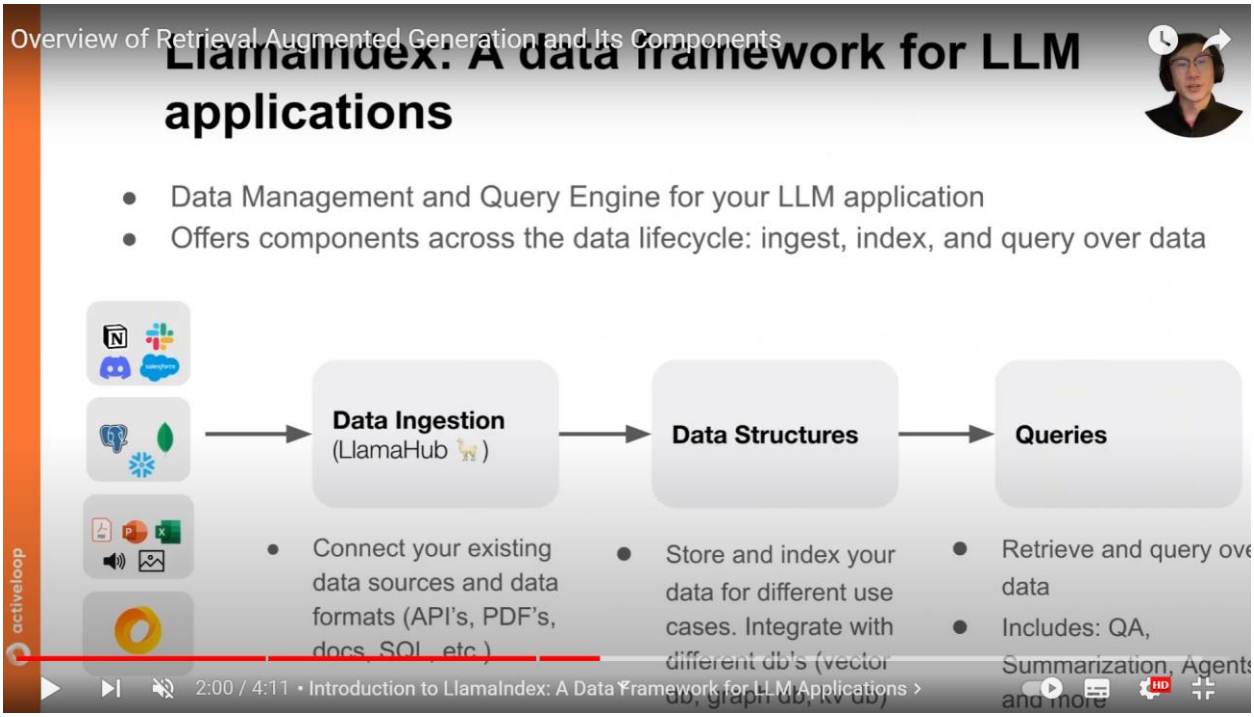
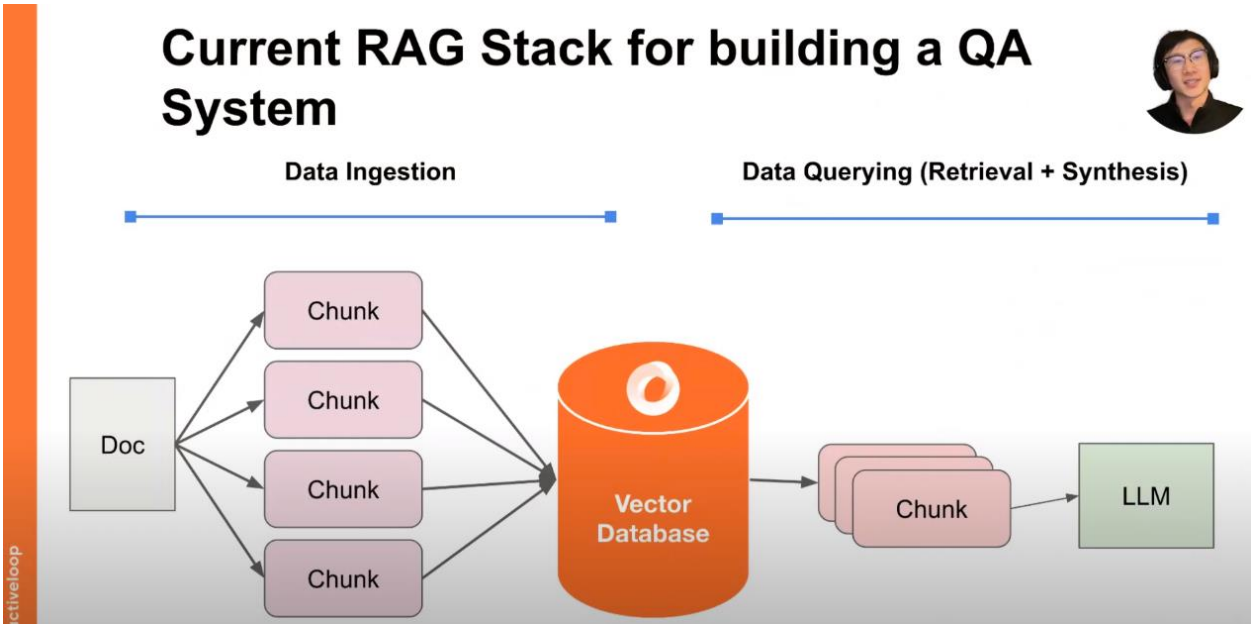
