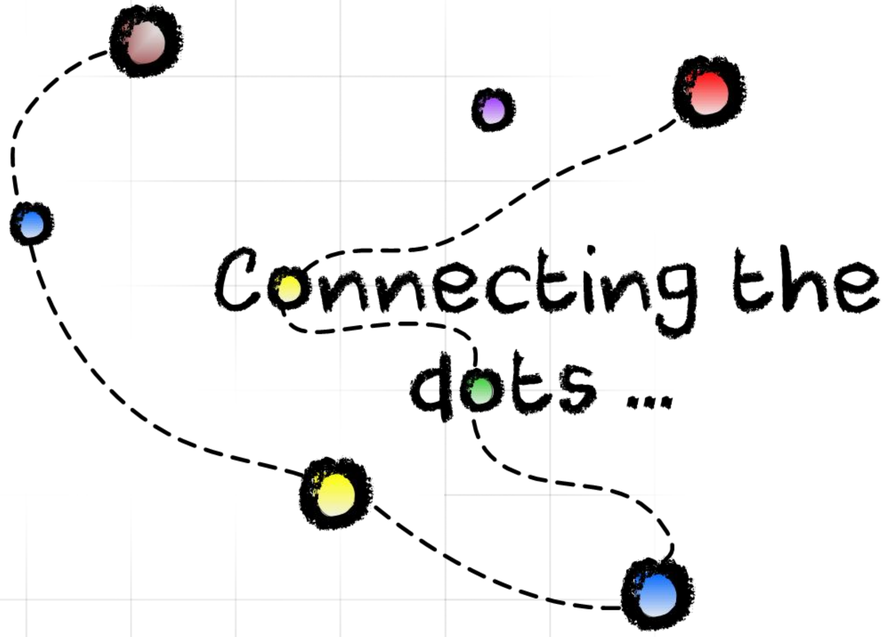
RAG



100x

Limitations of LLMs

- Knowledge cutoff
- No citations
- Incorrect data
- Limited context window

Emerging capabilities of LLMs: In-context learning

We can solve some of the limitations by giving the relevant context

Elements of a Prompt

A prompt is composed with the following components:

- Instruction
- Context
- Input Data
- Output Indicator

```
Classify the text into neutral,
negative or positive

Text: I think the food was okay.

Sentiment:
```

How to give LLMs relevant Context?

100x

- **Retrieve** most relevant data
- **Augment** user's prompt with context
- **Generate** response

Retrieve most relevant data => getWeather(City) -> 30 C

Augment user's prompt with context => Instruction + Context + Input

Augmented Prompt

- Instruction = You are a helpful weather assistant who can provide the weather details using the given context.
- Context = Data fetched from API ex. 30 C
- Input = User prompt ex. what’s the weather in bangalore?

Generate response

It is cloudy with temperature of 30 C in bangalore

Connected dots of RAG

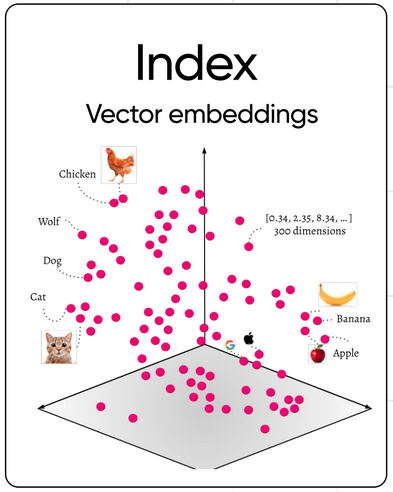
Your Data

Database
Structured

Documents
Unstructured

API
Programmatic

Load



Query 💬

User

response

Prompt + query
+ relevant data

LLM 🤖
Model like GPT, llama

100x

Stages of RAG

1. Load
2. Index/embed
3. Query
4. Generate
5. Evaluate