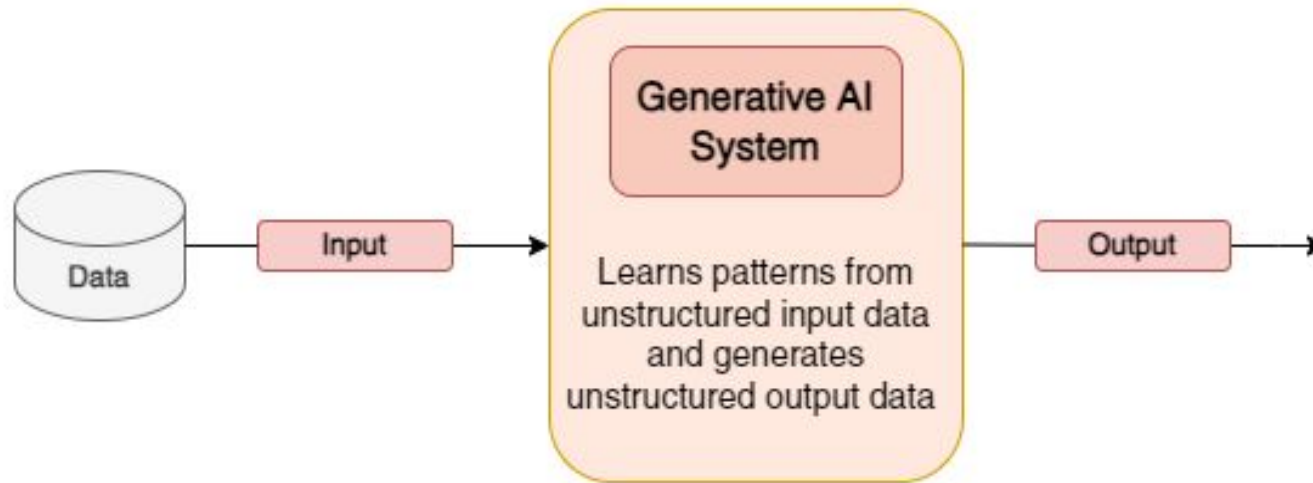




Generative AI Essentials

Dipanjan (DJ) Sarkar

What is a Generative AI System



**Is Not Generative AI
when Output is a:**

Number

Class Label

Probability Value

**Is Generative AI when
Output is:**

Text

Image

Audio

Video

Why Generative Language Models?

