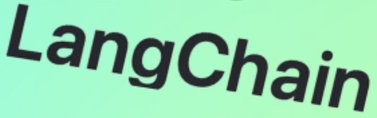
LLM Based Medical Support Chatbot using LangGraph



LangGraph

LCEL

LangChain

Prompt Engineering

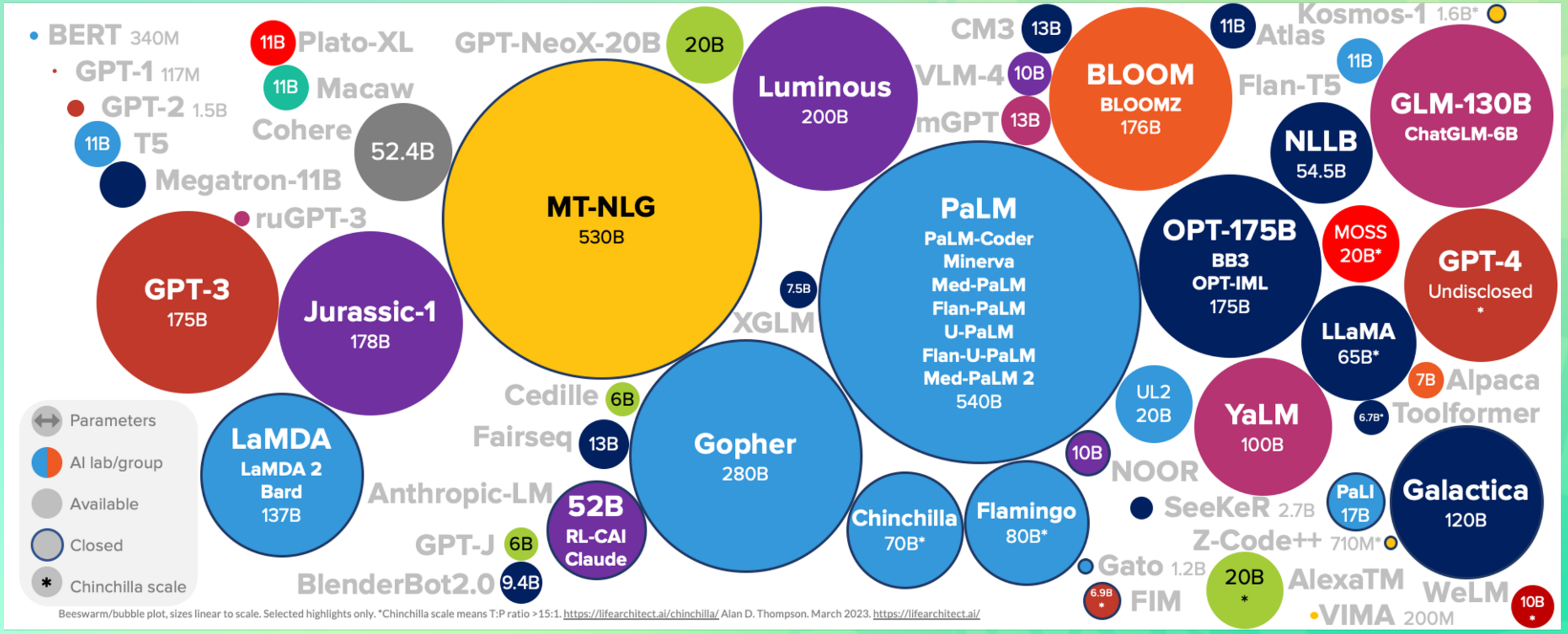
LLaMa-3-70B

TavilySearch

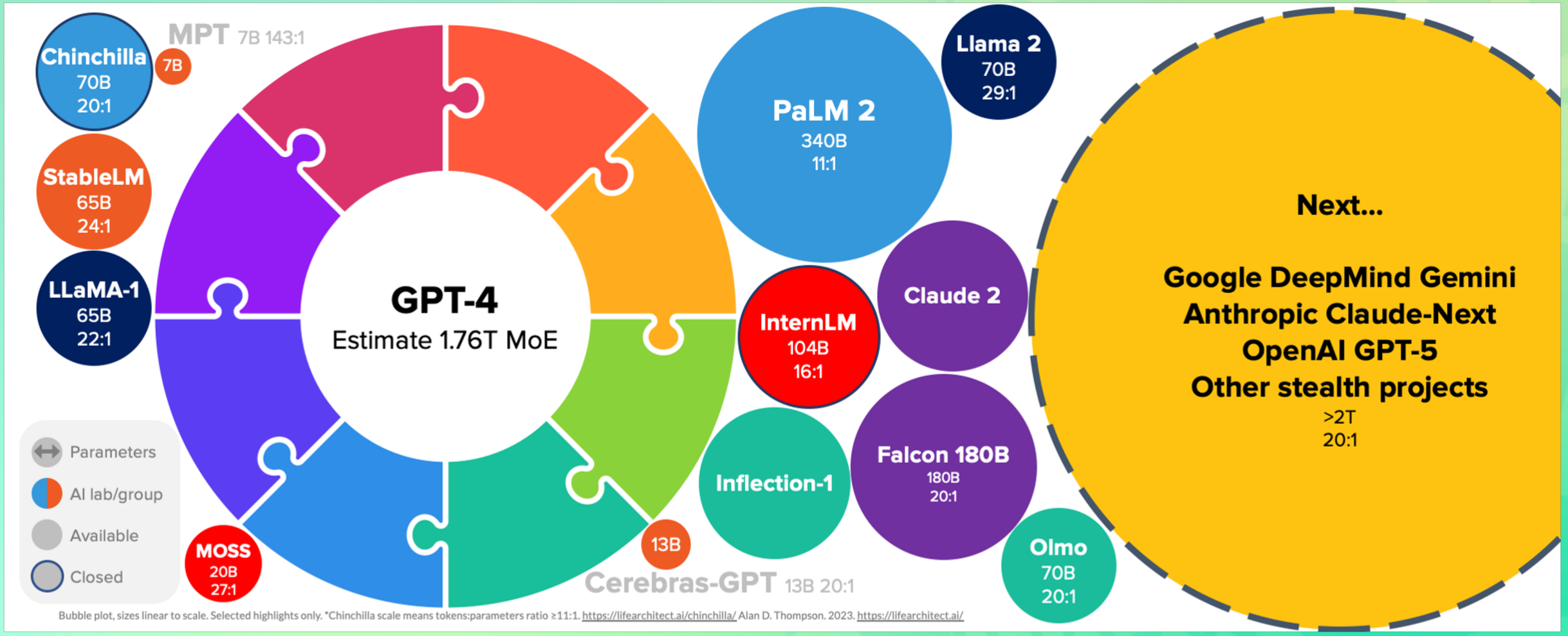
FAISS

Pydantic

What and Why of LLMs

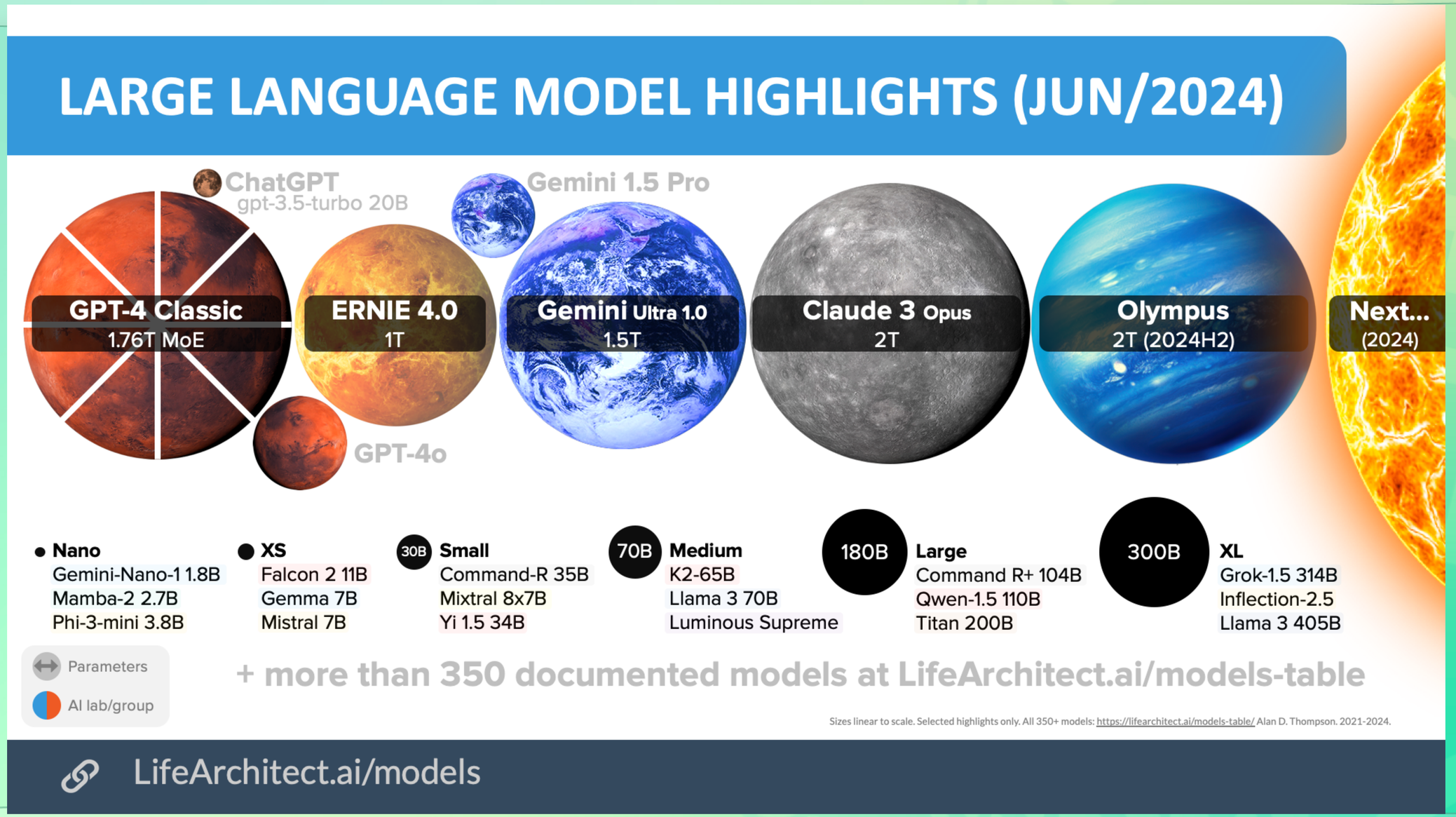


What and Why of LLMs

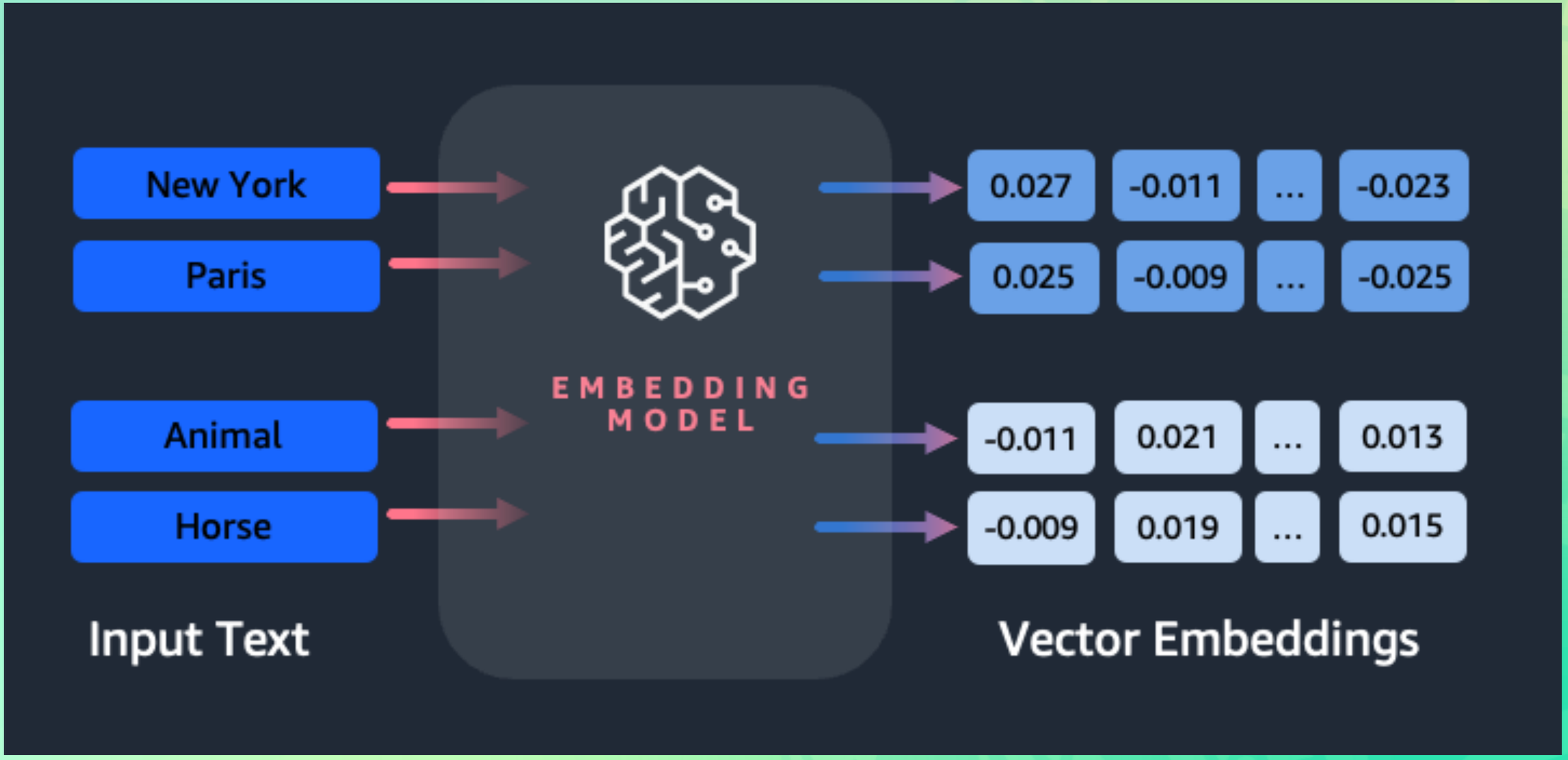


https://lifearchitect.ai/models/

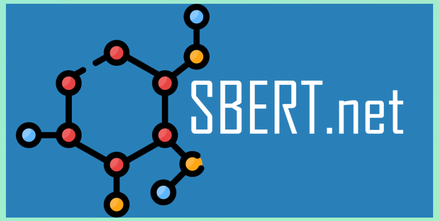
What and Why of LLMs



Embeddings


New York
Paris
Animal
