DAY 05 MODULE

End to end RAG Pipeline

The foundations are set. Now, we build.

By,

Shikha Tyagi

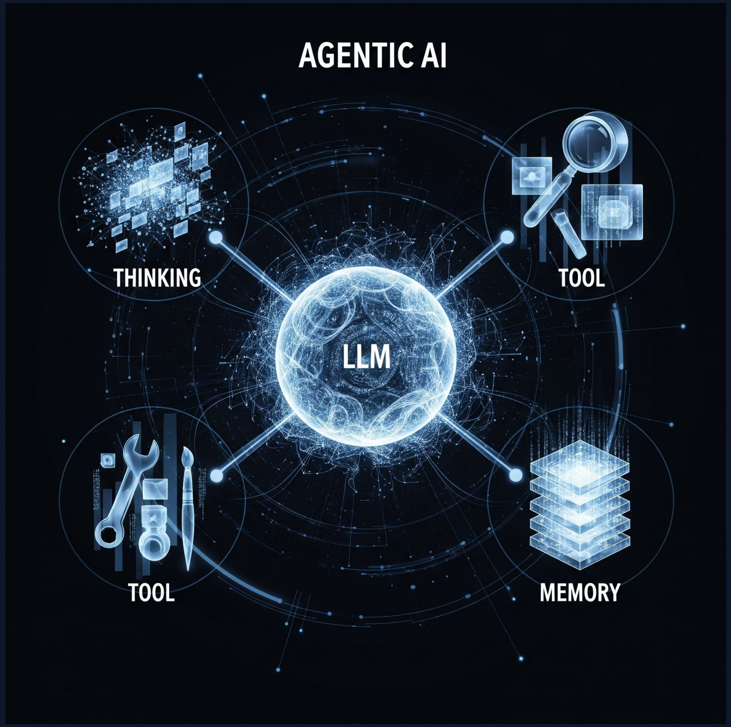
Founder - AI JAMIC (AI Research and Consulting)

Education: IIT Delhi (M.Tech.)

Guidelines

- Attendance is mandatory for all 5 sessions
- Hands on activity is mandatory
- 15 min break at 10:30PM
- QnA session at the end (10-15 min)
- Feel free to drop your questions in chat
- There will be quizzes in-between, drop your answers in chat

5 day roadmap



✓ 1

Shift

Agentic Thinking vs. Chatbots

✓ 2

Brain

LLM Types & Prompting

✓ 3

Hands

Function Calling & Tools

✓ 4

Memory

RAG & Vectors

🏁

Build

End to end pipeline and Projects

Today’s Agenda

01 **RAG Pipeline components**

02 **Hands on**

03 **Project details** (pick a track)

Quiz - 1

Which model can convert text into numbers

1. all-MiniLM-L6-v2
2. Gemini-flash

Quiz - 2

Which database is used to store vectors

1. Sqlite
2. Chromadb

Quiz - 3

Query and document should be embedded using

1. Same embedding model

2. Different embedding model

Quiz - 4

RAG pipeline is suitable for

1. Structure data
2. Unstructured data

Quick exercise

Open Deepseek and upload

https://csc-knu.github.io/sys-prog/books/Andrew%20S.%20Tanenbaum%20-%20Modern%20Operating%20Systems.pdf

Review: The Full RAG Pipeline

To build a "Knowledgeable" agent, we must master this flow:

