



Meta AI chief Yann LeCun

“ Don’t work on LLMs”

<https://x.com/llopatin/status/1800997014924107966>

LangChain & RAG

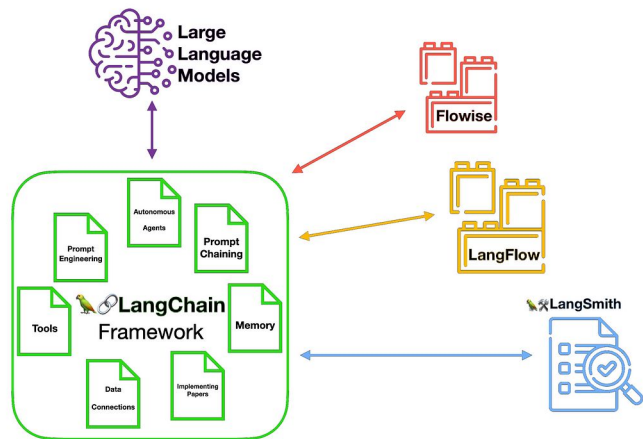
Reyero Joaquin

Objetivos

1. Entender qué es LangChain.
2. Introducir el concepto de RAG.
3. Aprender cómo integrarlos para construir aplicaciones confiables y escalables.

¿Qué es LangChain?

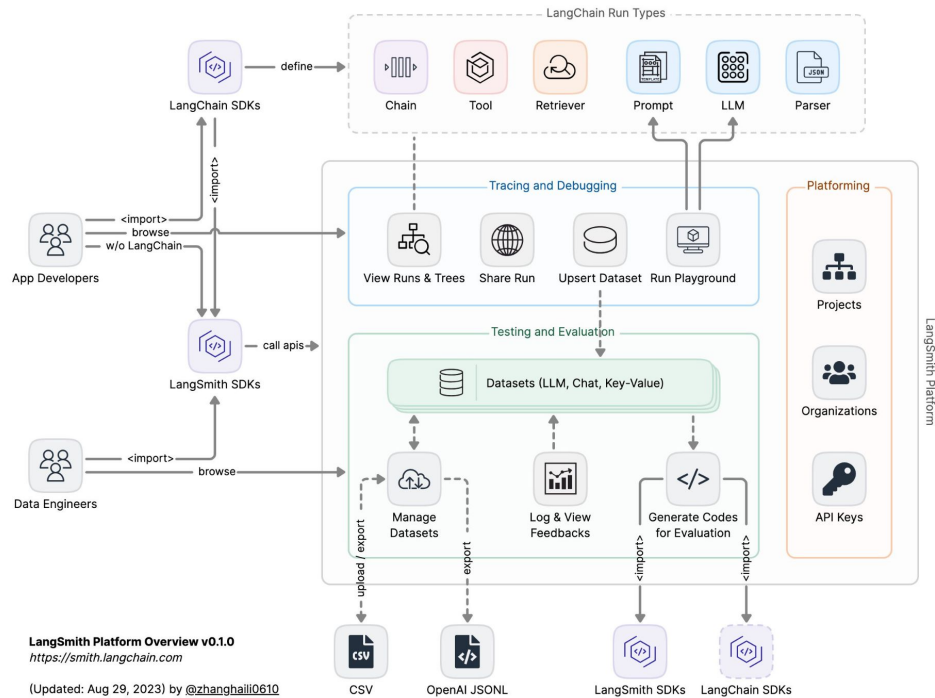
LangChain Ecosystem



www.cobusgreyling.com

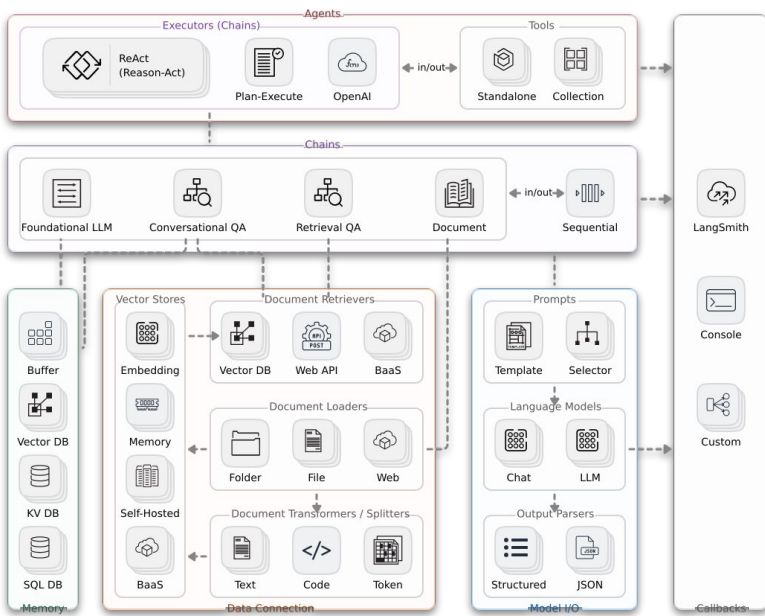
LangChain es un framework open source para construir aplicaciones sensibles al contexto.