Horse
EMBEDDING MODEL
0.027
-0.011
...
-0.023
0.025
-0.009
...
-0.025
-0.011
0.021
...
0.013
-0.009
0.019
...
0.015
Input Text
Vector Embeddings

Embeddings



Rank	Model	Model Size (Million Parameters)	Memory Usage (GB, fp32)	Embedding Dimensions	Max Tokens	Average (56 datasets)	Classification Average (12 datasets)	Clustering Average (11 datasets)	PairClassification Average (3 datasets)	Rera Aver (4 data
1	NV-Embed-v1	7851	29.25	4096	32768	69.32	87.35	52.8	86.91	60.5
2	voyage-large-2-instruct			1024	16000	68.28	81.49	53.35	89.24	60.0
3	Linq-Embed-Mistral	7111	26.49	4096	32768	68.17	80.2	51.42	88.35	60.2
4	SFR-Embedding-Mistral	7111	26.49	4096	32768	67.56	78.33	51.67	88.54	60.6
5	gte-Qwen1.5-7B-instruct	7099	26.45	4096	32768	67.34	79.6	55.83	87.38	60.1
6	voyage-lite-02-instruct	1220	4.54	1024	4000	67.13	79.25	52.42	86.87	58.2
7	GritLM-7B	7242	26.98	4096	32768	66.76	79.46	50.61	87.16	60.4
8	e5-mistral-7b-instruct	7111	26.49	4096	32768	66.63	78.47	50.26	88.34	60.2

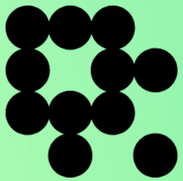
https://huggingface.co/spaces/mteb/leaderboard