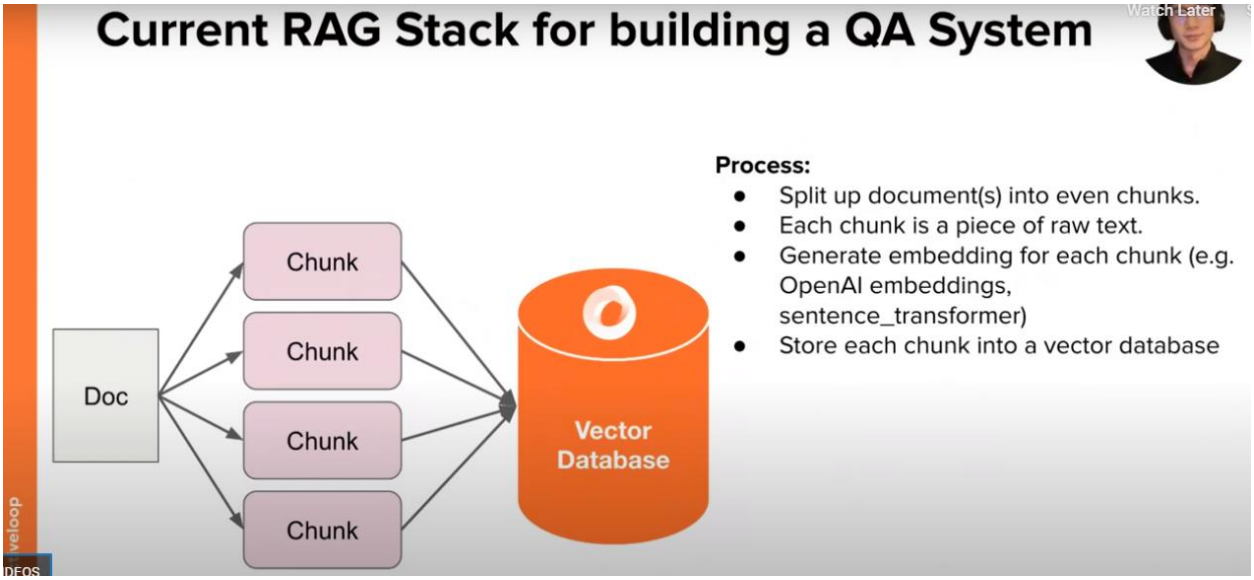
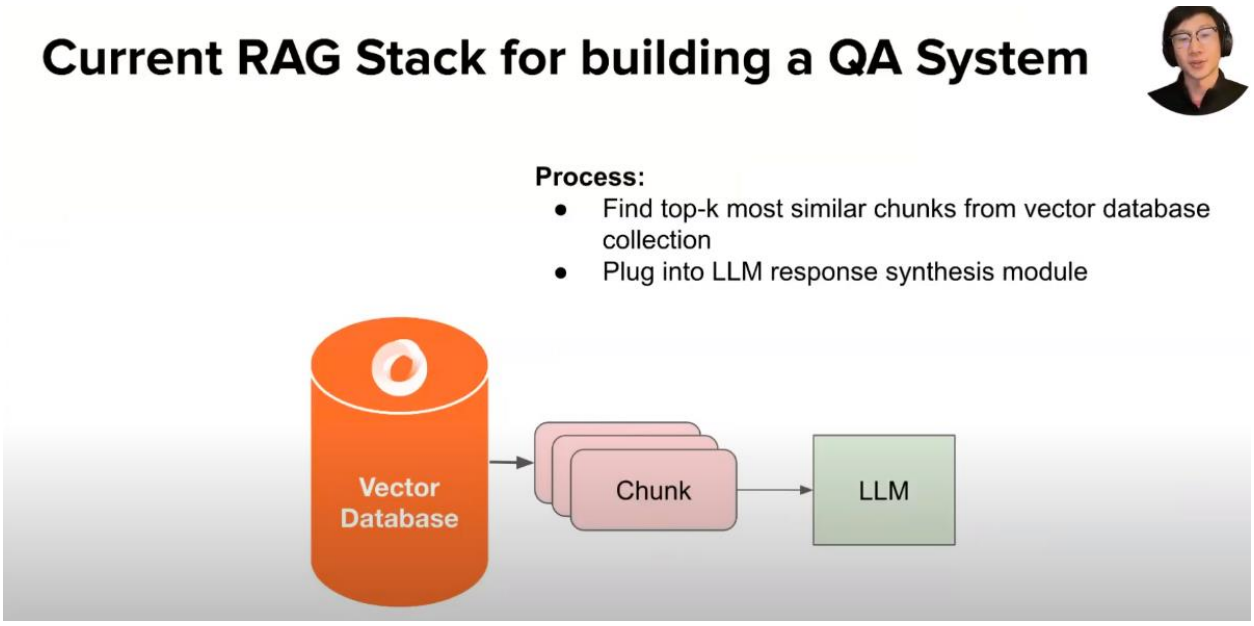