1. Ingest

Load Raw Data

2. Chunk

Split into segments

3. Embed

Create Vectors

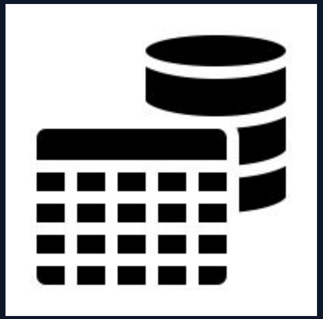
4. Retrieve

Semantic Search

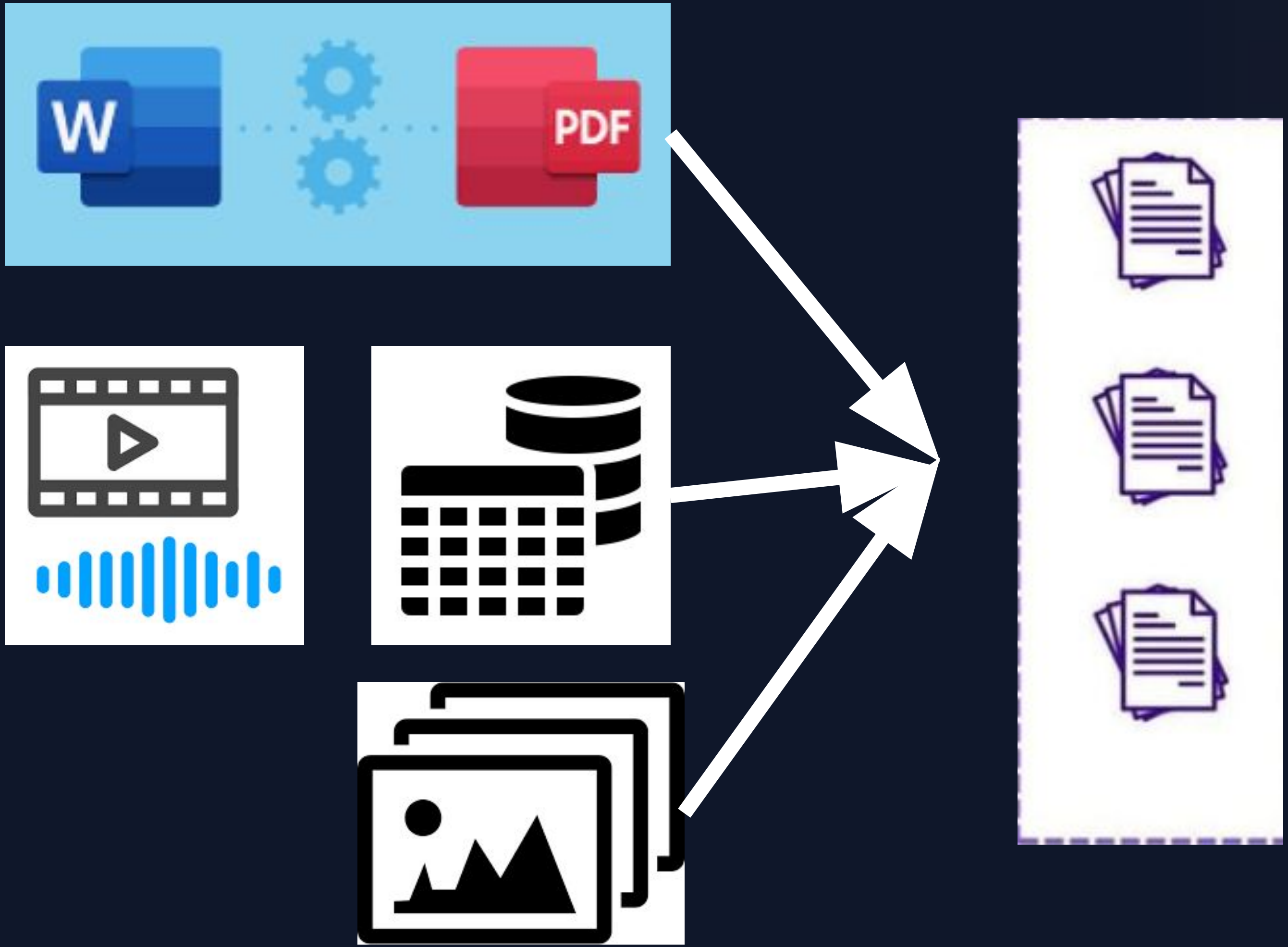
