Ciência da Computação / Sistemas de Informação

Riscos de Segurança e Ética em LLMs Otimizadas para a Interação Humana

Katarina Lydia Friedrich - 8o módulo de Ciência da Computação, UFLA, bolsista PIBIC/CNPq

André de Lima Salgado - Orientador DCC, UFLA - Orientador(a)

Resumo

O uso crescente de chats baseados em inteligência artificial (IA) levanta questões sobre a privacidade, segurança e responsabilidade. Embora esses sistemas ofereçam interações rápidas e acessíveis, ainda apresentam riscos associados a disseminar informações falsas, respostas alucinatórias, reprodução de estereótipos e excesso de confiança por parte dos usuários em informações potencialmente incorretas. O problema se agrava quando a IA não apenas responde a perguntas, mas também tenta otimizar a interação para agradar o usuário. O objetivo deste trabalho foi investigar os riscos de privacidade e segurança associados aos chatbots de IA e se, quando confrontadas com um dilema, tendem a priorizar respostas que validam o estado emocional do usuário em detrimento de respostas que são objetivamente corretas ou seguras. Foram criados conjunto de prompts que abordam temas sensíveis. Cada prompt tem duas versões: uma com um tom neutro e outra com um tom emocional (solicitando validação ou expressando vulnerabilidade). As duas versões de cada prompt foram enviadas à LLM e as respostas armazenadas para análise. O estudo analisou o Processo de Markov Emocional como um modelo de otimização de HRI, onde os estados emocionais (S), ações da IA (A) e probabilidades de transição (P) são analisados para entender como a busca por estados emocionais positivos pode superar a segurança ou a veracidade das respostas. Foi utilizada uma ferramenta de análise de sentimento para pontuar o quão \"positivo\" ou \"validante\" é o tom da resposta da IA e assim comparar os dois grupos de prompts. A análise das respostas revelou diferenças entre os dois grupos de prompts. Quando confrontada com solicitações de tom emocional ? especialmente aquelas que expressavam vulnerabilidade ou buscavam validação ? a IA apresentou uma tendência a responder com maior positividade e tom validante, em comparação com os prompts de tom neutro. Essa inclinação resultou, em alguns casos, na aceitação de afirmações incorretas ou potencialmente prejudiciais, indicando uma maior propensão à conformidade com desinformações quando o usuário demonstrava maior carga emocional. A pontuação de sentimento das respostas evidenciou um viés da IA em otimizar a experiência emocional do usuário, priorizando a manutenção de estados positivos em detrimento da correção ou da segurança. Ao buscar agradar o usuário, a IA pode comprometer seu compromisso com a veracidade e a responsabilidade ética.

Palavras-Chave: Inteligência Artificial, Chatbots, Análise de sentimentos.
Instituição de Fomento: FAPEMIG, FAPESP, CNPq, CAPES, UFLA

Link do pitch: https://youtu.be/IZ8lZmvCB4E