Day 3

Agenda

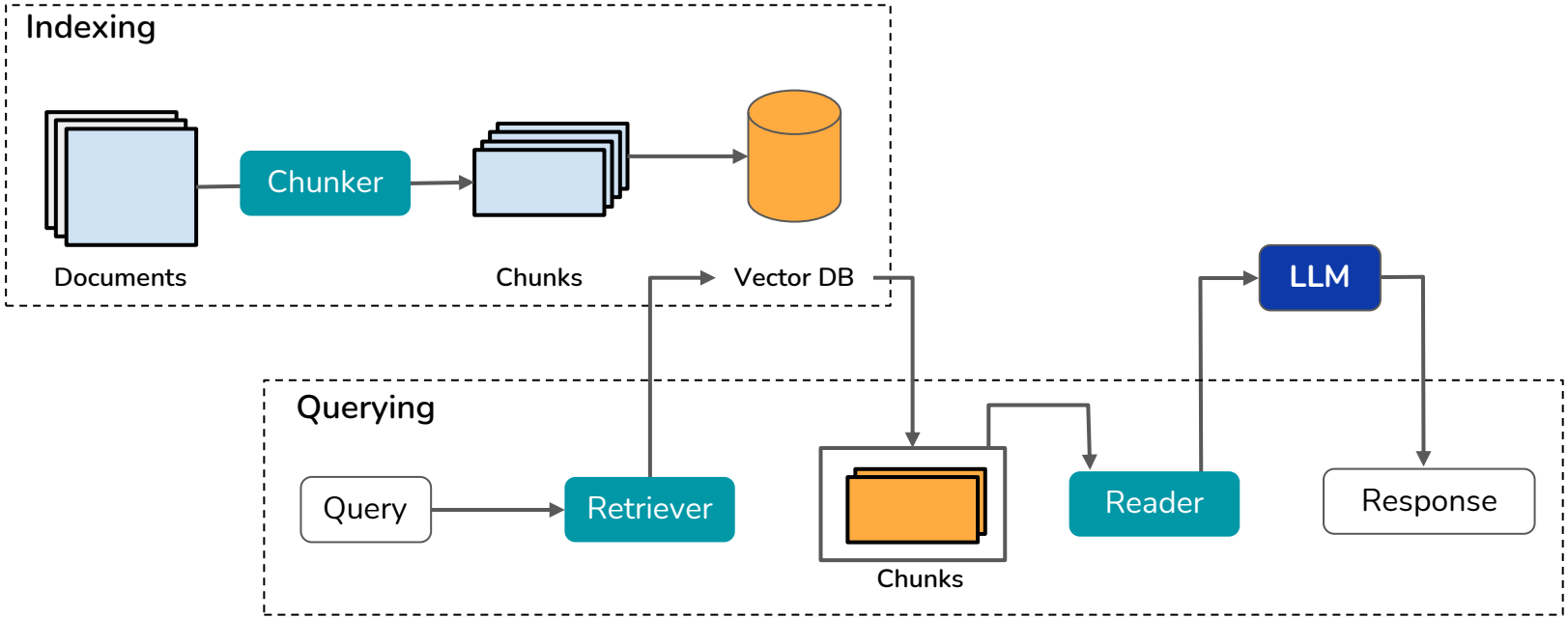
In this session, we will discuss:

- Chunking Strategies for RAG
- Evaluating RAG outputs for groundedness and relevance
- Hands-on Implementation of RAG

