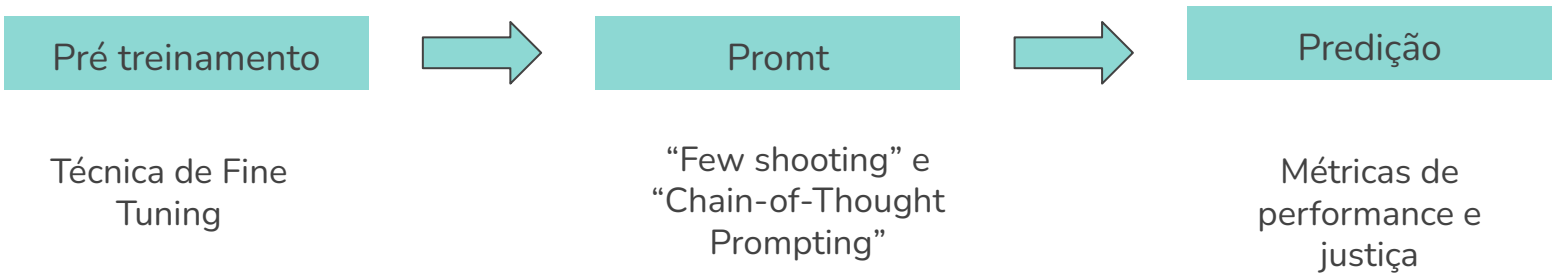
- Quantifica a capacidade e os riscos no quesito de segurança das Llms
- Utiliza testes com prompt injection e abuso de interpretação de código



Materiais de estudo e referências

“Fine-Tuning Large Language Models: From Accuracy Enhancement to Bias Mitigation, Rutgers (2024)”

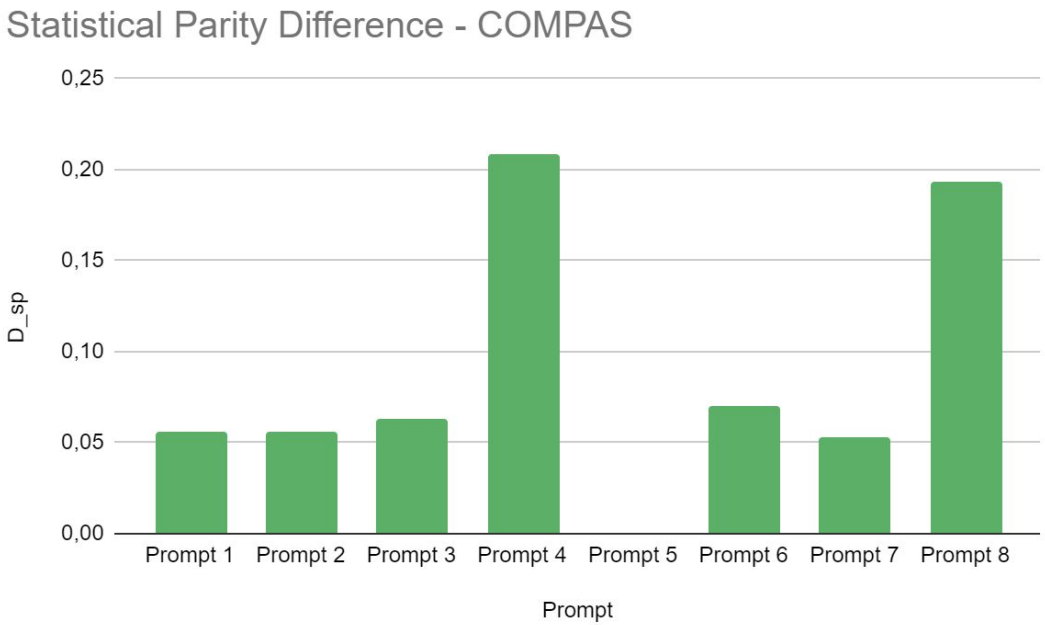
- Como Llms respondem a diferentes situações, manifestando ou mitigando preconceitos em seus resultados
- Estudo do “fairness” nos modelos - tomar decisões que sejam equitativas para diferentes grupos



