Retrieval Augmented Generation (RAG) Primer

Irtiza Chowdhury

Outline

- RAG Overview
  - LLM Limitations
  - Retrieval Augmented Generation
- RAG Steps
  - Basic RAG pipeline
  - Indexing
  - Retrieval
  - Generation
- Advanced RAG
  - Indexing strategies
  - Retrieval strategies
  - Generation strategies
- Notebook walkthrough

RAG Overview

LLM Limitations

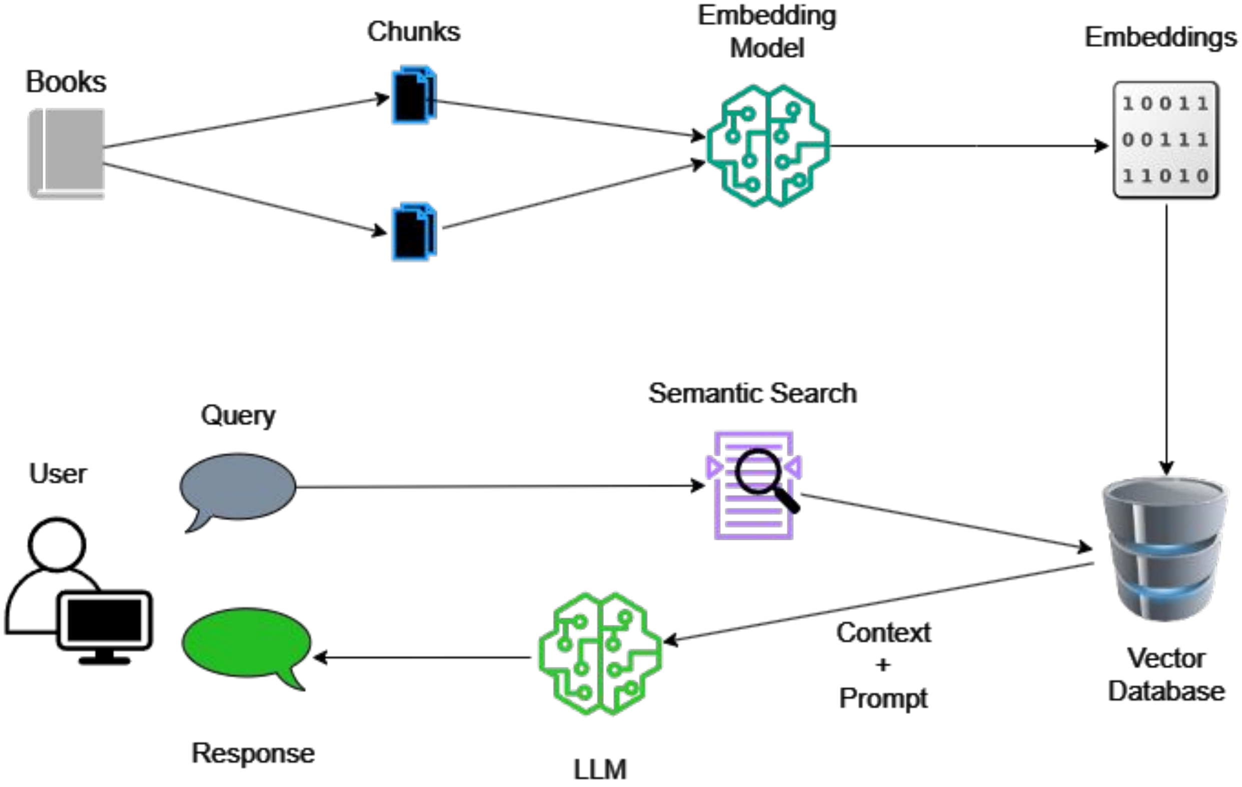
- Lack of domain specific knowledge
- Out of context answers
- No access to confidential data
- Knowledge cut-off period
- Fine-tune?

