

Aplicaciones LLM

LlamaIndex a fondo

LlamaIndex goal

- Unleash the power of LLMs over YOUR private data.
 - Load your data
 - Prompt over it

LlamaIndex options

- Start from scratch
- Use a starter kit
 - create-llama
 - Llama Packs

Documentation: starter tutorial

- Load data
- Create vector database
- Prompt over your data

Documentation: high-level concepts

- RAG Stages:
 - Loading
 - Chunks are called Nodes
 - Indexing
 - Mostly vector DB, also summary and knowledge graph
 - Storing
 - Store to avoid re-indexing
 - Querying
 - Steps: retrieval, postprocessing and response synthesis.
 - Observability
 - Tracing and debugging. 3rd party tools (W&B, etc)
 - Evaluating
 - Response evaluation, retrieval evaluation, cost prediction.

Optimization: advanced techniques

- This is the most interesting part of LlamaIndex
- General techniques
 - Decoupling retrieval and synthesis chunks
 - Structured retrieval
 - Dynamic retrieval of chunks
 - Context embedding optimization
- Specific techniques
 - Long list