- The promise: one single model to solve many NLP tasks

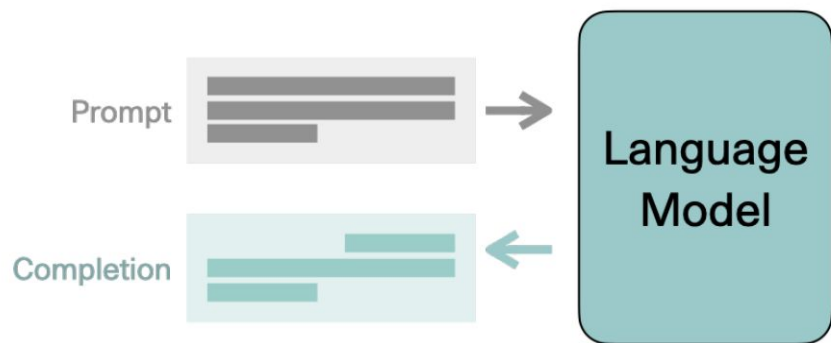


Image credit: Jay Alammar

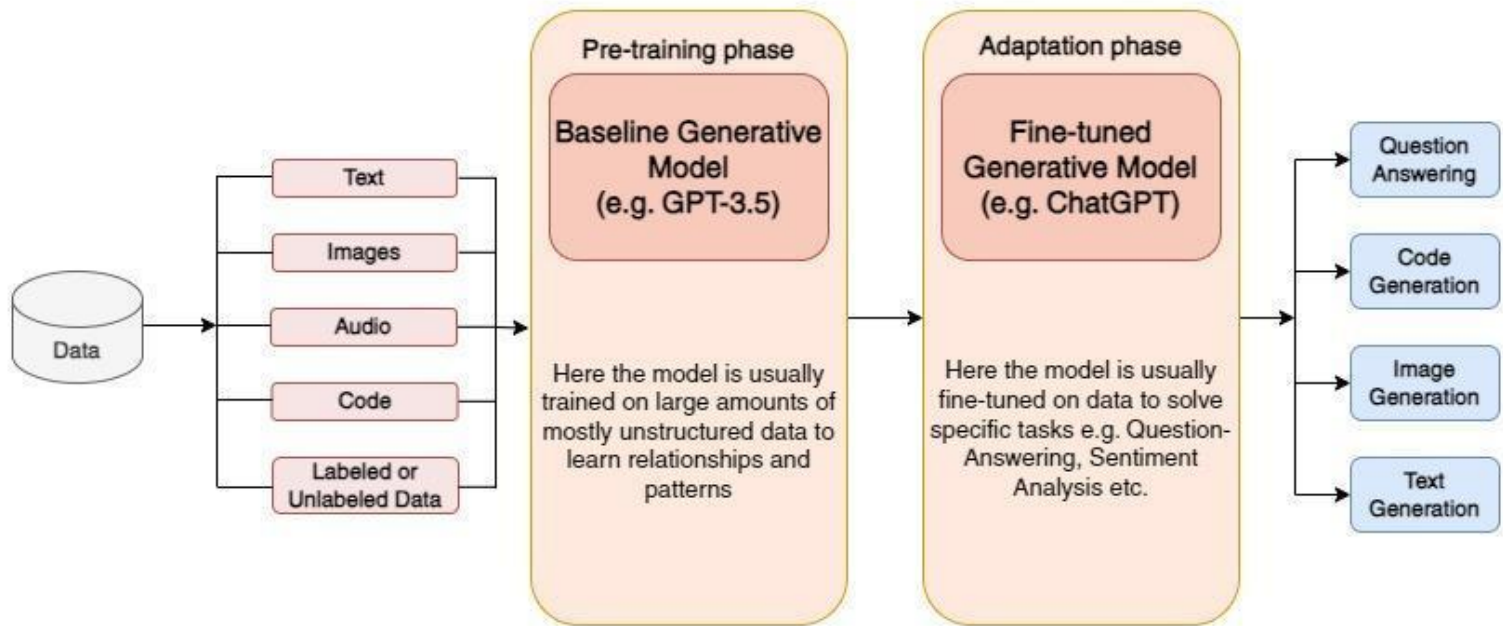
- 80% of data in enterprises is usually unstructured text data
- Various real-world applications in the industry are based on text data including translation, chatbots, search, classification and more.
- Generative Language Models can be easily trained to learn patterns, concepts and relationships from large amounts of unlabeled unstructured text data

Models? Language Models

- Language Modeling is the task of predicting what word comes next



Generalized Learning Systems - Generative AI



Gen AI principle is to adapt a single model to various tasks

Foundation vs. Fine-tuned Generative Models

Foundation Gen AI Model

Predicts next word, based on text training data

Once upon a time, there was a unicorn that lived in a magical forest with all her unicorn friends

What is the capital of France?
What is France's largest city?
What is France's population?
What is the currency of France?

Fine-tuned Gen AI Model

Tries to follow instructions

Fine-tune on instructions and good attempts at following those instructions.

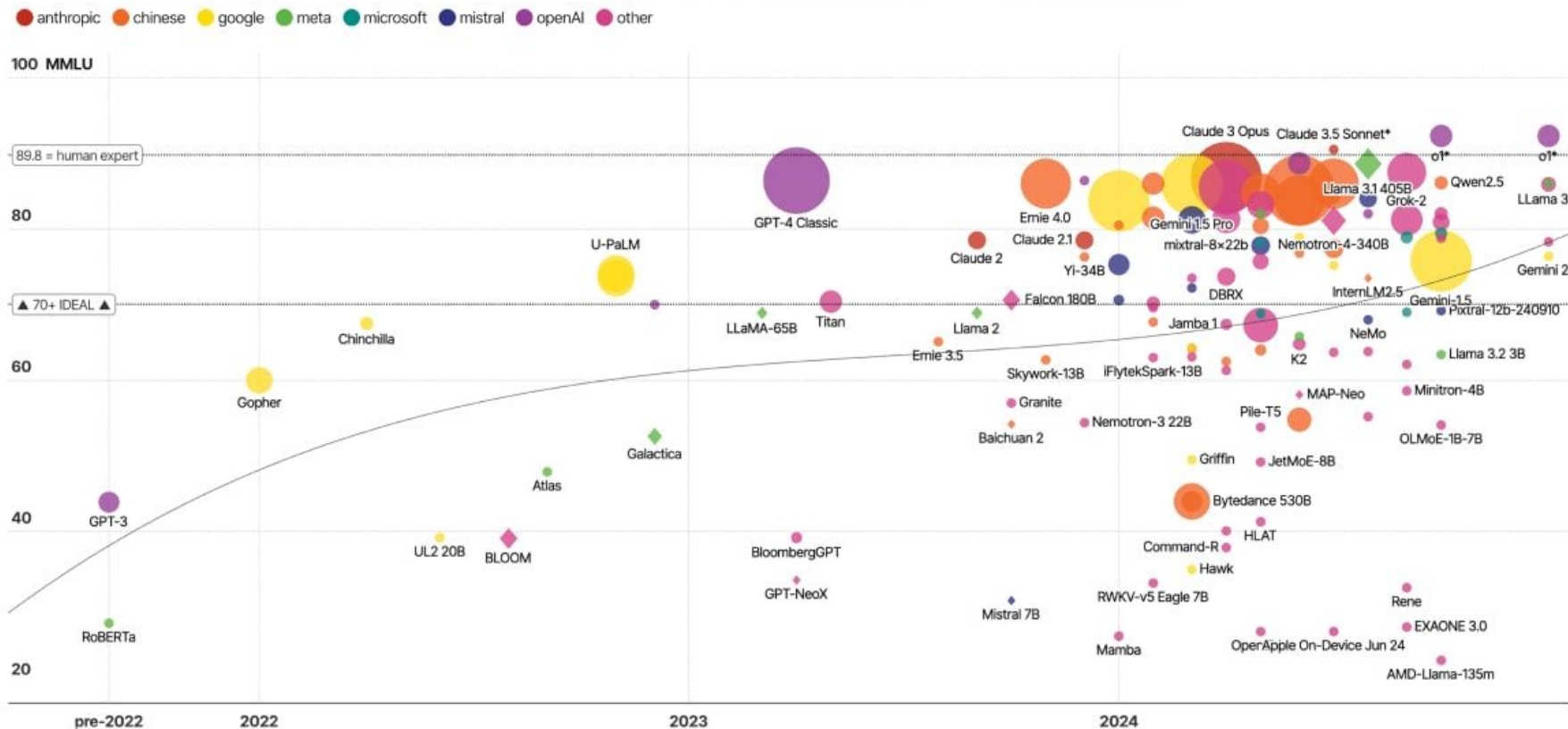
RLHF: Reinforcement Learning with Human Feedback