5. Generate

Answer with Context

Ingest Data



Text Extraction



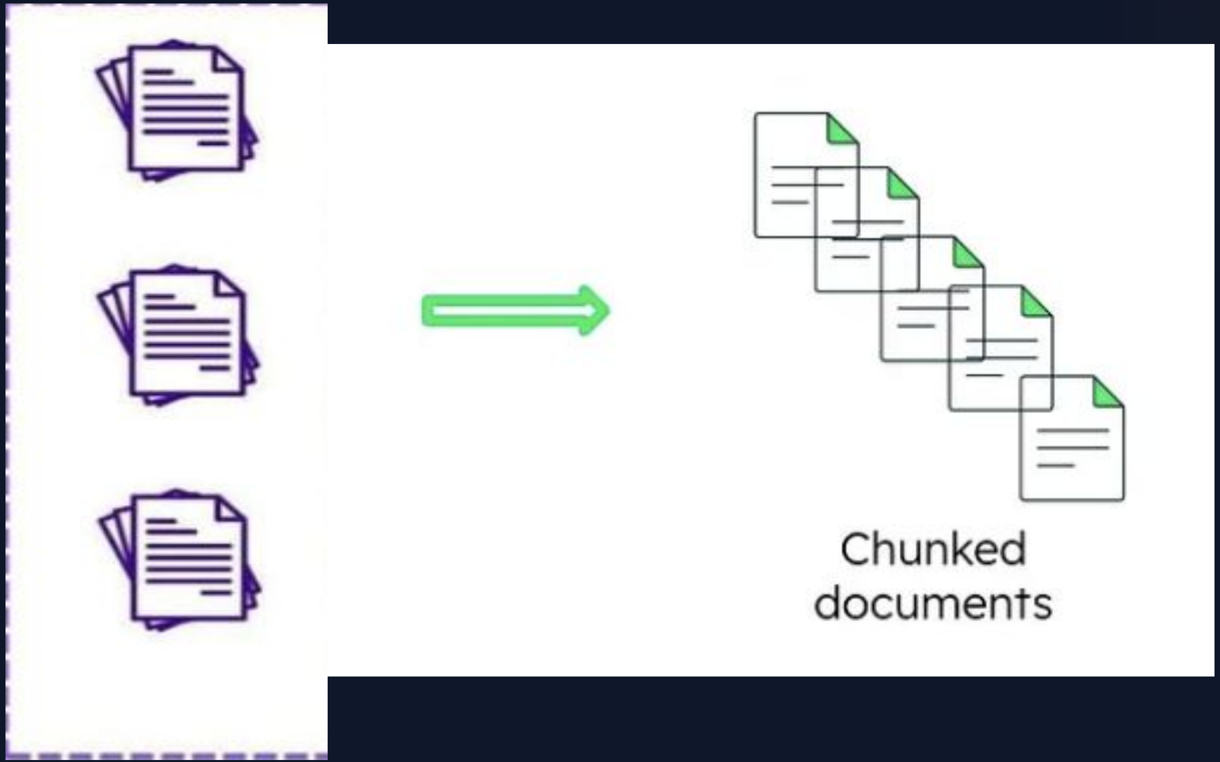
Quiz

Can all AI models handle image data?

Quiz

Which AI model can take image as input?