- Permite la integración sencilla de diferentes modelos y fuentes de datos.
- Soporta workflows de trabajo complejos y personalizables.



LangSmith

Características de LangChain



LangChain (JS) Modules Overview v0.2.0 for npm:langchain@0.0.114 (Last Updated: July 22, 2023) by @zhanghaoli0610

Componentes

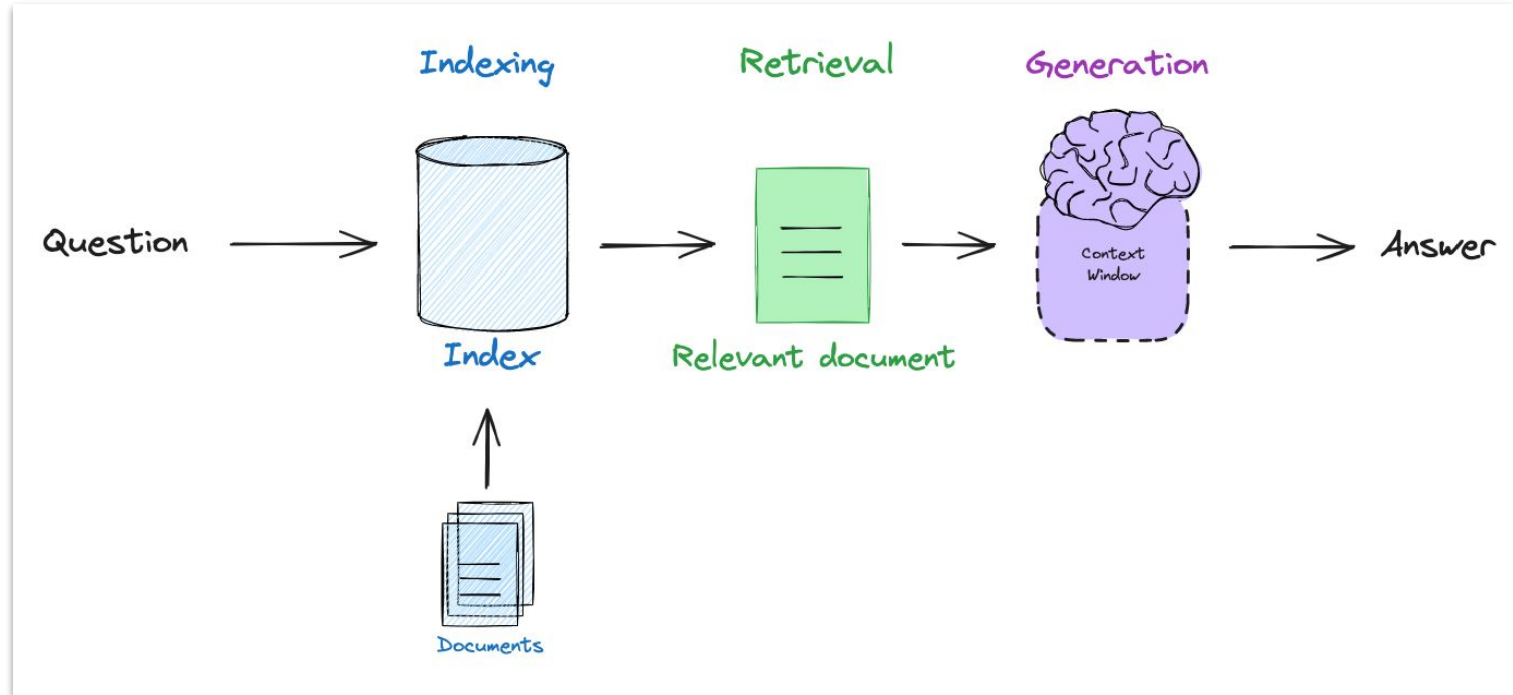
- **Model I/O**
 - Models(LLM/Chat/Embeddings)
 - Prompts(Templates, ExampleSelectors5, OutputParsers)
- **Retrieval**
 - Document Loaders
 - Text Splitters
 - Retrievers
 - Vector Stores
 - Indexers
- **Memory**