Helpful, Honest, Harmless

What is the capital of France?
The capital of France is Paris.

Image Credits: Deeplearning.ai

Latest LLMs and their performance



How to use LLMs like ChatGPT?

- The inputs to LLMs are referred to as "prompts"
- Creating a prompt is essentially how you "program" a Large Language Model.
- This is usually done by providing instructions or some examples of how to successfully complete a task
- Systematic ways of creating prompts are known as prompt engineering patterns

Input

Andrew is free from 11 am to 3 pm, Joanne is free from noon to 2 pm and then 3:30 pm to 5 pm. Hannah is available at noon for half an hour, and then 4 pm to 6 pm. What are some options for start times for a 30 minute meeting for Andrew, Hannah, and Joanne?

Output

Andrew: 11 am - 3 pm

Joanne: 12 pm - 2 pm, 3:30 pm - 5 pm

Hannah: 12 pm - 12:30 pm, 4 pm - 6 pm

Common availability for a 30-minute meeting: 12 pm - 12:30 pm

What is LangChain

- An open framework & ecosystem to develop applications powered by LLMs using a suite of tools

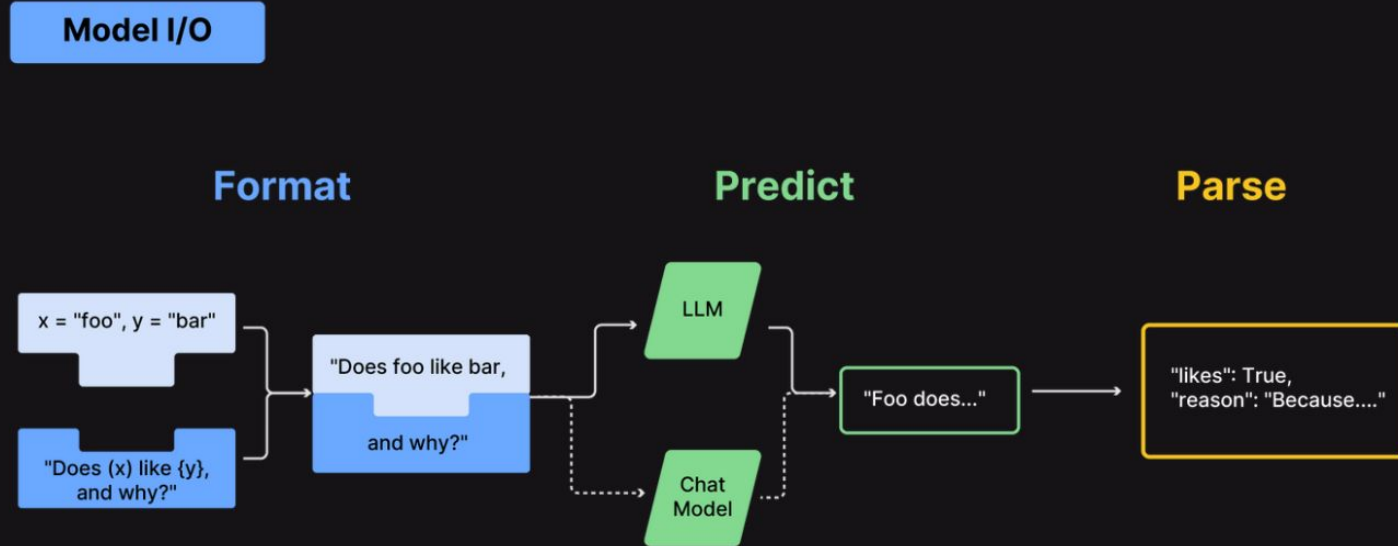


What is LangChain

- Its core framework consists of libraries to connect with LLMs using prompts and format outputs using output parsers.
- Multiple steps can be chained and complex conversational applications, RAG systems, and agents can be built.

