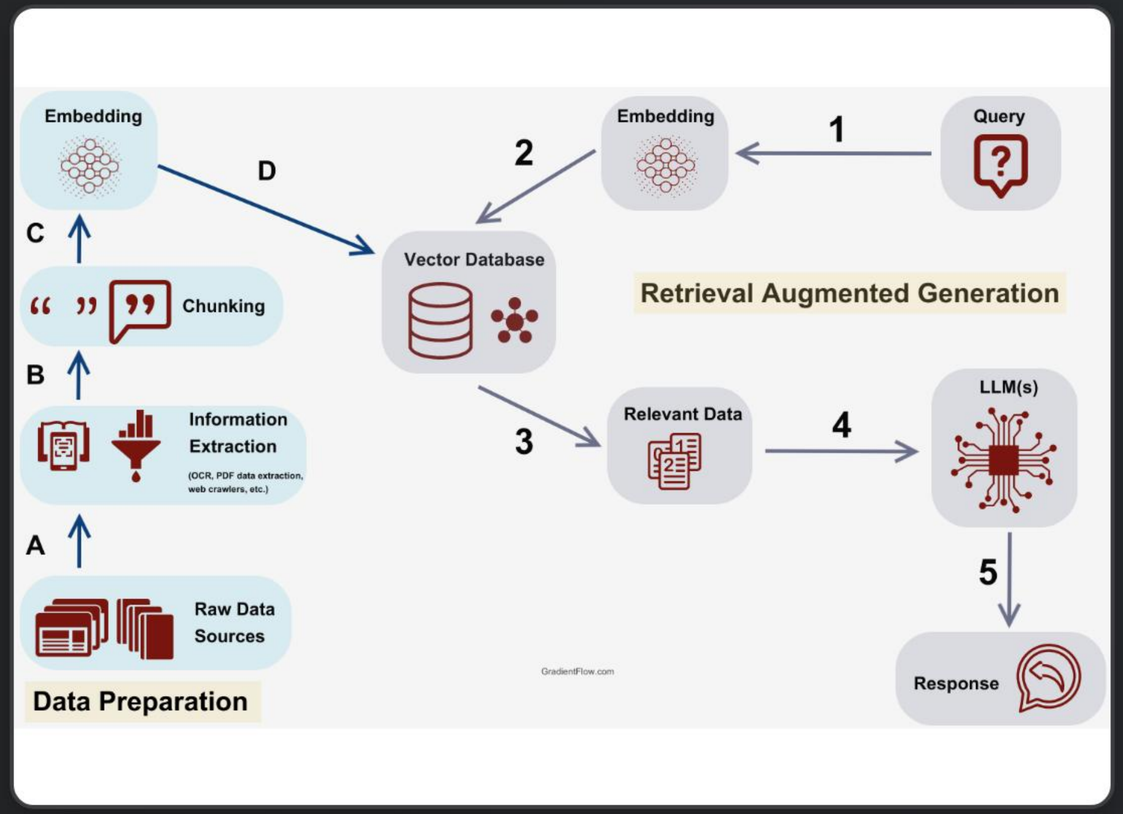
O Problema: Limitação de Conhecimento

Modelos de Linguagem (LLMs) como o GPT ou T5 são treinados em dados públicos da internet. Eles enfrentam dois grandes desafios:

- **Falta de Acesso Local:** Eles não conhecem seus arquivos privados, como o `Elite_da_tropa.txt` mencionado no código.
- **Alucinação:** Se forçados a responder sobre algo que desconhecem, eles podem inventar informações convincentes, mas falsas.
- **Janela de Contexto:** Não é possível simplesmente "colar" gigabytes de texto no prompt de uma só vez.

A Solução: Arquitetura RAG



O código implementa **RAG (Retrieval-Augmented Generation)**, que conecta seus dados ao LLM em três etapas:

- **1. Ingestão (Chunking):** O arquivo de texto é carregado e dividido em pedaços menores (chunks) de 500 caracteres para facilitar a busca.
- **2. Recuperação (Retrieval):** Quando uma pergunta é feita, o sistema busca no banco de dados vetorial (FAISS) apenas os trechos mais relevantes semanticamente.
- **3. Geração:** O modelo FLAN-T5 recebe a pergunta + os trechos recuperados e gera uma resposta fundamentada no contexto real.

Stack Tecnológico do Código

A

Embeddings

Library: `SentenceTransformer`

Converte os textos em vetores numéricos (embeddings). O modelo `all-mpnet-base-v2` captura o significado semântico das frases, permitindo que o computador "entenda" o conteúdo.

Vector Store

Library: `FAISS`

Uma biblioteca de alta performance do Facebook para busca de similaridade. Ela armazena os vetores gerados e encontra os vizinhos mais próximos (top-k) da pergunta do usuário instantaneamente.

LLM Generation

Library: `Transformers (FLAN-T5)`

O cérebro da operação. O `google/flan-t5-base` é um modelo Seq2Seq que recebe o prompt enriquecido e gera a resposta final em linguagem natural.

Image Sources

https://cdn.analyticsvidhya.com/wp-content/uploads/2020/11/python-libraries-for-data-science.webp

Source: www.analyticsvidhya.com

https://gradientflow.com/wp-content/uploads/2023/10/newsletter87-RAG-simple.png

Source: gradientflow.substack.com