20240509Meeting

312707003 黃鈺婷

Retrieval-Augmented Generation(RAG)

- 檢索式模型+生成式模型
- 利用外部知識來增強生成的文本，使生成的文本更加準確
- 不需要重新訓練模型
- 適用於有大量資料，但多數資料未分類或標記
- 開放領域的問答任務、資訊檢索

RAG vs. Fine-tuning

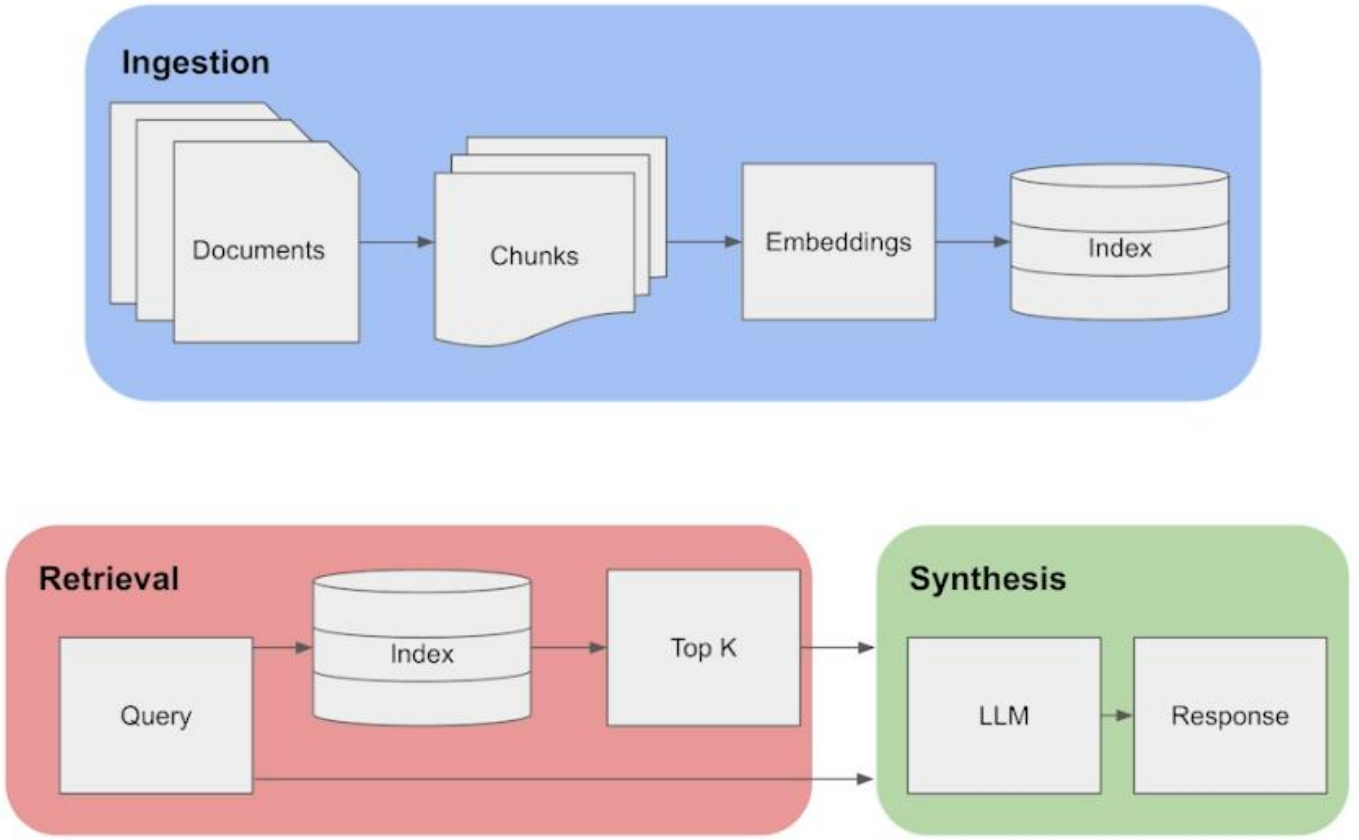
RAG	Fine tuning
直接更新檢索知識庫	需要重新訓練
需要少的數據處理	有限的資料集無法有顯著效能的提升
答案可以追溯到特定的資料來源	像黑盒子，無法知道模型為何以某種方式做出反應
無法自訂模型的行為或風格	允許根據特定語氣或術語調整模型的風格

Neural Retrieval

- Encode the query and documents into dense vector representations
  → compute cosine similarity
  → determine which documents are most relevant to a query
- Advantage:
  adept at dealing with long and complex queries
- Challenge:
  performance depends on the data they are trained on

Process of RAG

1. Vector Database Creation
2. User Input
3. Information Retrieval
4. Combining Data
5. Generating Text



Chunking

- Fixed-size chunking
- Context-aware chunking

1. Sentence Splitting
   - Naïve Splitting
   - NLTK
   - spaCy
2. Recursive Chunking