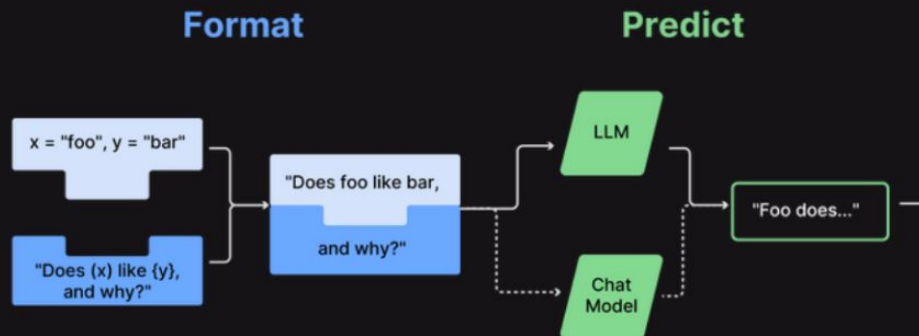


LLM Input/Output with LangChain



LLMs and Chat Models

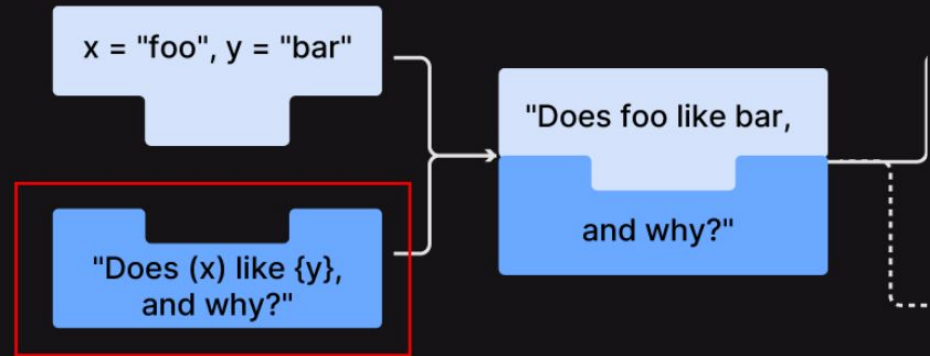
- LangChain provides two different ways of interfacing with LLMs
 - Regular LLMs for general prompting
 - Chat Models for conversational prompting
- We can interface with these LLMs using Prompt Templates in LangChain



What is a Prompt Template?

- Prompt templates are predefined recipes or templates for generating prompts for LLMs
- A template may include instructions, few-shot examples, and specific context and questions appropriate for a given task
- Usually has placeholder variables where you can substitute your data at runtime

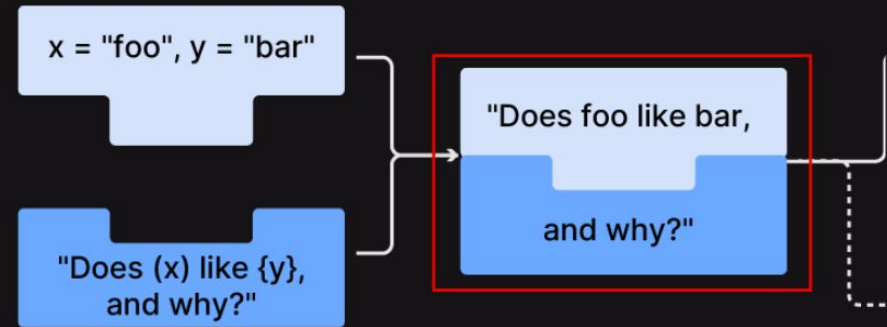
Format



What is a Prompt?

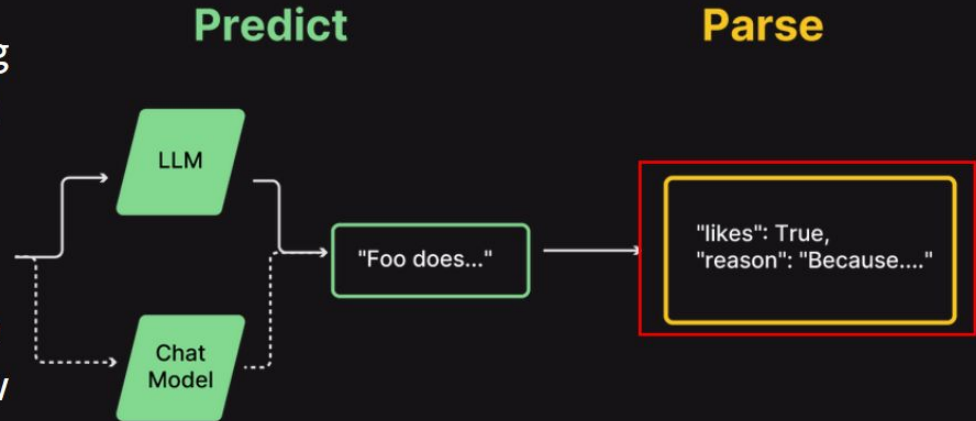
- A prompt is a set of instructions provided by a user to guide the LLM to respond in a specific way
- Prompt = Prompt Template + Input Data
- We usually pass our own data into a prompt template to create a prompt before sending it to an LLM

