Document Query RAG Pipeline

Process Walkthrough

1A. User uploads documents into database for storage

1B. User selects documents to reference and submits query (AI-query enhancement optional)

2A. Scan document(s) for the **underlying meaning and concepts** behind the query

2B. Scan document(s) for **exact keyword matches and specific terms** found in the query

3. Extracted results are ranked for query relevancy into one combined list (Results on both lists rank higher)

4. Cross-Encoder assigns confidence scores to chunks on combined list for relevancy

- IF no confidence scores reach the relevance threshold:

    6B. No Confidence to Answer

- ELSE

    5. Extractor reviews high scoring results for key quotes to reduce "noise"

    6A. Program sends: User Query, Best Excerpts, and Prompt Instructions into LLM for response generation

1B. Submit Query

Optional: Query Transform

ON/OFF

User Query

User Query

2A. Contextual Document Search

meaning

2A. Dense Search (Bi-Encoder)

Dense Search Results

3. Combine Results

1A. Upload Documents

Document(s)

4. Score Results

Combined Results

5. Extract Key Quotes

High Scoring Results

Best Excerpts

6A. Response Generation

Document(s)

User Query

keywords

2B. Sparse Search (BM25)

2B. Keyword Document Search

Sparse Search Results

Low Scoring Results

6B. Low Confidence / No Response