Use-Cases

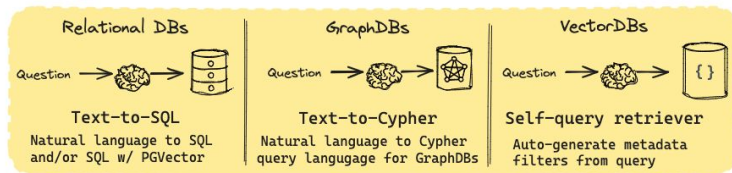
- Chains
- Agents
 - Tools
 - AgentExecutor

¿Qué es RAG?

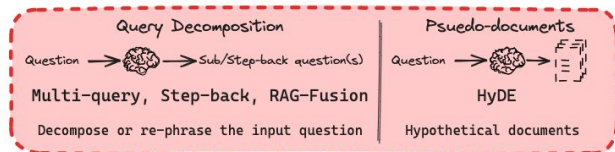
retrieval augmented generation



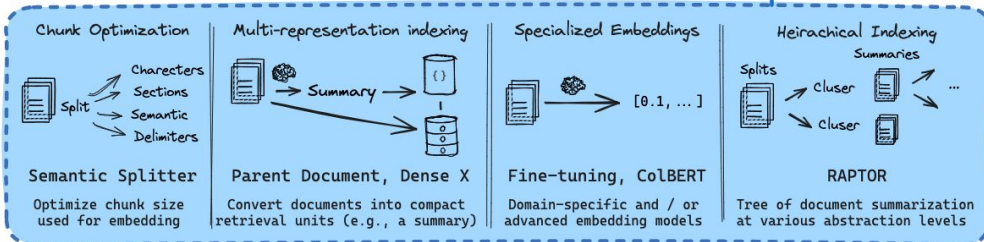
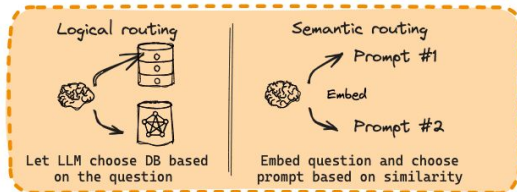
Query Construction



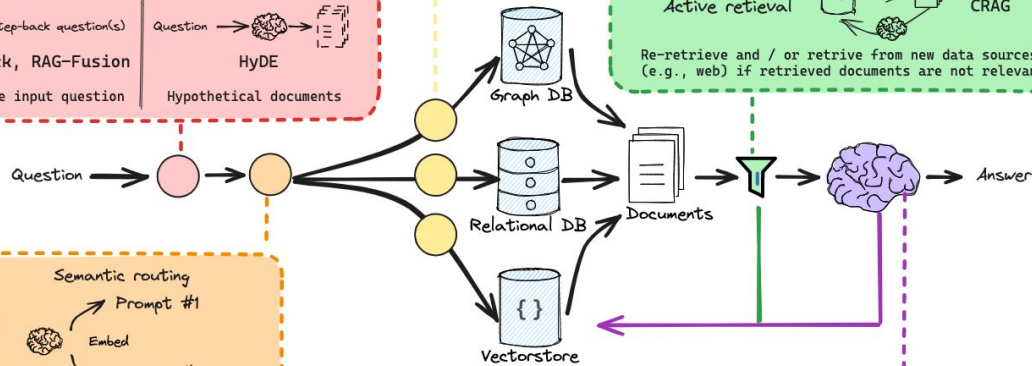
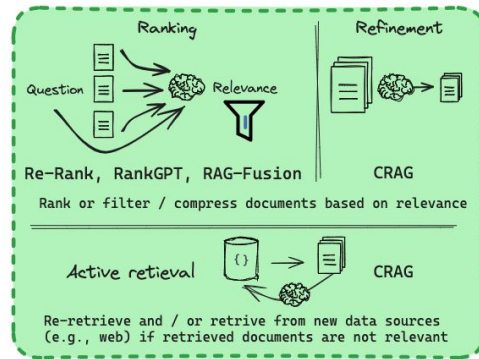
Query Translation



Routing



Retrieval



Indexing

Generation



Let's build

<https://github.com/joaquinreyero/simple-docuquery-ai>



LangChain



New Thread

⌘ K

🔍 Home

🔍 Discover

📖 Library

🔑 Sign In

Sign Up

Try Pro

Upgrade for image upload,
smarter AI, and more
Copilot.