Format

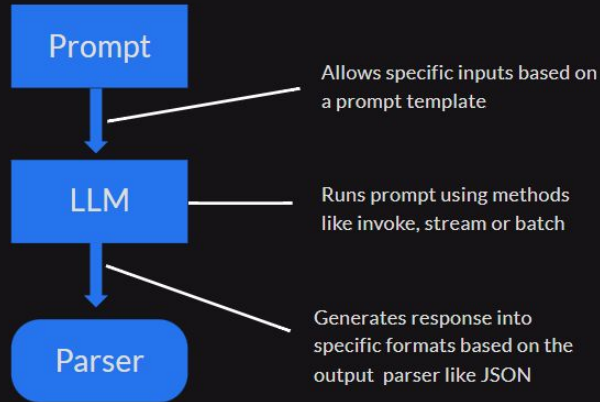


What is an Output Parser?

- Output parsers are responsible for taking the output of an LLM and transforming it to a more consumable format
- Very useful when we want to get structured or semi-structured outputs to store them or use them in some workflow

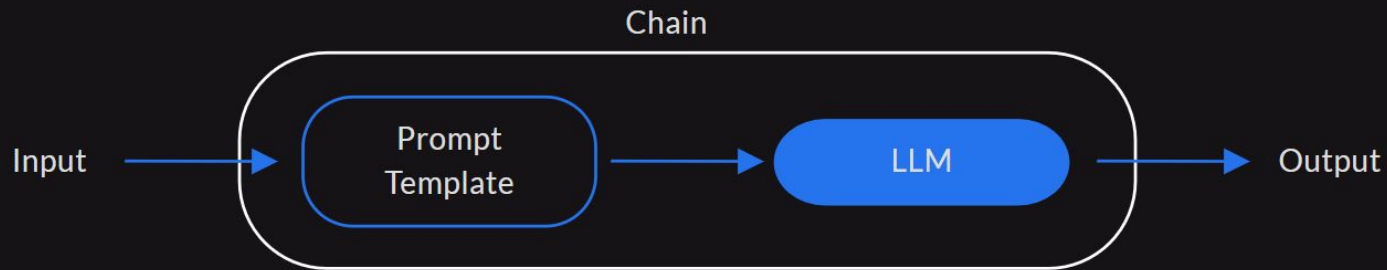


LangChain Expression Language - LCEL Chains



- In LCEL you can chain multiple steps using the overloaded vertical bar or pipe '|' operator
- Example LCEL chain on the left can be created as:
 - `chain = Prompt | LLM | Parser`
 - This signifies the prompt flows into the LLM which generates the response which is formatted using the Parser rules

Simple LLM Chain with LCEL



What is a Conversation?

Explain to me Generative AI in 3 bullet points

HumanMessage

- **Creation of New Content:** Generative AI uses machine learning models to create new data similar to existing data. It can generate text, images, music, and other types of content by learning patterns from large datasets.
- **Advanced Neural Networks:** It leverages advanced neural networks, such as Generative Adversarial Networks (GANs) and Transformer models (like GPT-4), to produce realistic and high-quality outputs that mimic human creativity and intelligence.
- **Wide Applications:** Generative AI has diverse applications, including content creation, design, entertainment, drug discovery, and more, providing innovative solutions across various industries by automating and enhancing creative processes.

AIMessage

Conversation History
or Memory

Now do the same for Deep Learning

HumanMessage

- **Hierarchical Learning:** Deep learning uses neural networks with multiple layers (deep neural networks) to automatically learn hierarchical representations of data. Each layer extracts increasingly complex features from the raw input data.
- **State-of-the-Art Performance:** It achieves state-of-the-art performance in various tasks such as image and speech recognition, natural language processing, and autonomous driving by leveraging large amounts of data and computational power.
- **End-to-End Training:** Deep learning models are trained end-to-end, meaning they can learn directly from raw data to final output without the need for manual feature extraction, allowing for more accurate and efficient processing and prediction.

AIMessage

Now do the same for Machine Learning

HumanMessage

New Prompt or Query

Need for Conversational Memory

With Conversational Memory

I am interested in integrating LLMs with external knowledge

LLMs are great at generating human-like text. Yet, integrating external knowledge can enhance their capabilities even more.

What are the different possible methods for doing this?

You could use pre-existing knowledge graphs, allow LLMs access to tools like APIs, or retrieval augmentation with vector DBs!

..... Conversation History

Interesting! What was it I wanted to know about?

You were interested in integrating LLMs with external knowledge.