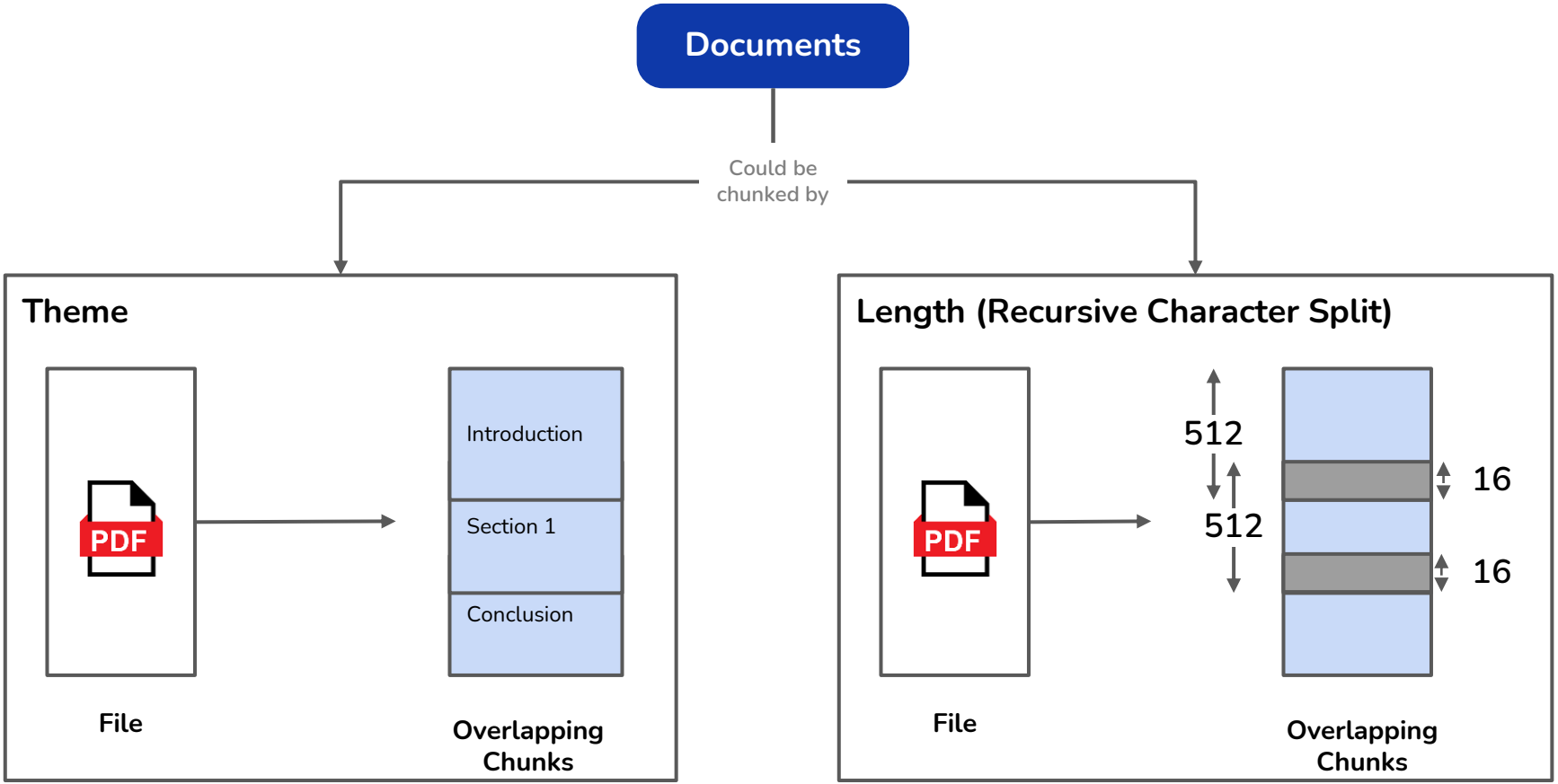
A Workflow for RAG

A Workflow for Retrieval-Augmented Generation

Retrieval-augmented generation (RAG) is a technique that enhances generative AI models by incorporating external data sources to improve accuracy and relevance in text generation