Vector Stores

Weaviate

ChromaDB

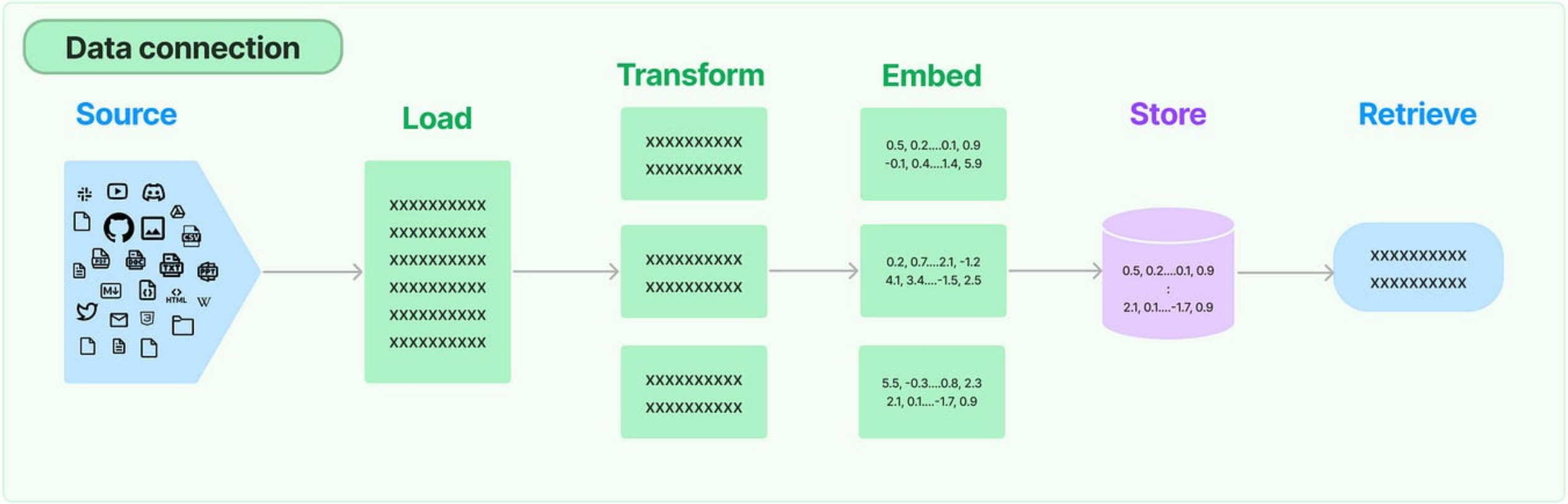
Milvus



LanceDB

PostgreSQL

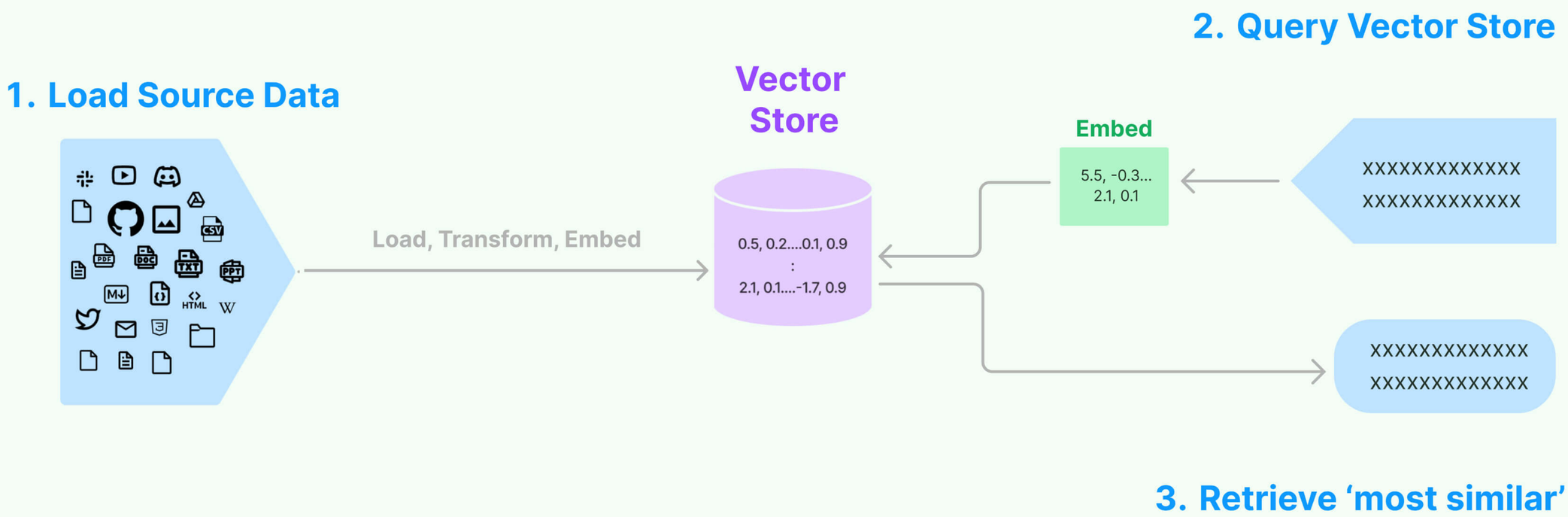
Vector Stores