Without Conversational Memory

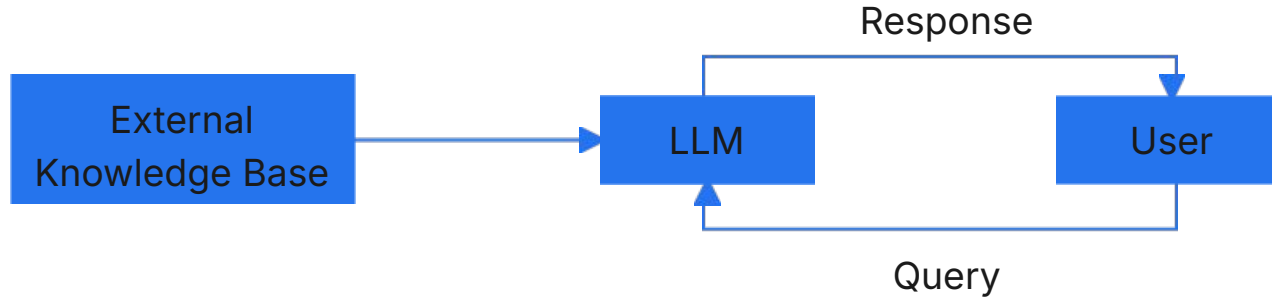
(No conversation history is stored)

..... Conversation History

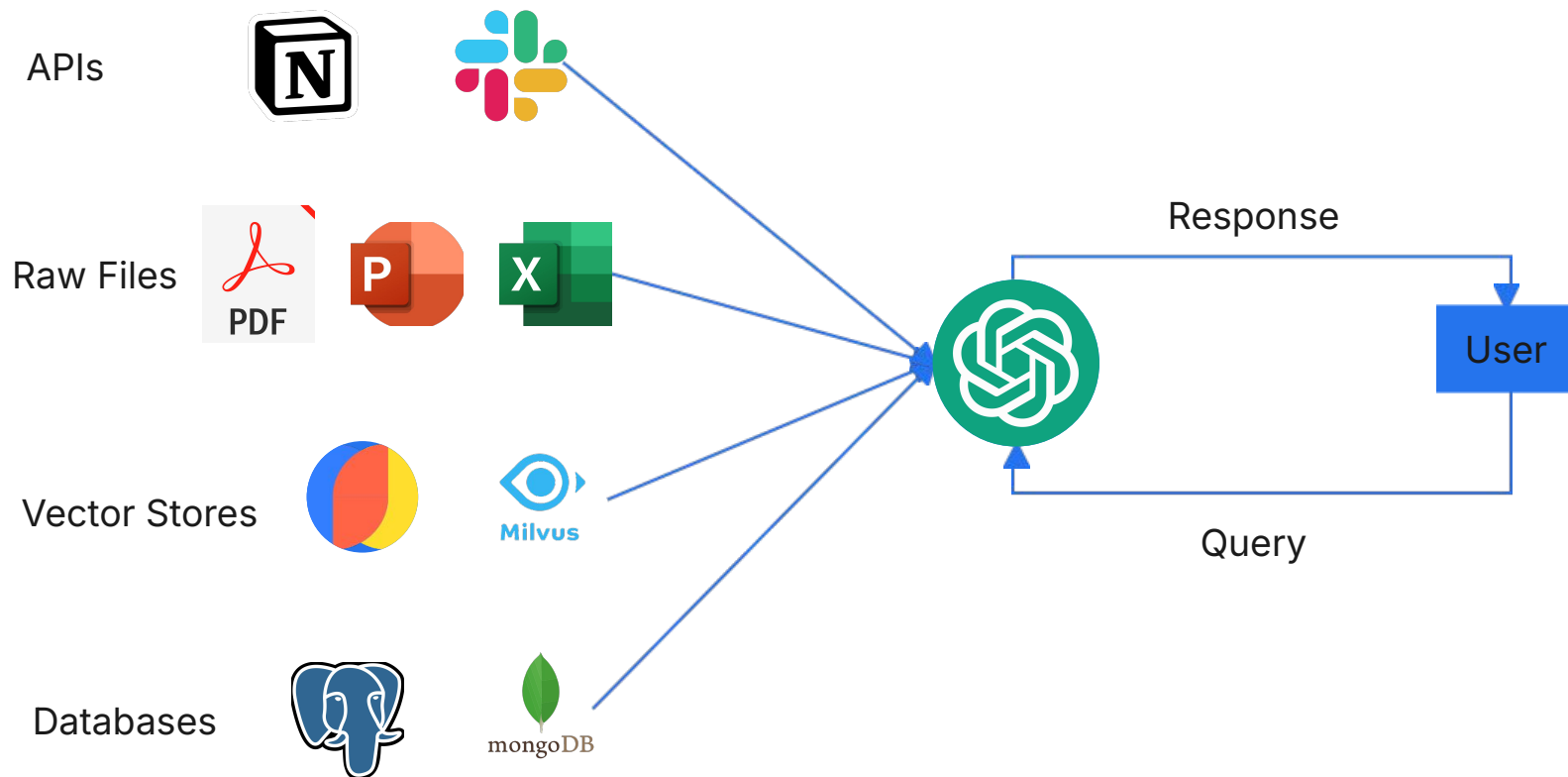
Interesting! What was it I wanted to know about?