Retrieval Augmented Generation (RAG)

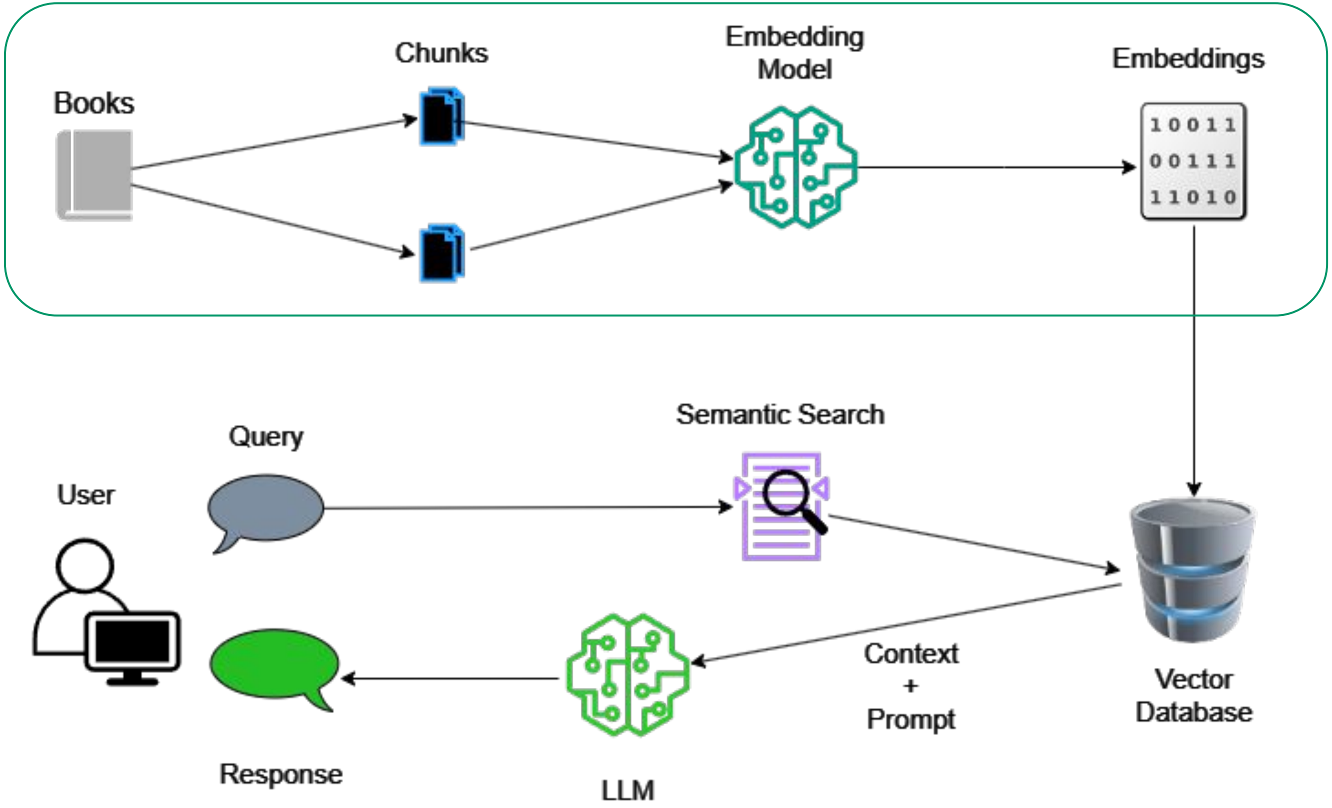
- Adds personalized context to LLMs
- Generate responses on custom data
- Improve factual accuracy of the LLM
- Without altering existing knowledge base