How RAG Works




Current RAG Stack for building a QA System
Watch Later
Process:
Split up document(s) into even chunks.
Each chunk is a piece of raw text.
Generate embedding for each chunk (e.g. OpenAI embeddings, sentence_transformer)
Store each chunk into a vector database
Doc
Chunk
Chunk
Chunk
Chunk
Vector Database

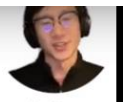

Current RAG Stack for building a QA System
Process:
Find top-k most similar chunks from vector database collection
Plug into LLM response synthesis module
Vector Database
Chunk
LLM

Challenges with "Naive" RAG

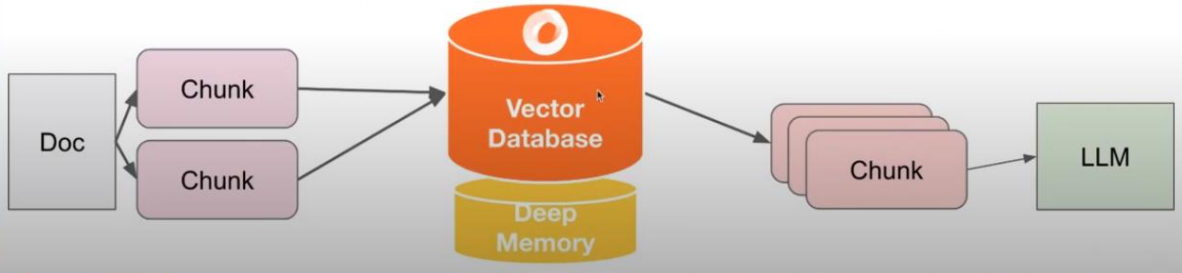
Challenges with Naive RAG

- **Bad Retrieval**
  - **Low Precision:** Not all chunks in retrieved set are relevant
    - Hallucination + Lost in the Middle Problems
  - **Low Recall:** Now all relevant chunks are retrieved.
    - Lacks enough context for LLM to synthesize an answer
  - **Outdated information:** The data is redundant or out of date.
- **Bad Response Generation**
  - **Hallucination:** Model makes up an answer that isn't in the context.
  - **Irrelevance:** Model makes up an answer that doesn't answer the question.
  - **Toxicity/Bias:** Model makes up an answer that's harmful/offensive.

