Minimal Local RAG System for Policy Q&A

Retrieval-Augmented Generation using Ollama + ChromaDB

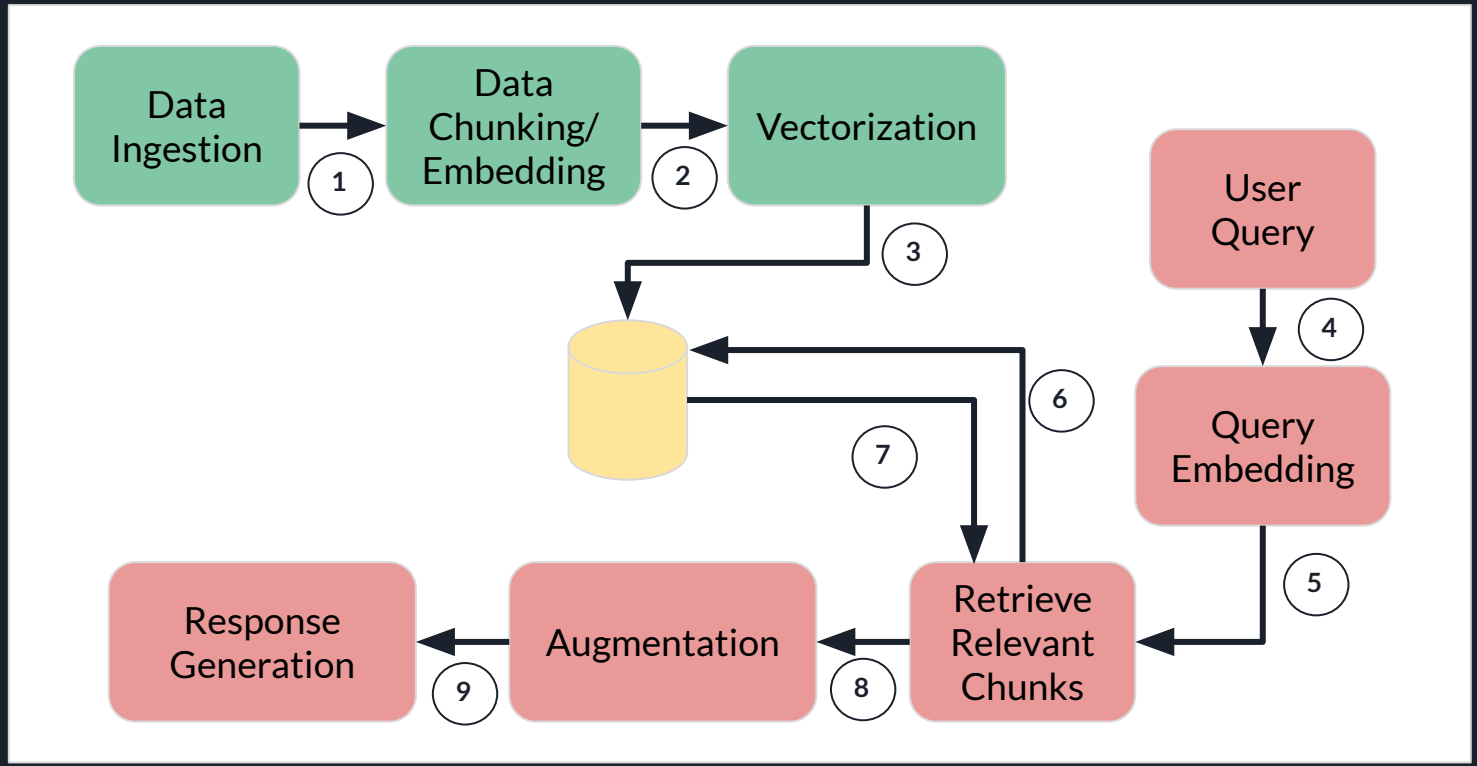
Jose Escobar · Technical Take-Home Project

Problem & Motivation

Why RAG?

- LLMs may hallucinate policy details
- Policy answers must be grounded and traceable
- Users need to know where an answer comes from

High-Level Architecture



Key Design Decisions

Design Choices

- Fully local execution
- Ollama used for:
  - embeddings (nomic-embed-text)
  - generation (mistral:7b-instruct)
- ChromaDB as a local persistent vector store
- Citations required in all answers

Data & Ingestion

Dataset and Preprocessing

- 10 text files (one per chapter)
- Repeated boilerplate headers removed (programatically)
- Fixed-size chunking with overlap
- Low-quality chunks filtered out

Retrieval & Prompt Augmentation

Retrieval and Prompting

- Dense vector similarity search (top-K)
- Retrieved chunks inspected before generation
- Structured prompt enforces:
  - use only retrieved context
  - always include citations
  - Ideally no hallucinations

Tradeoffs & Demo Transition

Tradeoffs and Next Steps

- Not production-ready
- Simple chunking strategy
- No access control or monitoring (next version)
- Easy to extend

DEMO

Thank you!