

LLM Apps

Technical components of RAG

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: Importance of the RAG technique

- RAG is essential for creating LLM applications.
- That's why we will focus on mastering this technique.

Technical components of RAG

- We will see them in detail throughout the program:
 - Embeddings
 - Vector Databases
 - Semantic Similarity
 - Query
 - Indexing
 - Orchestration