LLM Apps

Vector Databases

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.

Remember: 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).

Vector Databases

- Vector databases are specialized in working with hundreds of millions of embeddings, they are much faster than conventional databases.

- Optimized for:
  - Storing.
  - Indexing.
  - Retrieving.

Semantic Similarity

- Databases group embeddings by their semantic similarity. For example, the embeddings of "perro" (dog) and "gato" (cat), which are semantically similar as both are animals, will be grouped together in the vector database.