Assistente de Compliance

Um projeto de Retrieval-Augmented Generation (RAG) com Hugging Face

Larissa Pereira Ibiapino

O problema e o caso de uso

Problema: alucinações

Modelos de linguagem comuns (LLMs) não conhecem dados privados de empresas. Ao serem questionados sobre regras internas (como nesse caso o código de ética interno), eles alucinam ou inventam respostas. Isso é inaceitável para áreas que exigem 100% de precisão, como Jurídico e Compliance.

Caso de uso: agente de dúvidas de compliance.

Criar um assistente de IA que funcione como uma ferramenta de consulta segura. O sistema deve ser capaz de ler o Código de Ética da empresa, que esta no formato de PDF e responder perguntas baseando-se exclusivamente no conteúdo do documento, sem inventar informações.

Arquitetura da Solução (Hugging Face)

1. Processamento (pypdf)

O PDF do Código de Ética é lido, limpo e fatiado em blocos de 1200 caracteres para análise.

2. Modelo usado: TinyLlama

O flan-t5 (512 tokens) falhou por ter memória curta. Troquei pelo TinyLlama-1.1B, com janela de 2048 tokens, capaz de ler os blocos de texto inteiros sem cortes, entendendo o contexto completo.

3. O Prompt

Usei um "System Prompt" para forçar o Llama a falar sempre Português, utilizar para respostas apenas o contexto passado e usar uma resposta padronizada caso não tenha contexto suficiente, se atentar em palavras-chave como "Proibido" ou "Vedado".

Conclusão e aprendizados

- **Missão Cumprida:** O RAG provou ser a arquitetura correta. O sistema final responde com precisão, baseado apenas no documento-fonte, eliminando alucinações, com exceção do idioma, mesmo com o prompt para que responsa apenas em Português do Brasil, ele se perde algumas vezes e responde a perguntas em inglês.
- **Desafio do Modelo:** O modelo é pequeno mas exige muita máquina, então a depender da máquina, pode levar bastante tempo para responder as perguntas. E, tambem, trabalhar com idioma português foi um desafio, já que em determinadas palavras o contexto pode ficar desconexo.
- **Valor Real:** A IA deixa de ser um "chatbot genérico" e se torna uma ferramenta de negócios poderosa, capaz de fornecer respostas de Compliance precisas e recebendo perguntas mais abertas desviando dos bloqueios de um chatbot desenvolvido por árvore de decisões, por exemplo.