Hands-on: RAG Pipeline

💻 LangChain RAG Implementation

```
from langchain.vectorstores import Qdrant
from langchain.embeddings import OpenAIEmbeddings
from langchain.llms import OpenAI
from langchain.chains import RetrievalQA

# 1. Setup Vector Store
vectorstore = Qdrant(
    embeddings=OpenAIEmbeddings(),
    collection_name="medical_kb"
)

# 2. Create Retrieval Chain
qa_chain = RetrievalQA.from_chain_type(
    llm=OpenAI(temperature=0),
    retriever=vectorstore.as_retriever(
        search_kwargs={"k": 5}
    ),
    return_source_documents=True
)

# 3. Query with Citations
result = qa_chain({
    "query": "Treatment for Type 2 Diabetes?"
})

print(result['answer'])
print(result['source_documents'])
```

📋 Implementation Steps

1. Load medical documents (PDFs, text)
2. Chunk documents (512 tokens with 50 overlap)
3. Generate embeddings (BioBERT recommended)
4. Index in vector database (Qdrant/Weaviate)
5. Configure retrieval (hybrid search, top-k)
6. Test with medical queries and evaluate