https://python.langchain.com/v0.1/docs/modules/data_connection/

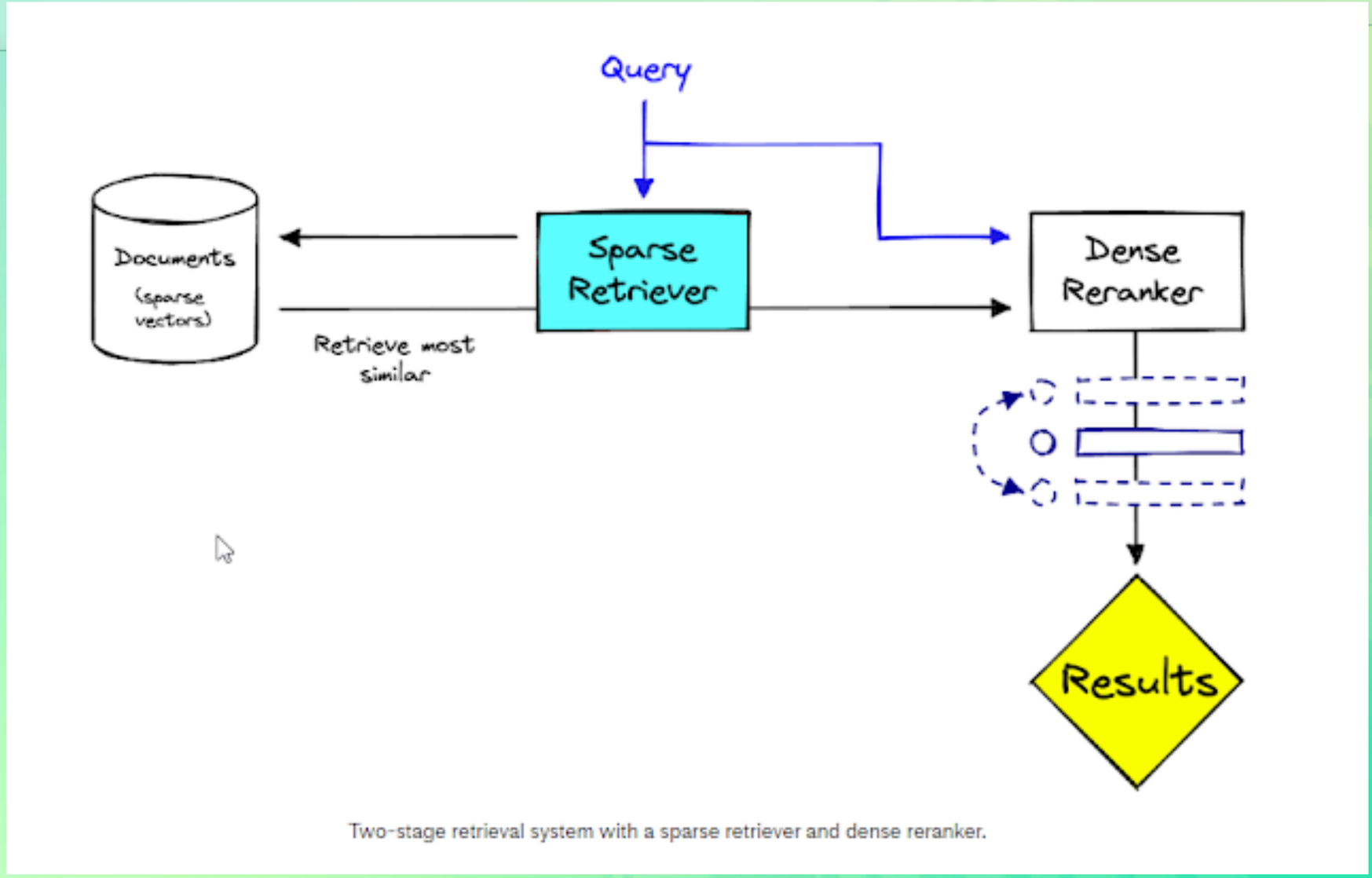
Retrievers

Vector Stores



https://python.langchain.com/v0.1/docs/modules/data_connection/vectorstores/

Retrievers



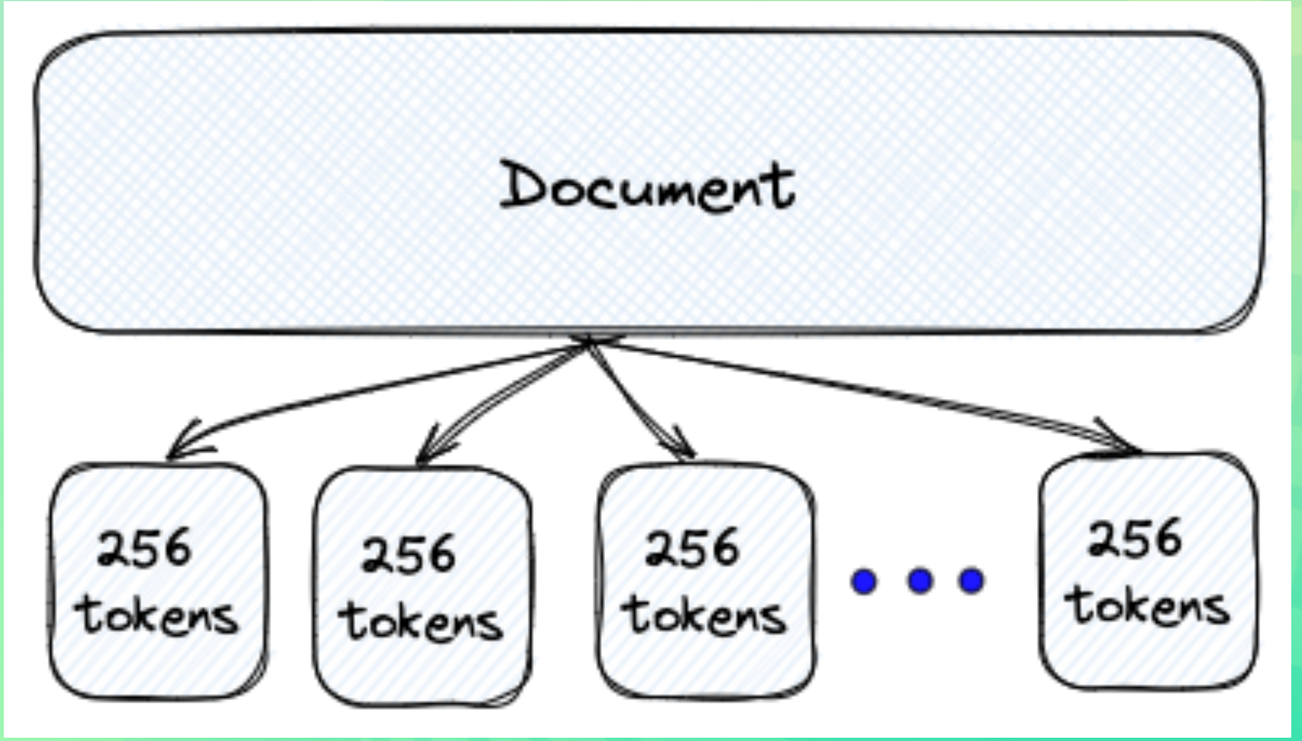
Two-stage retrieval system with a sparse retriever and dense reranker.