Chunking - Split a document into sub-documents



Chunking - Managing Large Amount of Information

- Chunking Strategies

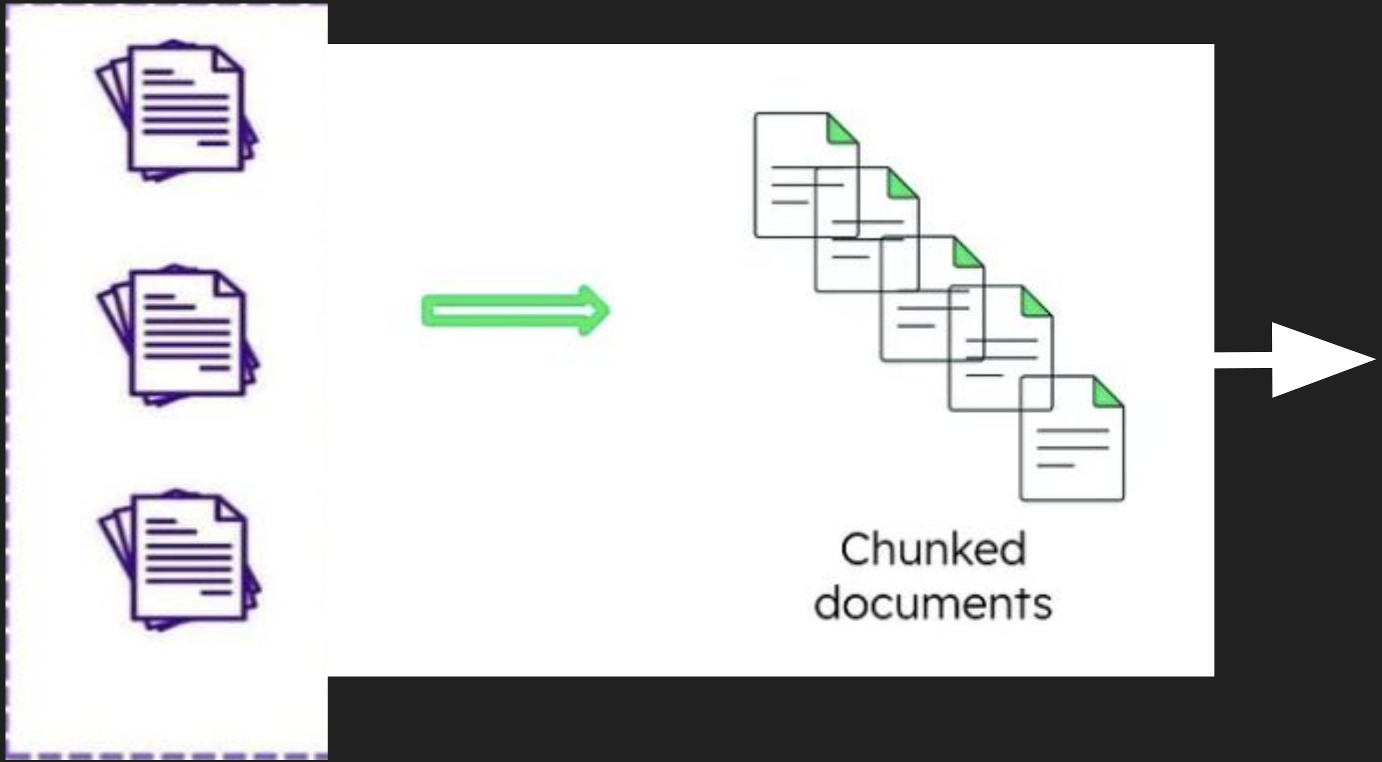
- Document segmentation approaches:
  - Fixed-size chunks (token or character count)
  - Semantic chunking (paragraphs, sections)
  - Recursive chunking with hierarchical representation

https://python.langchain.com/docs/concepts/text_splitters/

- Overlap Techniques

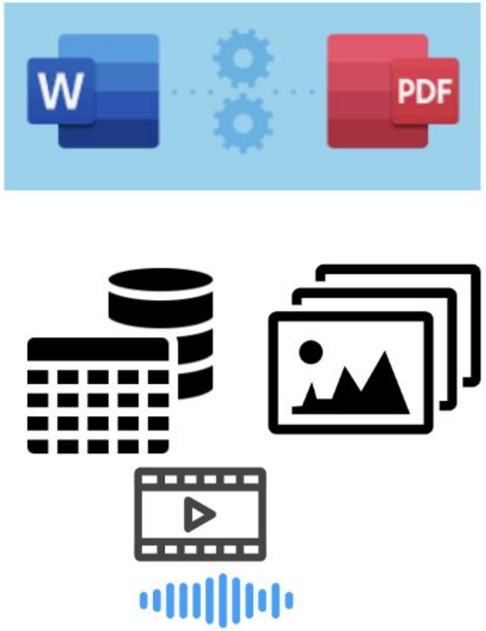
- Sliding window with overlap
- Handling cross-reference information
- Preserving context at chunk boundaries