RAG Steps

Basic RAG Pipeline

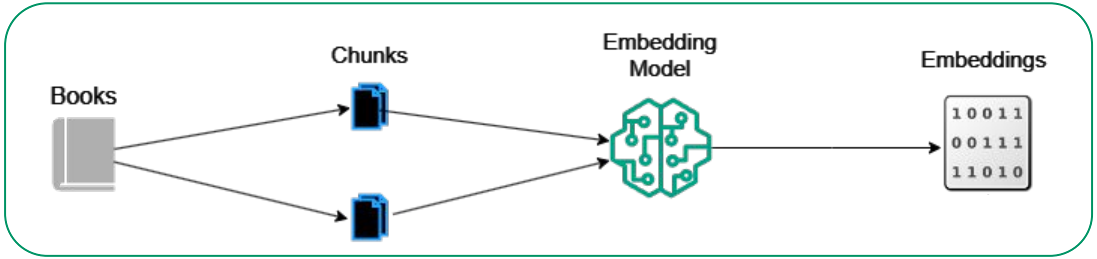


Indexing

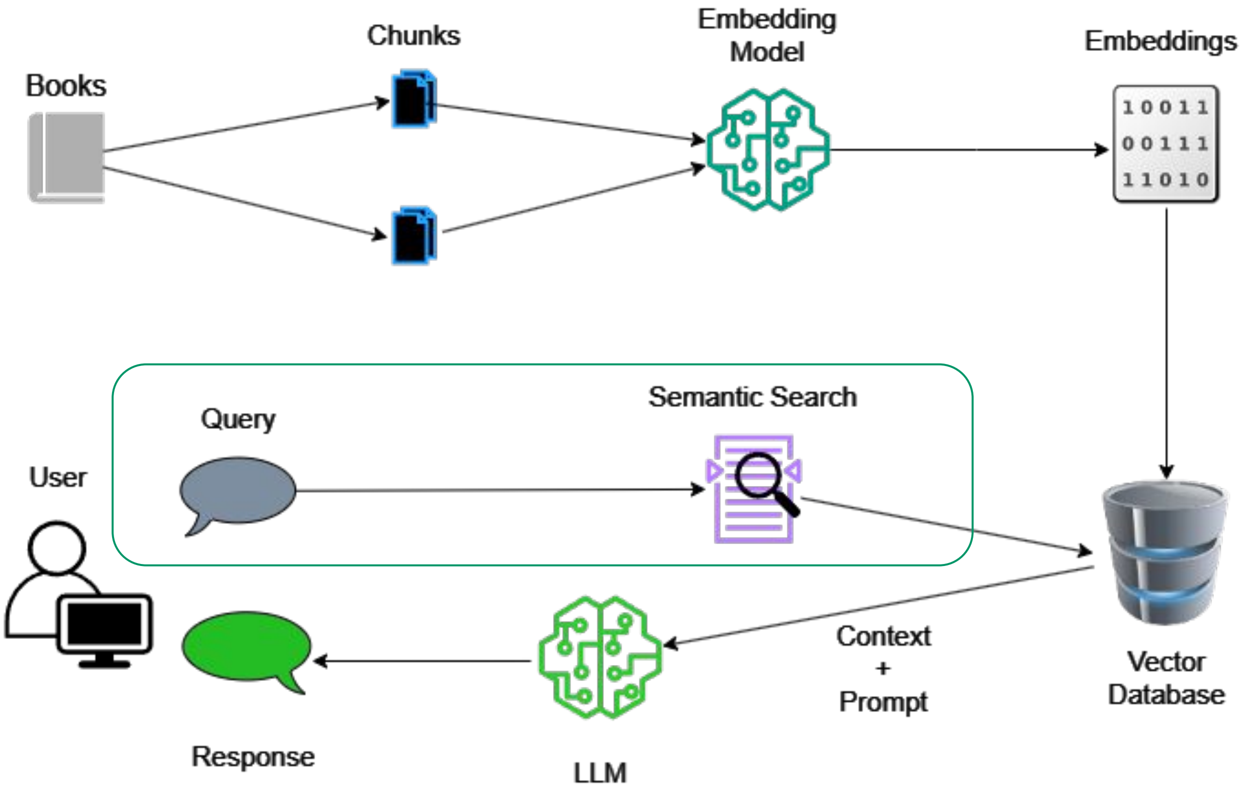

Books
Chunks
Embedding Model
Embeddings
1 0 0 1 1
0 0 1 1 1
1 1 0 1 0
Semantic Search
Query
User
Vector Database
Context + Prompt
LLM
Response