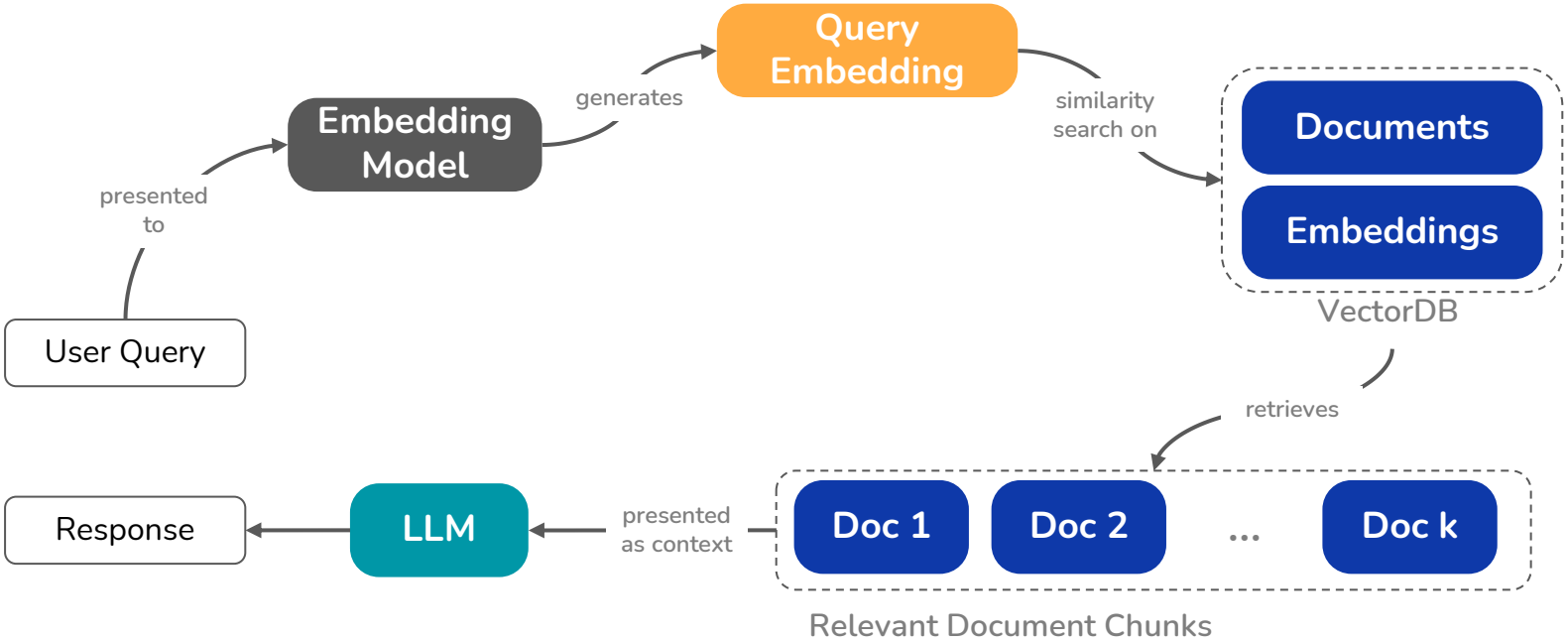
Chunking Strategy



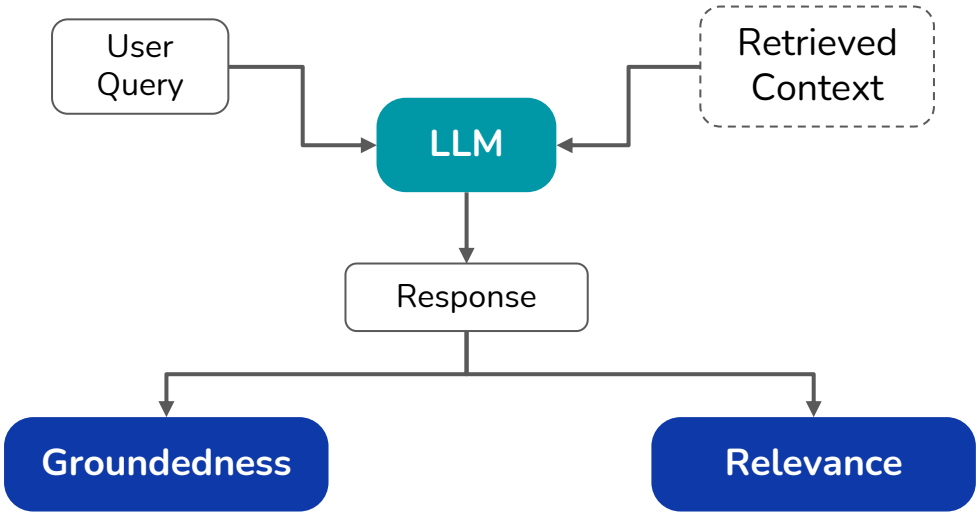
https://chunkviz.up.railway.app/

Retrieval & Generation

Vector databases are specialized in storing and retrieving vectors associated with unstructured data. Given input queries, the database can retrieve relevant documents using similarity search.

