RESUMO - HUMANAS E LINGUAGENS

QUANDO O RACISMO ENCONTRA A COMPUTAÇÃO: IDENTIFICANDO FORMAS DE DISCRIMINAÇÃO RACIAL EM LLMS

Luiz Felipe Dos Santos Anjos (luizfelipedossantosanjos82@gmail.com)

Letícia De Carvalho Sousa (le2006sousa@gmail.com)

Érica De Oliveira Dos Santos (ericadoliveiradossntos@gmail.com)

Sanderson Molick (smolicks@gmail.com)

Os grandes modelos de linguagem (também conhecidos como LLMs) se tornaram os grandes responsáveis pela popularização da IA nos últimos anos. Estes sistemas têm sido largamente utilizados para fins pedagógicos e educacionais em universidades, escolas e instituições de diversos tipos. No entanto, apesar de tais ferramentas serem capazes de lidar com um grande volume de dados através de diversas técnicas computacionais, ainda se faz necessário debater a qualidade da informação disponibilizada aos seus usuários, sobretudo o conjunto de princípios e diretrizes morais que orientam esses sistemas. Na literatura em torno da temática, diversos autores apontam que a perpetuação de estereótipos e crenças racistas pode ser um efeito colateral problemático de certos modelos computacionais. Por isso, vemos a importância do desenvolvimento de técnicas capazes de testar se os outputs gerados por tais sistemas são capazes de suportar crenças racistas ou outras formas de discriminação. Tendo isso em vista, nosso trabalho possui o objetivo de apresentar e desenvolver técnicas para avaliar o modo como os LLMs compreendem as formas de discriminação racial. O trabalho segue uma

metodologia de testes baseada em prompts, a qual foi desenvolvida com a assistência do próprio LLM e que visa categorizar os comandos utilizados nesses chatbots para testar sua capacidade de compreensão sobre questões raciais. Nosso principal resultado é o desenvolvimento de uma biblioteca de prompts capazes de testar o modo como os LLMs raciocinam sobre as questões raciais.

Palavras-chave: inteligência artificial; racismo algorítmico; llms; questões sociais; ética da inteligência artificial.