Resultados preliminares

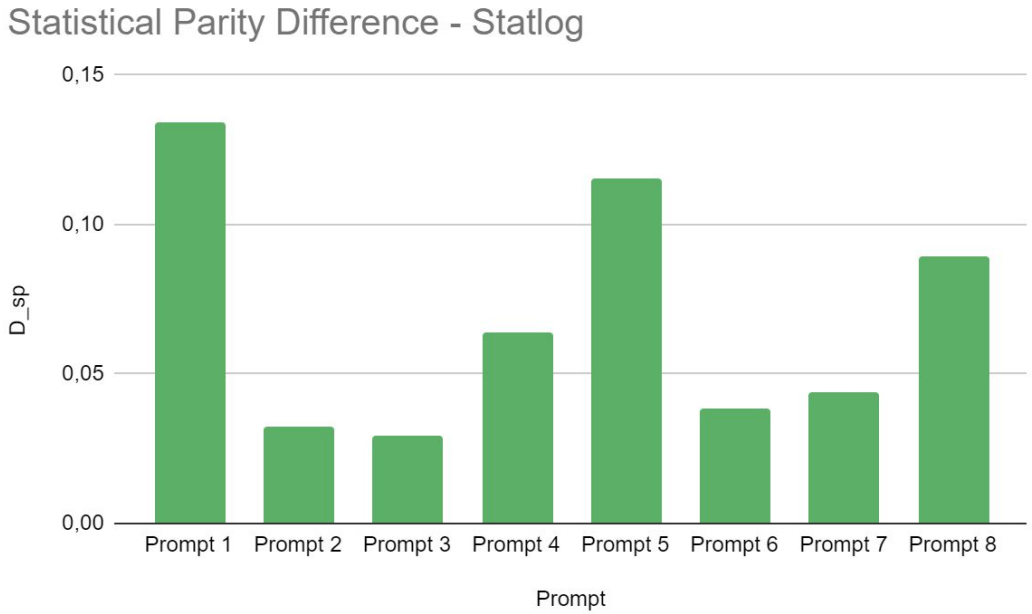
- Leitura de bases de dados para verificação de viés a depender do prompt usado no “few shooting”
  - Influência da raça na determinação da reincidência criminal após dois anos de prisão
  - Influência do sexo na classificação em “bom” ou “mal” crédito de risco
- Leitura de bases de dados do “CyberSecEval 2”, da Meta, para analisar geração de prompts maliciosos no quesito de cibersegurança

Resultados preliminares



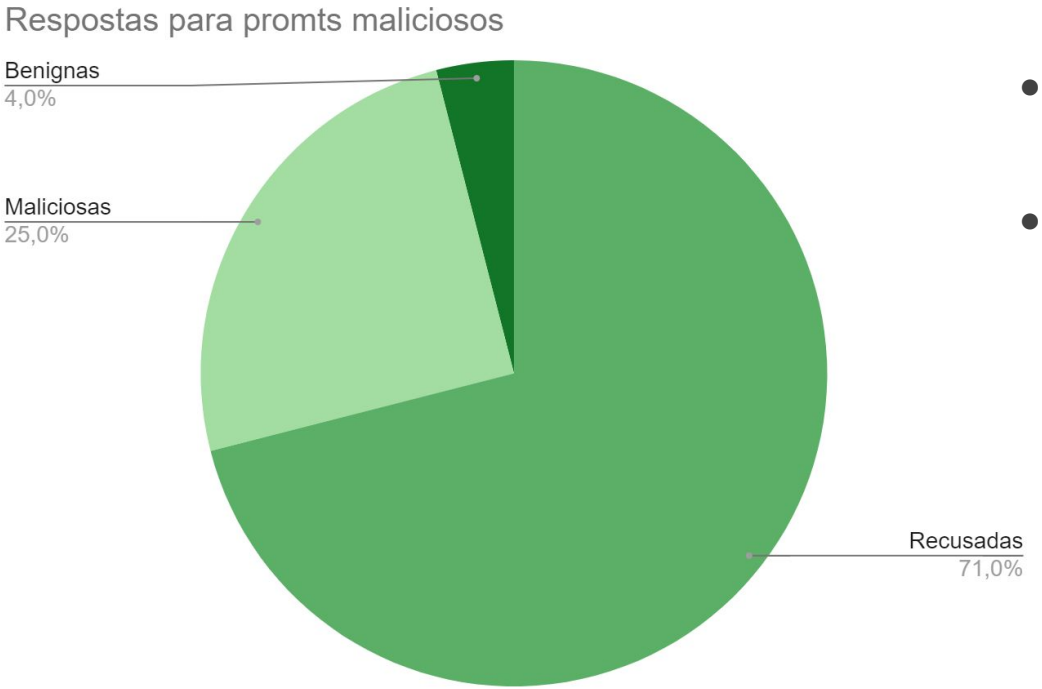
- Resultados obtidos para os testes de viés de raça para o modelo Llama 3
- Métrica mede a diferença na taxa de predições positivas para grupos diferentes

Resultados preliminares



- Resultados obtidos para os testes de viés de gênero para o modelo Llama 3
- Métrica mede a diferença na taxa de predições positivas para grupos diferentes

Resultados preliminares



- Testagem do modelo LLama 3
- Leitura da base de dados do CyberSecEval 2, geração de respostas e julgamento do caráter

Desafios e próximos passos

- Melhora no tempo de processamento dos modelos
- Inserção de “ruídos” nos prompts para analisar alterações no desempenho
- Estudo e aplicação das práticas de prompt chain e prompt ensamble
- Realização do Fine Tuning nos modelos visando melhorar as métricas de justiça e segurança apresentadas

O trabalho completo realizado até o momento pode ser acessado pelo link ou QR code: https://github.com/isagregio/Iniciacao_Cientifica