https://aarontay.medium.com/boolean-vs-keyword-lexical-search-vs-semantic-keeping-things-straight-95eb503b48f5

Chunking and Why

1. LLMs have limited input size (Context Window)

2. We only want the Relevant Pieces of Information



https://app.iki.ai/content/link/1705592

What are LLMs Good and Bad At?

LLMs are Good at

- Text Generation
- Reasoning
- Summarization
- Paraphrasing
- Data Extraction from Context

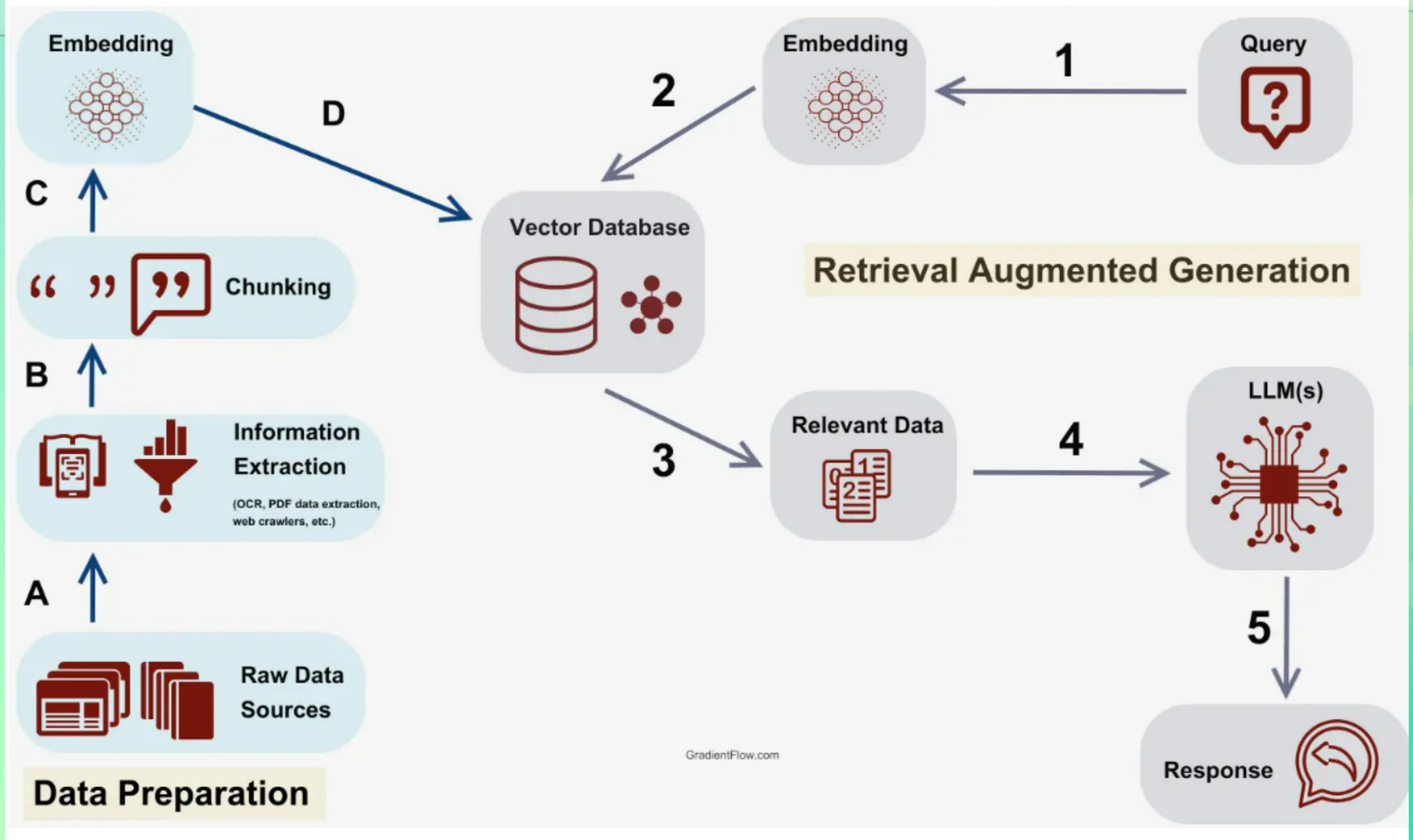
LLMs are Bad at

- Knowing about New Data
- Knowing about Our Data
- Hallucination
- Remembering Long-tail Information

What are LLMs Good and Bad At?