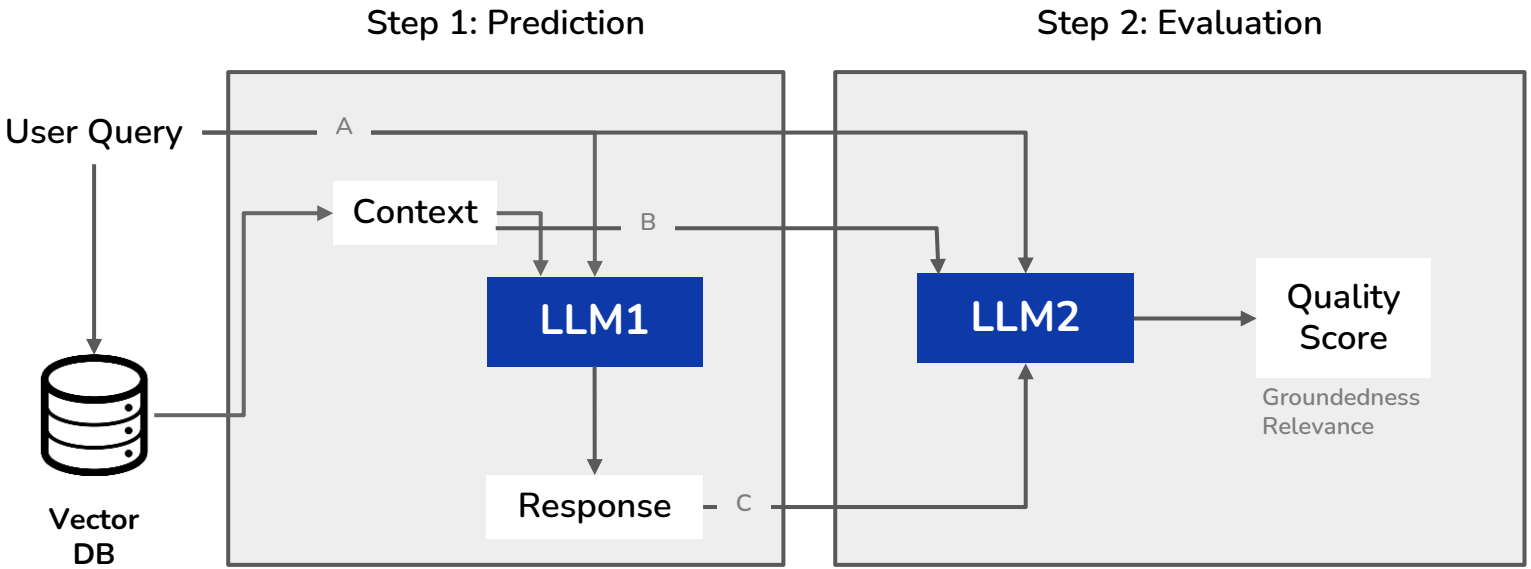


Evaluating RAG Applications

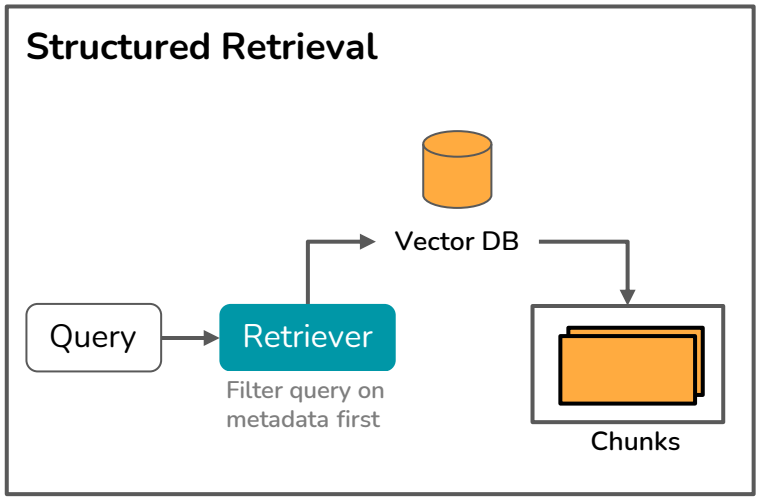


Evaluating RAG with LLM-as-a-Judge

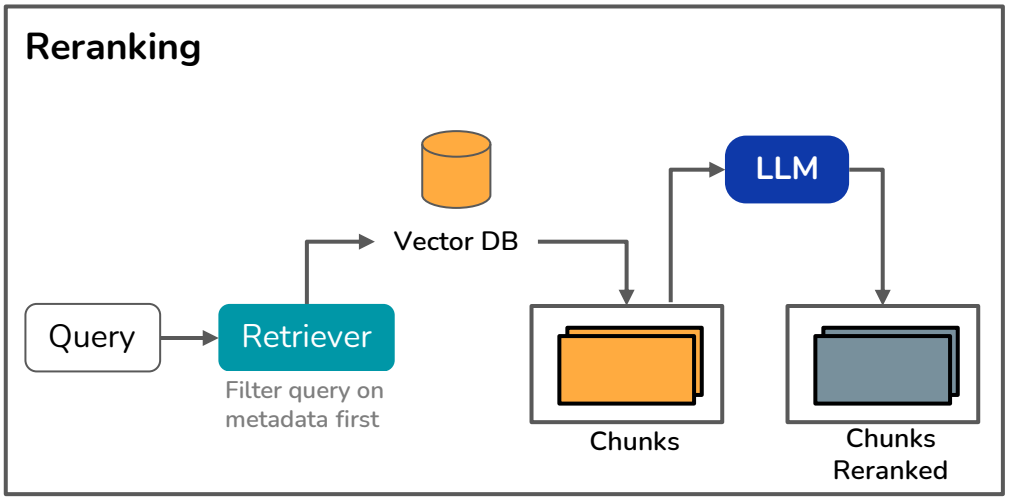
Given input query, context and LLM response, rater LLMs judge whether: (a) The response is grounded in the context, (b) The response is relevant to the query



Advanced Retrieval Techniques



Used when there are many similar documents



Used when evaluation reveals poor relevance scores

RAG Hands-on

- Building a Vector DB on Tesla 10-k documents
- Prompt Engineering for RAG
- Evaluating RAG outputs for groundedness and relevance

Summary