➔ Learn More

📄 Download

✕

💬

how do large language models work

Sources

What are Large
Language Models? - LL...

amazon • 1

What are large language
models, and how do the...

stackexchange • 2

What is a Large
Language Model? | A...

elastic • 3

View 2 more



PRO Generate Image



Answer

Large language models (LLMs) are deep learning models that are pre-trained on vast amounts of data, typically using the transformer architecture, which consists of an encoder and a decoder ¹. These models are incredibly flexible and can perform a variety of natural language processing tasks, such as answering questions, summarizing documents, translating languages, and completing sentences ¹. Here's how LLMs work:

1. **Input:** LLMs receive an input, which can be a sequence of text or other content ³.
2. **Encoding:** The input is encoded by the encoder, which extracts meanings from the sequence of text and understands the relationships between the words ¹.
3. **Decoding:** The decoder then decodes the encoded input to produce an output prediction ³.

+ Ask follow-up...



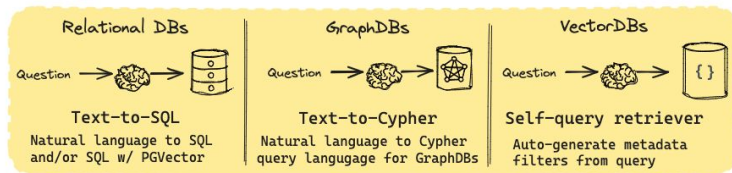
Copilot



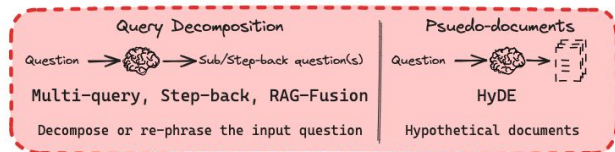
5. **Flexibility:** LLMs can be used for various purposes, such as information retrieval,

?

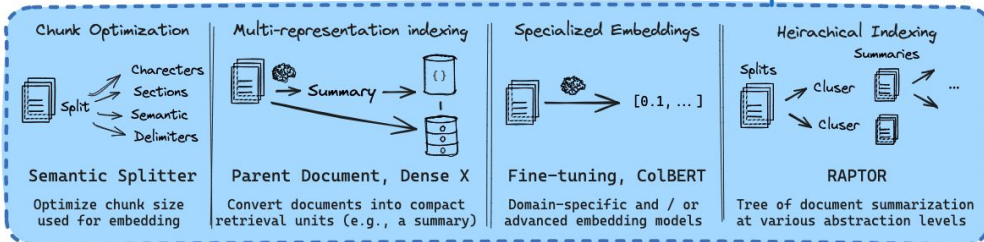
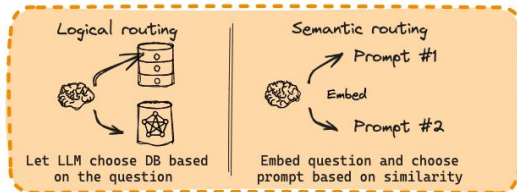
Query Construction



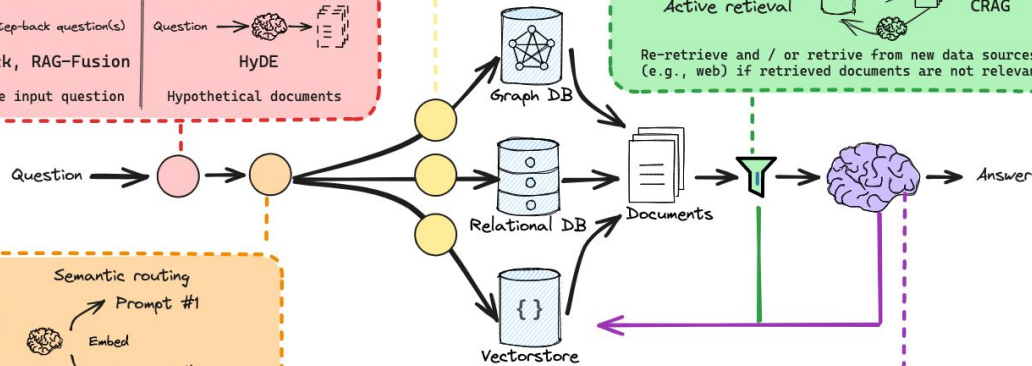
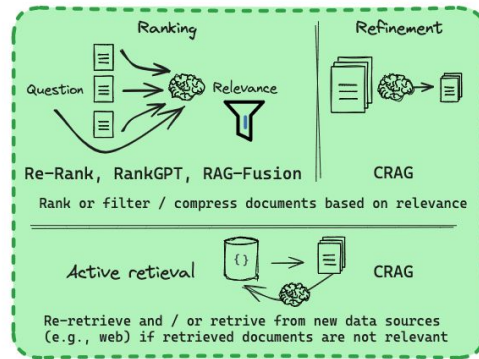
Query Translation