Sorry, I have no idea what you're talking about!

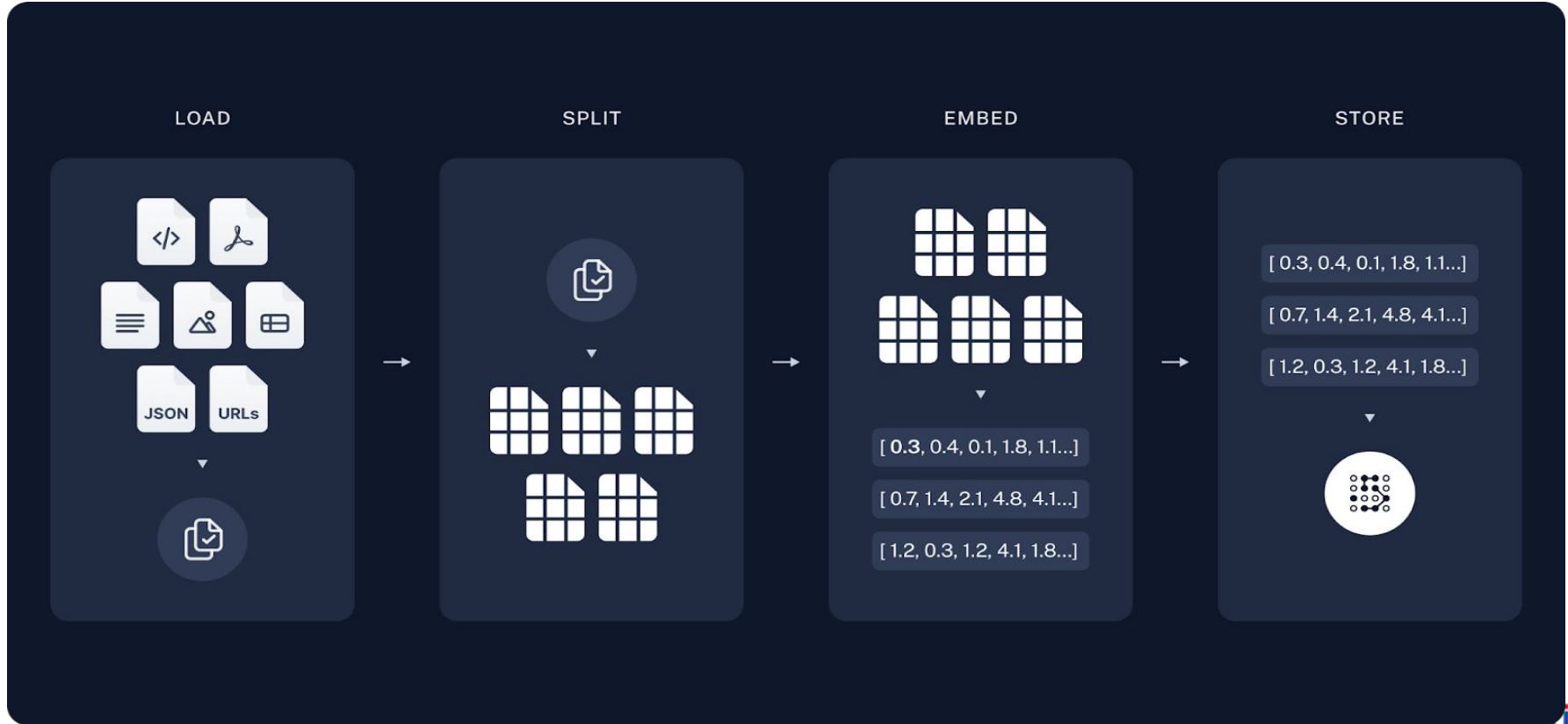
What is RAG?



