RAG with DeepSeek R1

Reasoning AI Chatbot

Project Layout

Phase1: Setup UI with Streamlit

- Setup Upload PDF functionality
- Chatbot Skeleton (Question & Answer)

Phase2: Setup Vector Database (Memory for LLM)

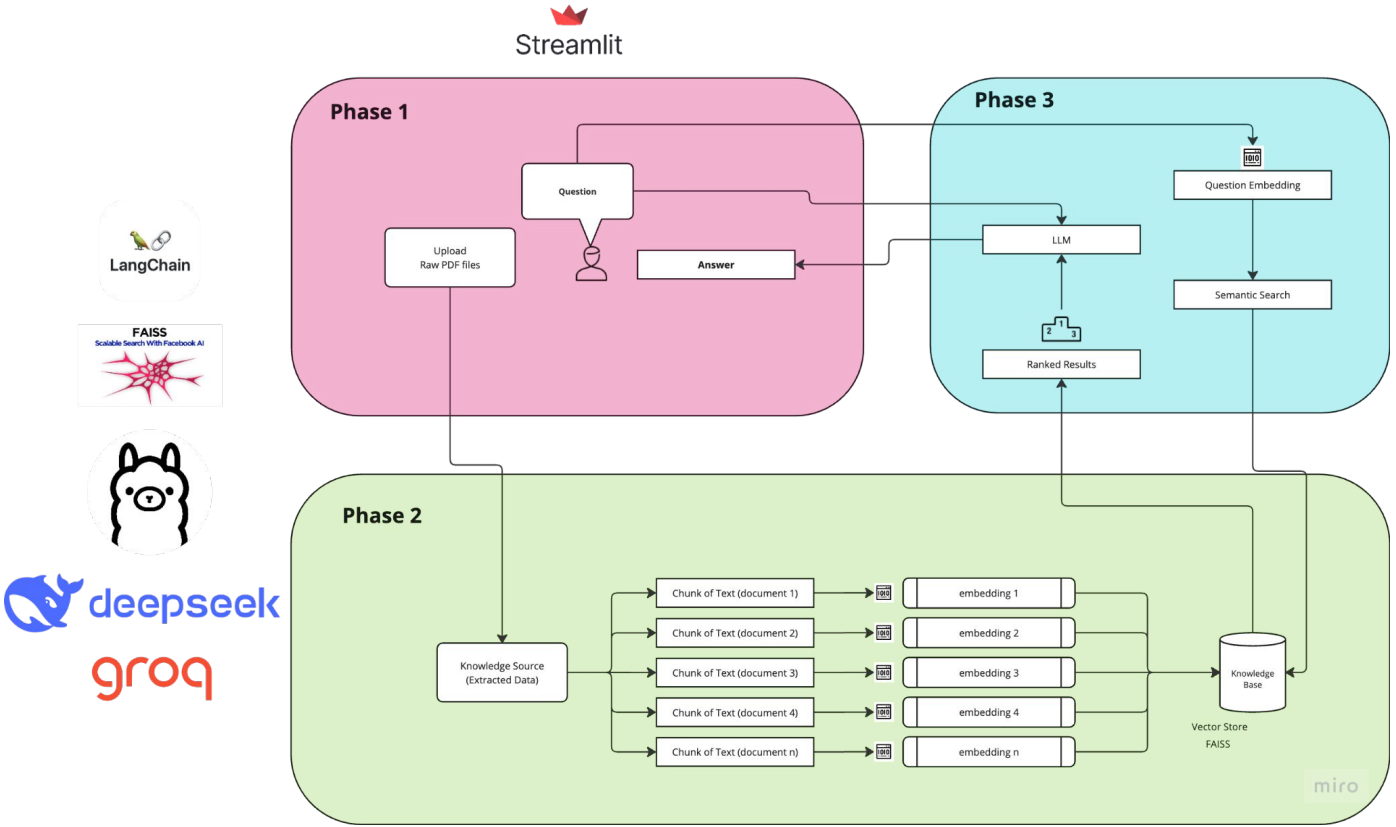
- Upload & Load raw PDF(s)
- Create Chunks
- Setup Embeddings Model **(Use DeepSeek R1 with Ollama)**
- Index Documents **Store embeddings in FAISS (vector store)

Phase3: RAG Pipeline (Connect Memory with LLM)

- Setup LLM **(Use DeepSeek R1 with Groq)**
- Retrieve Docs
- Answer Question

Phase4: Connect Everything Together

Technical Architecture



Tools and Technologies

- Deepseek R1 (Best Free CoT Reasoning Model)
  - With Ollama
  - With GROQ
- Ollama (Local LLM hosting/managing platform)
- Langchain (AI Framework for LLM applications)
- RAG, What is RAG and why do we need it?
- Streamlit (For Chatbot UI)
- FAISS (Vector Store)
- Pdfplumber and langchain loaders

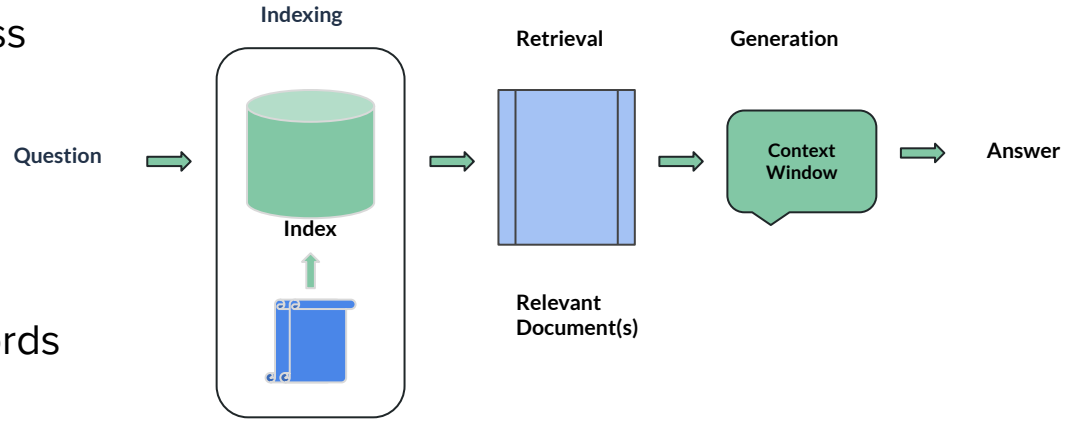
RAG–Retrieval Augmented Generation

Basics:

- Models are trained up to a limit
- No proprietary information access
- Answers from specific sources
- Reduce Hallucinations
- Vector Embeddings

Vector Embeddings:

- Numerical representations of words or phrases
- Capture meanings and relationships
- Helps machine learning models understand text effectively



RAG High Level Pipeline