Routing

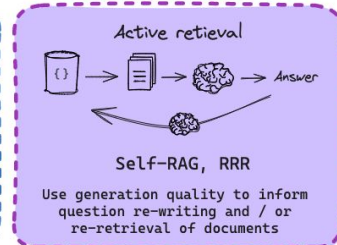


Retrieval

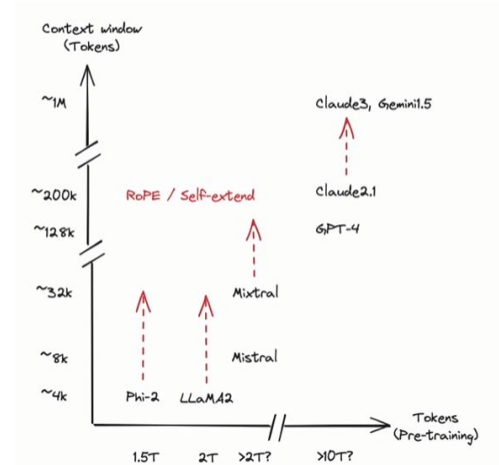


Indexing

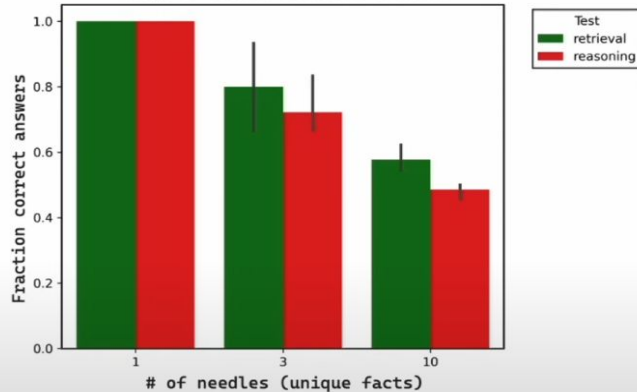
Generation



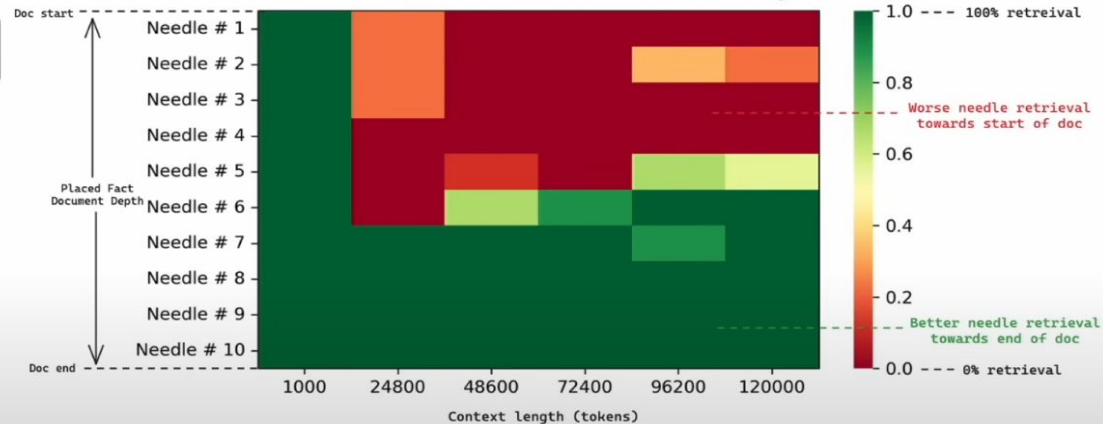
Es RAG el futuro?



Asking GPT-4 to retrieve or retrieve & reason
1, 3, or 10 needles (facts) in a single turn
120,000 token context window



Asking GPT-4 to retrieve 10 unique facts in 1 turn
Assess which needles are retrieved as context grows



Preguntas