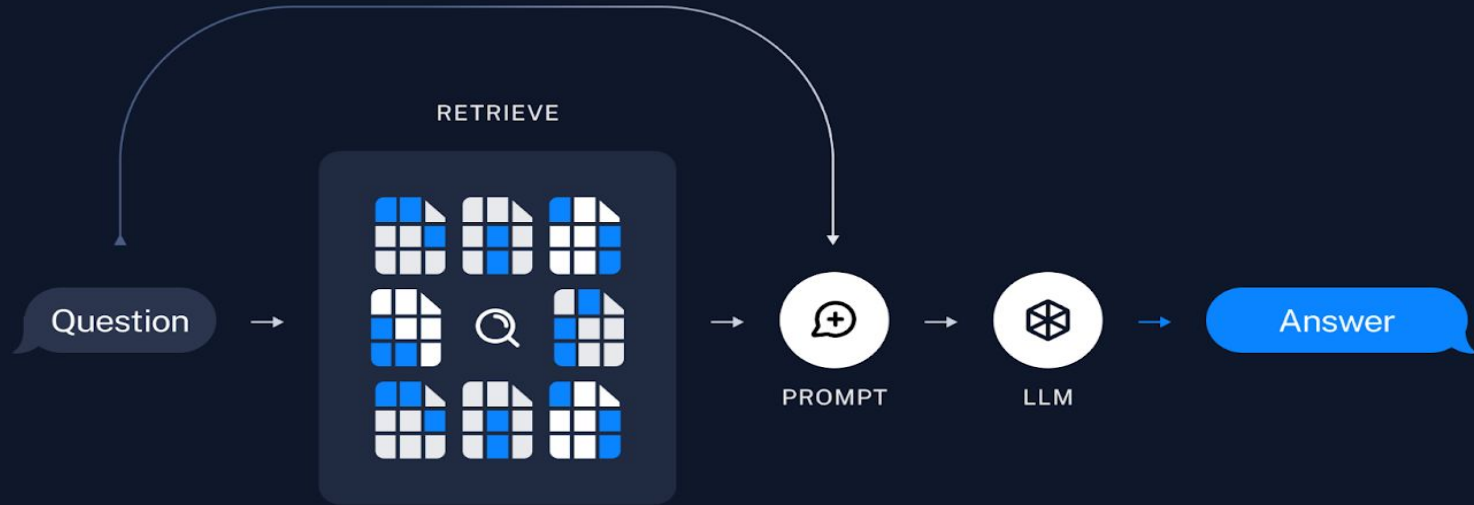
What is RAG?



RAG System Architecture

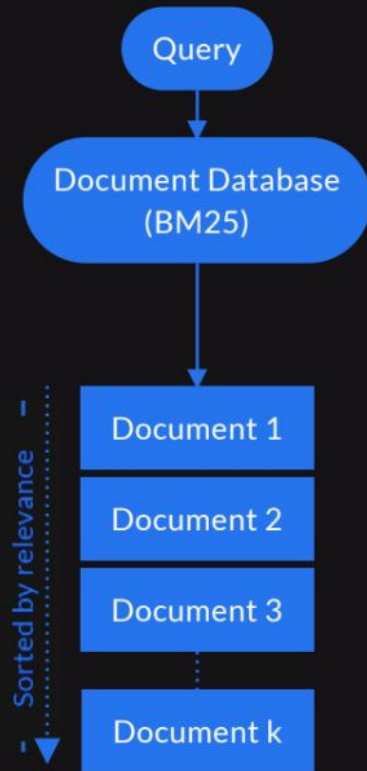


RAG System Architecture

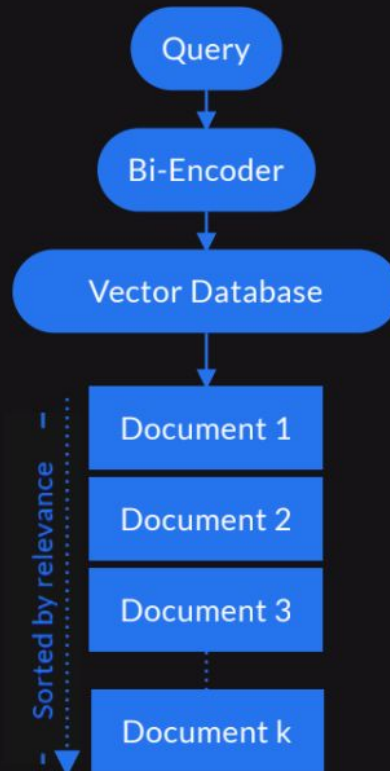


Types of Search Mechanisms in Vector DBs

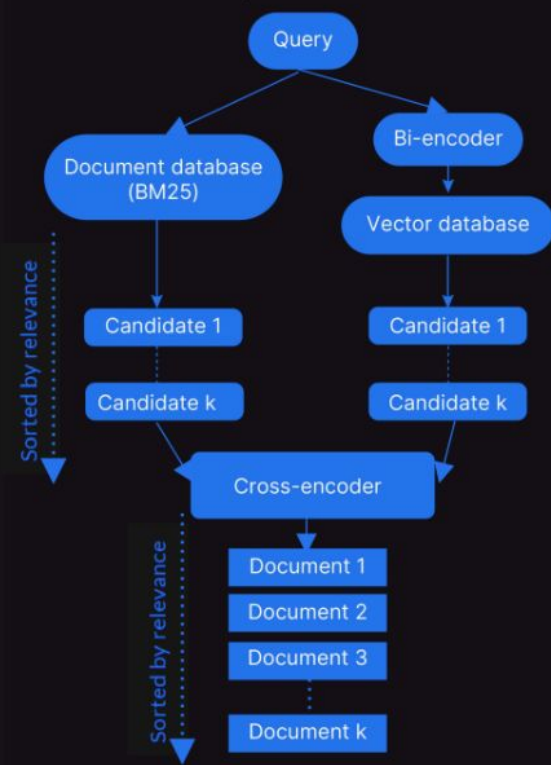
1. Keyword Search



2. Vector Search



3. Hybrid Search



Embedding Similarity - Useful to find similar documents