Embedding

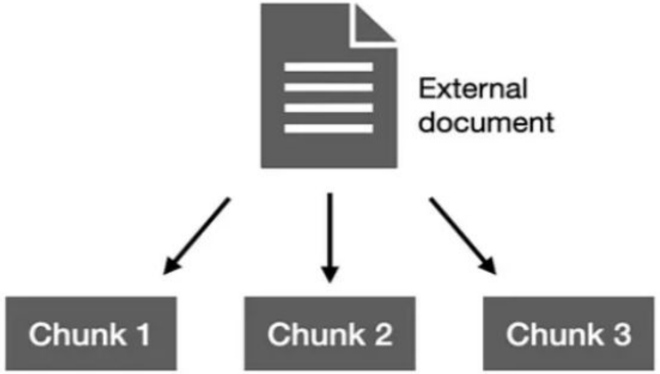


Each chunk is embedded using embedding model

Data Repository

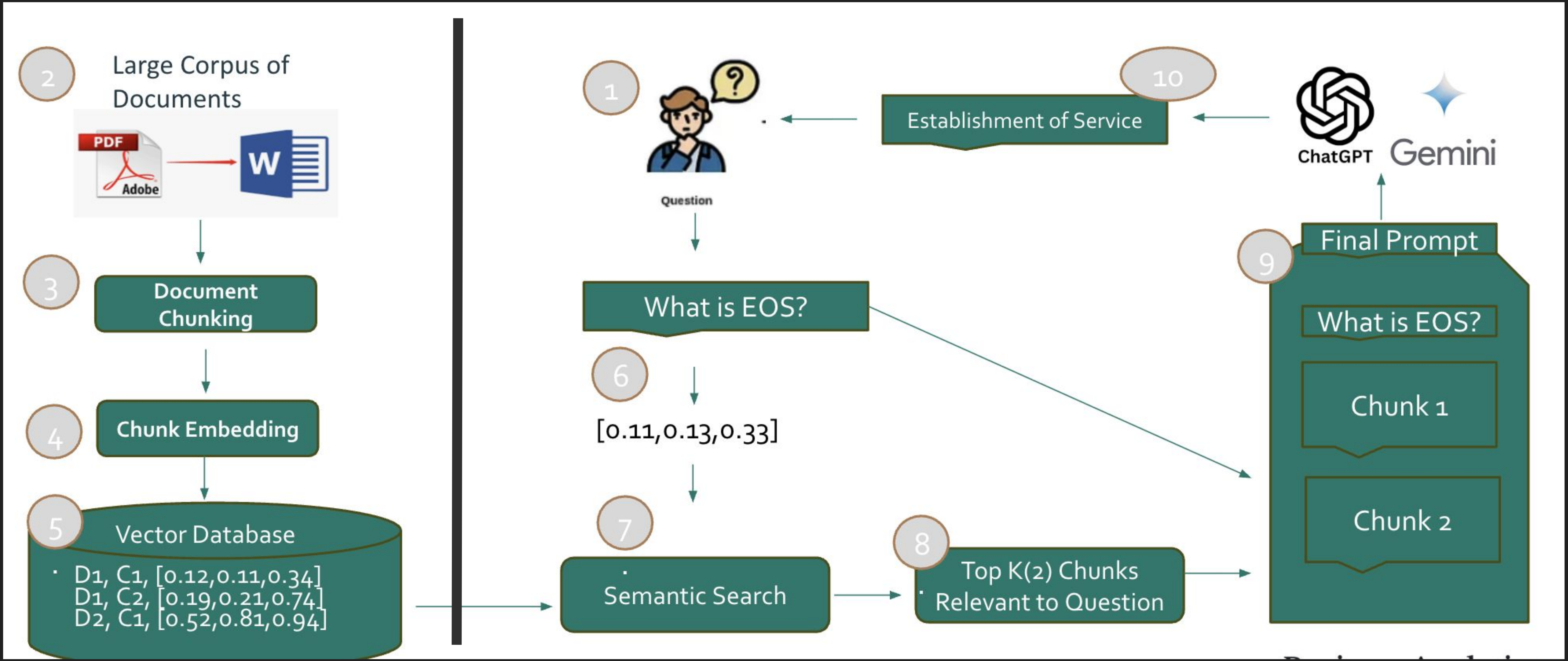


Chunker



Vector Database

```
-0.32643065,
-0.12308089,
-0.2873811 ,
-0.99628943,
-0.2503798 ,
 0.24311952,
 0.5662387 ,
 0.17282294,
-0.1109335 ,
 0.15209009,
 0 47017908,
-0.19270805,
```



2 Large Corpus of Documents

3 Document Chunking

4 Chunk Embedding

5 Vector Database

- D1, C1, [0.12,0.11,0.34]
  D1, C2, [0.19,0.21,0.74]
  D2, C1, [0.52,0.81,0.94]

1 Question

What is EOS?

6 [0.11,0.13,0.33]

7 Semantic Search

8 Top K(2) Chunks Relevant to Question

9 Final Prompt

What is EOS?

Chunk 1

Chunk 2

10 Establishment of Service

Quiz

How to decide top(k) chunks in RAG pipeline?

Quiz

Document ingestion in RAG pipeline should be an online process or offline process?

Quiz

Is RAG a right solution for large document summarization?

