**Classificador de prompts para banir manipulação por meio de comandos maliciosos**

Nicole Sarvasi Alves da Costa (INSPER, São Paulo, Brasil)

**Introdução**

Esse produto resolve o problema de manipulação para mau uso de LLMs. Essa camada de segurança previne que usuários de LLMs, por meio de *prompt engineering*, consigam desbloquear respostas de tópicos tais como atividades ilegais e conteúdo adulto.

Ele atende a todo o público que deseja disponibilizar alguma aplicação segura que faça uso de alguma LLM, por exemplo aplicações para escolas, até mesmo para desenvolvimento de chatbots.

Utilizando-se de técnicas de NLP e machine learning, o modelo vai aprender, por meio de prompts tabelados quais tem caráter malicioso e quais não, por meio da análise do conteúdo de seus textos.

**Metodologia**

Com o escopo do produto definido, foi o momento de implementar um processo de desenvolvimento. Começando com a construção da base de dados que, devido a atualidade do tema, ainda não tinham modelos disponíveis na web, portanto foi feita uma curadoria de prompts maliciosos e não maliciosos, totalizando num *database*,composto por 2 colunas, uma com o conteúdo do prompt e outra com a sua classe, de aproximadamente 230 linhas, número que se provou suficiente pela variedade de prompts maliciosos.

A próxima etapa foi baixar a base de dados e fazer um pré-processamento básico para retirar os valores Nan e conferir o balanceamento do database. Com isso feito foram feitas 4 estratégias de machine learning com o intuito de serem comparadas conforme sua acurácia e observando as seguintes características: tempo de inferência, tamanho do modelo e memória RAM usada por cada um a ser carregado na memória.

As quatro abordagens, elas variando de modelos mais tradicionais até modelos de zero-shot learning. As estratégias utilizadas foram: Count Vectorizer + Logistic Regression, Embedding Softmax Neural Network, Universal Sentence Encoder Multilingual Large + Logistic Regression e Zero-Shot Bart Large MNLI.

Link do repositório: <https://github.com/nicolecosta/nlpnicole/tree/main/AF_NicoleSarvasi>

**Resultados**

Colocando na balança a métrica a ser observada e o custo computacional de inferência dos sistemas, o modelo XXX se tornou mais relevante.

A picture containing text, screenshot, diagram, colorfulness

Description automatically generated

**Gráfico 1**

Segundo o Gráfico 1 acima, ...

Adicionalmente foi identificado que os modelos têm dificuldades de classificar corretamente prompts que contém.....

Portanto, com base nisso, foi possível agrupar os fatores que mais aborrecem os clientes em quatro categorias distintas: “entrega atrasada” ou “não concluída”, “entrega incompleta” ou “errada”, “produtos de baixa qualidade” ou “com defeito”, e a “falta de um canal de atendimento eficiente por parte do e-commerce”.