Indexing

- Data preparation
- Chunking
  - Segmenting into smaller texts
  - Fit the context window
  - Better semantic search
- Embeddings
  - Numerical representations (vectors)
  - Retain semantic information
  - Multilingual embeddings
  - Embedding model

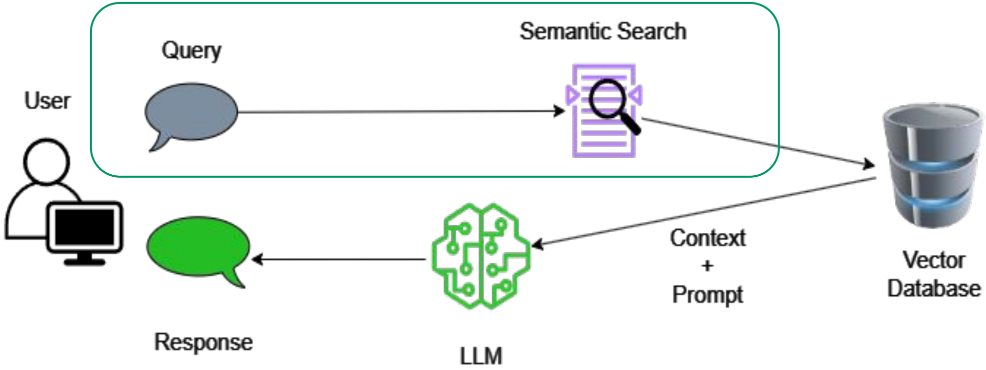


Retrieval

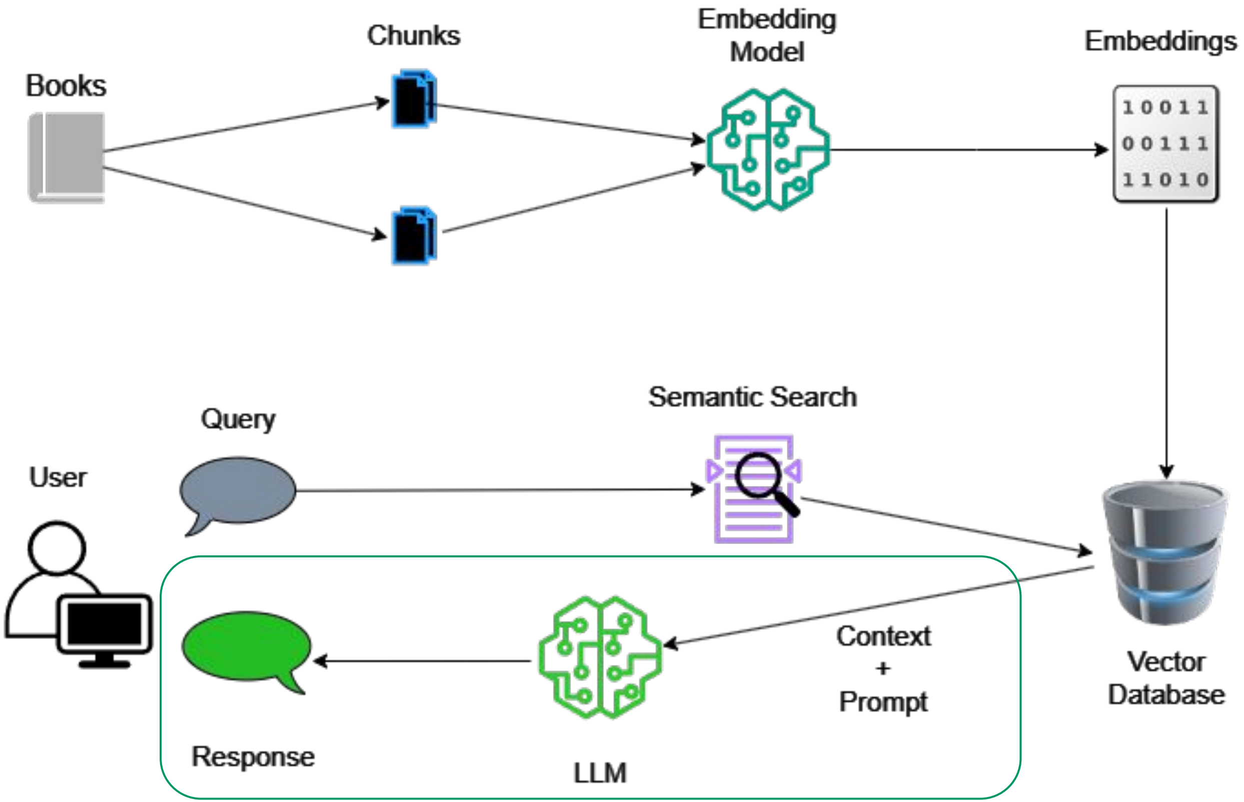

Books
Chunks
Embedding Model
Embeddings
1 0 0 1 1
0 0 1 1 1
1 1 0 1 0
Query
Semantic Search
User
Vector Database
Context + Prompt
LLM
Response

Retrieval

- Vector Database
  - Store the indexed embeddings
  - Prevent re-indexing
- Querying
  - Receive query from user
  - Encode query into vector
  - Similarity score with chunks
  - Return top ‘k’ relevant chunks
  - Used as context for LLM

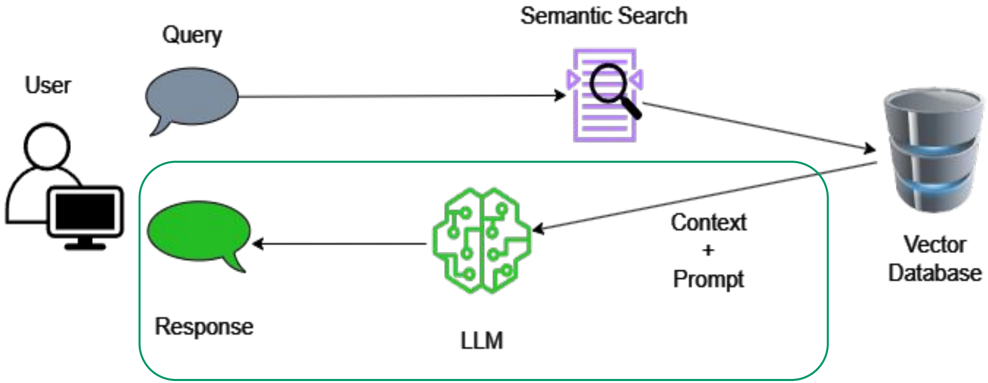


Generation


Chunks
Embedding Model
Embeddings
Books
1 0 0 1 1
0 0 1 1 1
1 1 0 1 0
Semantic Search
Query
User
Context
+
Prompt
Vector Database
Response
LLM

Generation

- Context from retrieval step
    - Relevant chunks
- Specified prompt
- Combined with query
- Provided to LLM
- Task/context-rich response



Advanced RAG

Indexing strategies

- Chunking
- Context too large
  - does not fit window
- Context too small
  - poor answers
- Fit the embedding model
- Choice of embedding models
  - Size, multi-lingual, performance, task etc.
- Metadata
  - Self querying

Retrieval strategies

- Parent-Child retriever
  - Smaller chunks for querying
  - Larger chunks for context
- Multi Query retriever
  - Query might not be semantically similar
  - Variations of the query
  - Use LLM to create these
  - Each query searched against index
- Embedding filter
  - Drop documents below a similarity threshold
- Hybrid Search

Generation strategies

- Choice of LLMs and Prompting
- Reranking
  - Vectors might lose semantic information
  - Use a transformer to compare
  - But comparatively slower
  - Instead retrieve relevant documents
  - Rerank their relevance
  - Cohere API
- Evaluation
- RAGAS
  - Multiple metrics e.g. retriever recall, precision

Notebook