LLM Apps

Embeddings

Remember: RAG overcomes the context window

- Preparations:
    - Divide your data into small segments.
    - Convert the small segments into numbers (“embeddings”).
    - Load the embeddings into a vector database.

- Now when you ask (“query”) the LLM:
    - The LLM goes to the vector database and searches (“indexing”) for data that only (“semantic similarity”) answers your question.

Embeddings

- Computers work with numbers.

- That's why they convert text into numbers.

- They do the same with images, audio, video, etc.

- Embeddings are vectors of numbers.
  - Example: "hola" is converted into an embedding like (1,4,6).