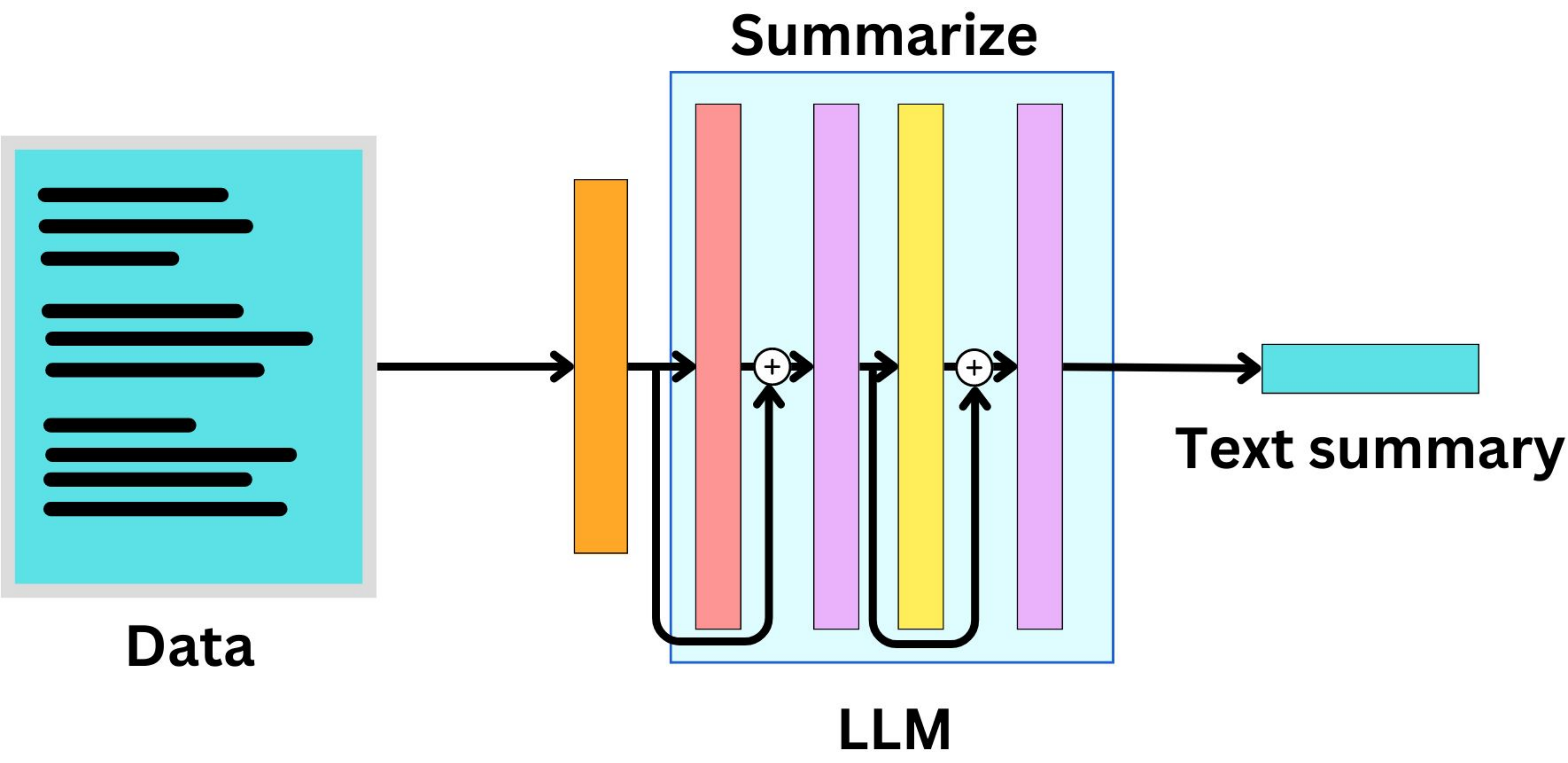
Query: Summarize this document -> embedding