What do we do?



- **Data:** Can we store additional information beyond raw text chunks? (**e.g. with multi-modal data store like Activeloop's Deep Lake)**
- **Embeddings:** Can we optimize our embedding representations?
- **Retrieval:** Can we do better than top-k embedding lookup? (**with systems like Activeloop's Deep Memory**)
- **Synthesis:** Can we use LLMs for more than generation?

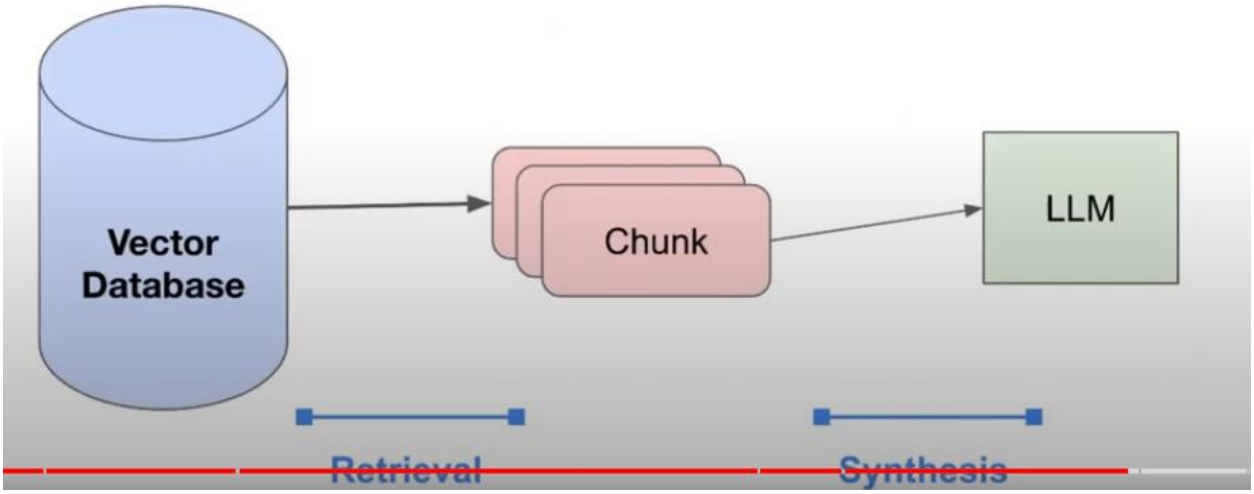


rove RAG Performance

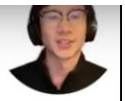
Evaluation

What do we do?

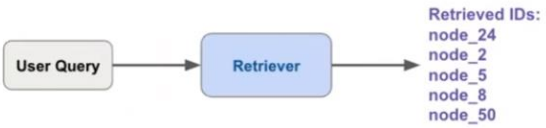
- How do we properly evaluate a RAG system?
  - Evaluate in isolation (retrieval, synthesis)
  - Evaluate e2e

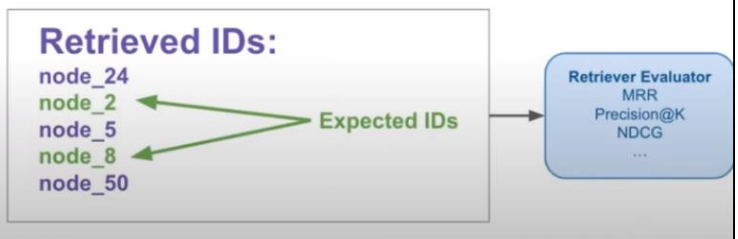


Evaluation in Isolation (Retrieval)

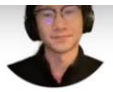


- Evaluate quality of retrieved chunks given user query
- Steps:
  - Create Deep Lake dataset
  - Run retriever over dataset
  - Measure **ranking metrics**

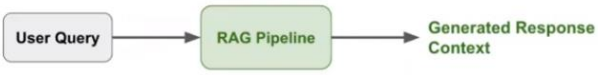




Evaluation in Isolation (E2E)



- Evaluation of final generated response given input



- **Steps**
  - Create a Deep Lake Dataset
  - Run through full RAG pipeline
  - Collect evaluation metrics