Projeto de Formatura – 2024

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**

Introdução

Após a popularização do Chat-GPT, o uso dos grandes modelos de linguagem (**LLMs**) para diversos tipos de tarefas ganhou destaque para as empresas e pessoas no mundo todo. Contudo, um fenômeno recorrente nesses modelos gera um empecilho para sua confiabilidade: **a alucinação.**

No contexto de modelos de inteligência artificial, a alucinação é uma resposta do modelo que ou diverge da realidade, ou diverge do requisição feita. Suas causas são diversas, abrangendo desde questões arquiteturais a questões como a qualidade e quantidade de dados de treinameno para o modelo.

Objetivo e Caso de Uso

O objetivo do projeto é desenvolver uma aplicação que utilize Geração Aumentada de Recuperação (RAG, em inglês) para mitigar a ocorrência das alucinações e fornecer uma resposta mais assertiva e confiável para cada prompt do usuário.

Para este fim, o caso de uso a ser estudado são as tentativas de aplicação de fraude em conversas de centrais de atendimento, onde a incorreta classificação do intuito da pessoa será a alucinação a ser mitigada.

Por se tratar de dados de característica privada, gerou-se um conjunto de conversas sintéticas que será a base do aprimoramento das respostas do LLM.

Solução

O projeto foi implementado sobre a biblioteca LangChain, que permite definir uma arquitetura RAG de maneira simples através dos seus objetos.

Utilizou-se o ChromaDB, um banco de dados para *embeddings*, para armazenar as conversas sintéticas geradas e, quando necessário, fornecê-la como contexto para as requisições futuras.

O LLM escolhido foi o LlaMa 2 (versão 7b chat) da Meta, por se tratar de um modelo open-source que poderia ser acessado e utilizado através da API do site Hugging Face de maneira gratuita.

Ao fazer um *prompt* com a conversa a ser avaliada, este é convertido em um embedding e uma busca por similaridade é executada no ChromaDB. As conversas recuperadas são concatenadas ao prompt original, gerando assim o prompt contextualizado que será enviado ao LlaMa.

Para fins de avaliação, gerou-se também uma resposta sem as conversas recuperadas. A figura 1 ilustra a descrição do fluxo.

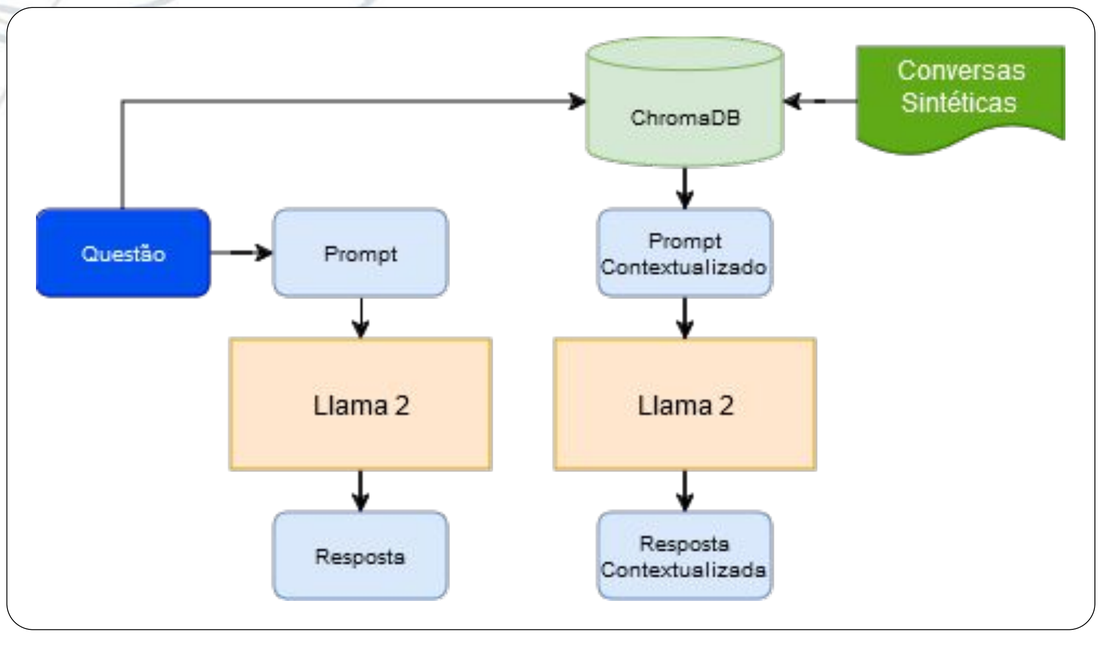

Figura 1: Fluxo de informação na arquitetura

Conclusão

A aplicação se mostrou promissora e, ao limitar o número de documentos fornecidos como contexto, conseguiu realizar uma boa busca mais precisa no banco de *embeddings*. O formato da classificação tendeu a imitar o formato contido nos arquivos base, e isso contribuiu para a clareza da informação.

Integrantes: Felipe Batista Arrais
Igor Souza Lima e Silva Caixeta
Vinicius de Castro Lopes

Profª. Orientadora: Profª. Drª. Anarosa Alves Franco Brandão
Co-orientador: Dr. João Paulo Aragão Pereira