Hands on

Summarization Strategies

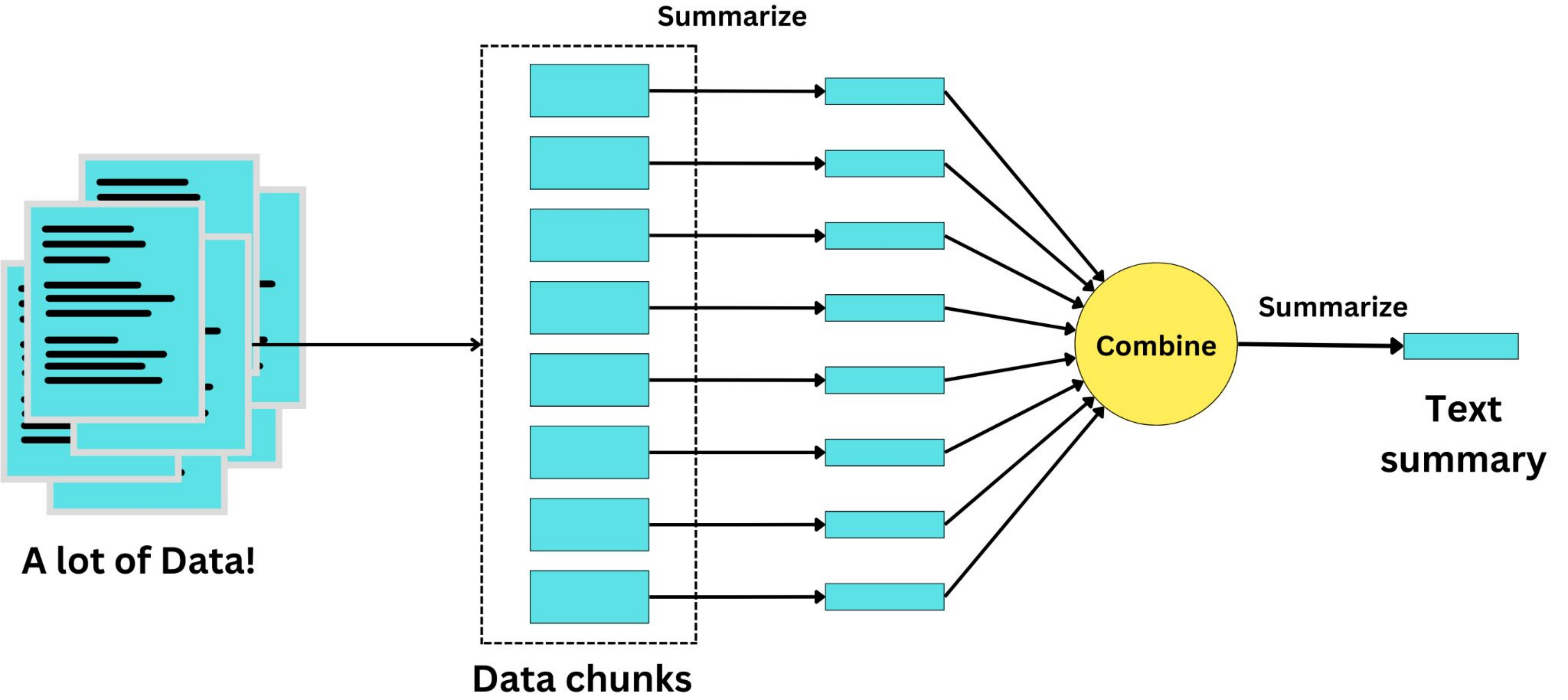
- Chaining
- Chunking

Chaining



Aiedge newsletter

Chunking



Aiedge newsletter

Hands on

Capstone

From learning to implementation

Track 1: E-Commerce Agent

The ”Personal Shopper”

Retail & Customer Support

Goal: Build an agent that helps users find products and track orders.

- **RAG:** Index a product catalog (PDF/CSV) to answer "What is your return policy?"

"User: What is your contact number?"
"Agent: +136571352"

Track 2: Academic Assistant

The ”Research Companion”

Education & EdTech

Goal: An intelligent study buddy that quizzes you on textbooks.

- **RAG:** Index a specific textbook chapter or research paper.

"User: What is photosynthesis."
"Agent: Sure! photosynthesis is ...."

Track 3: Legal Analyzer

The ”Risk Spotter” Legal & Compliance

Goal: An agent that reviews contracts and highlights risky clauses.

- **RAG:** Index standard NDAs or Employment Contracts.

"User: Check this NDA for non-compete"
"Agent: Warning: The 'Non-Compete' duration of 5 years is unusually long."

Q&A