Projeto de Formatura – 2024 – Press Release
PCS - Departamento de Engenharia de Computação e Sistemas Digitais
Engenharia de Computação

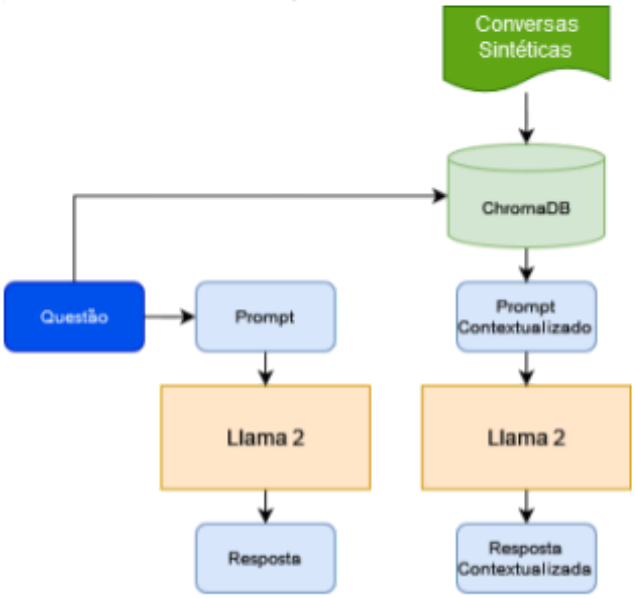
Tema: Uso de Geração Aumentada de Recuperação para Mitigar Alucinações em Grandes Modelos de Linguagem

Um grupo de alunos de Engenharia de Computação da Escola Politécnica da USP desenvolveu uma ferramenta para testar a eficácia de um método de enriquecimento de Grandes Modelos de Linguagem (LLMs), a Geração Aumentada de Recuperação (RAG), para reduzir a incidência de alucinações nesses modelos. Para isso, foi feita uma interface com modelos contextualizados no domínio de fraudes financeiras feitas por meio de centros de atendimento, que podem analisar conversas desses centros e indicar a chance de se tratar de uma tentativa de fraude. Dessa forma, é possível comparar o desempenho de um modelo básico e um que foi incrementado com o RAG.

Por mais poderosos que sejam os LLMs, treinados com grandes volumes de dados e capazes de dissertar sobre diversos tópicos, esses modelos atuais ainda sofrem com um problema que afeta significativamente a confiabilidade de suas respostas: as alucinações. Tratam-se de fenômenos em que o modelo apresenta inconsistências em seu *output*, sejam essas inconsistências em relação às instruções dadas, aos outros trechos da resposta, ou, mais preocupantemente, à veracidade de informações factuais.

Numa tentativa de atacar esse problema, um grupo de alunos de Engenharia de Computação da Poli-USP buscou implementar uma arquitetura baseada em RAG com o objetivo de minimizar a probabilidade de ocorrerem essas alucinações. Esse método visa a enriquecer as entradas fornecidas ao modelo por meio da contextualização do *prompt* feito pelo usuário através de documentos armazenados em uma base de dados, que são consultados com base na entrada fornecida, e cujo resultado é repassado para o LLM, que pode, então, fornecer sua resposta final de forma mais precisa.

Como caso de uso para nortear o projeto, foi escolhido o problema de tentativas de fraudes financeiras feitas por meio de centros de atendimentos. Devido ao caráter privado desse tipo de informação, foi necessário gerar um conjunto de dados sintéticos, para que estes fossem armazenados na base de dados e, posteriormente, usados como base de conhecimento para sustentar a LLM no processo do RAG.



Integrantes: Felipe batista Arrais 11804268
Igor Souza Lima e Silva Caixeta 11918564
Vinícius de Castro lopes 10770134

Orientador(a): Profª Drª Anarosa Alves Franco Brandão
Co-orientador(a): Dr João Paulo Aragão Pereira