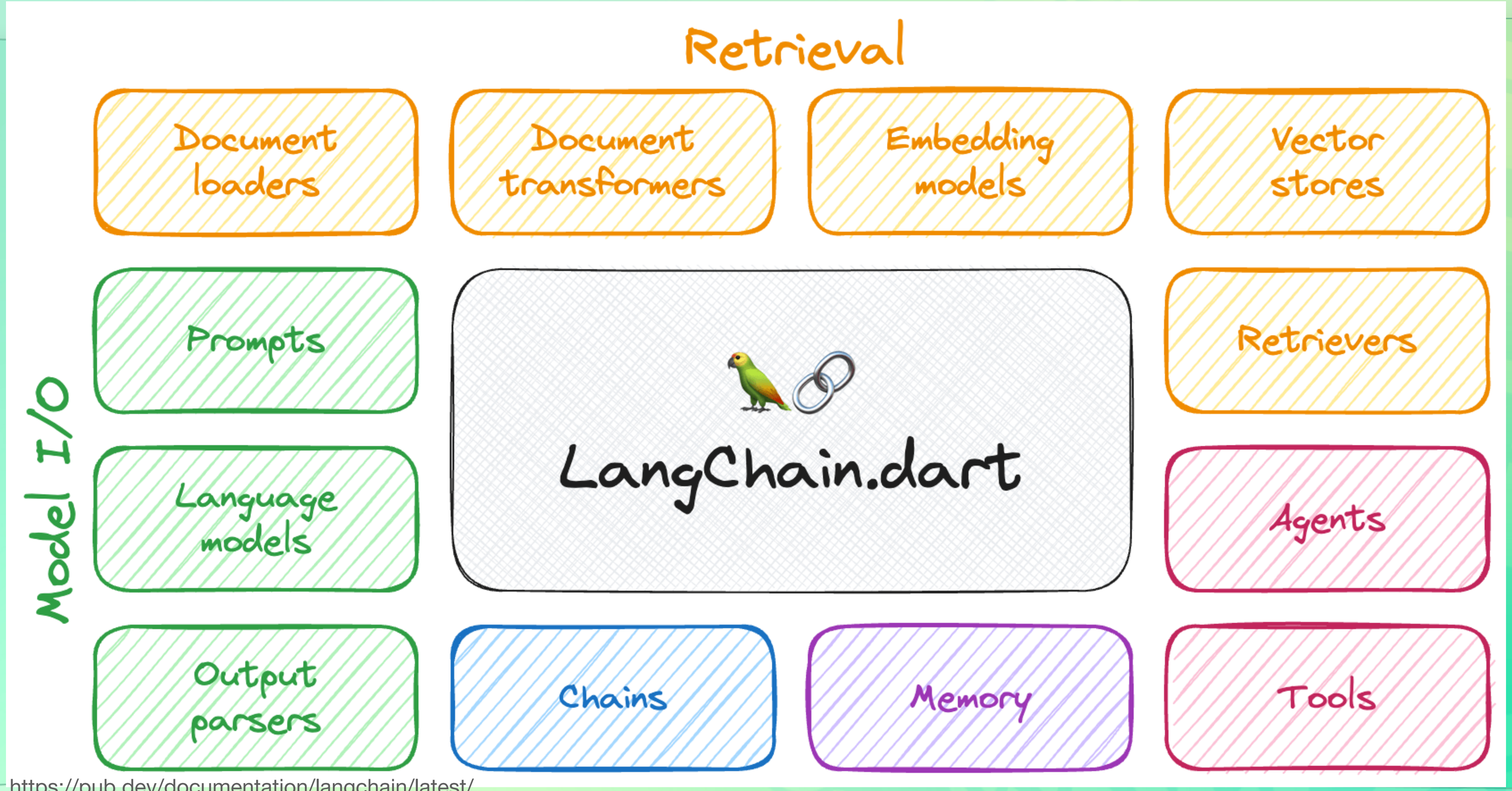
1. Our Chatbot must have the latest information
2. Our Chatbot must know about our Business Specific Data
3. Our Chatbot should be trustworthy

What is RAG?



https://www.linkedin.com/pulse/what-rag-why-should-you-care-aymen-noor-eyymf/

What is LangChain?

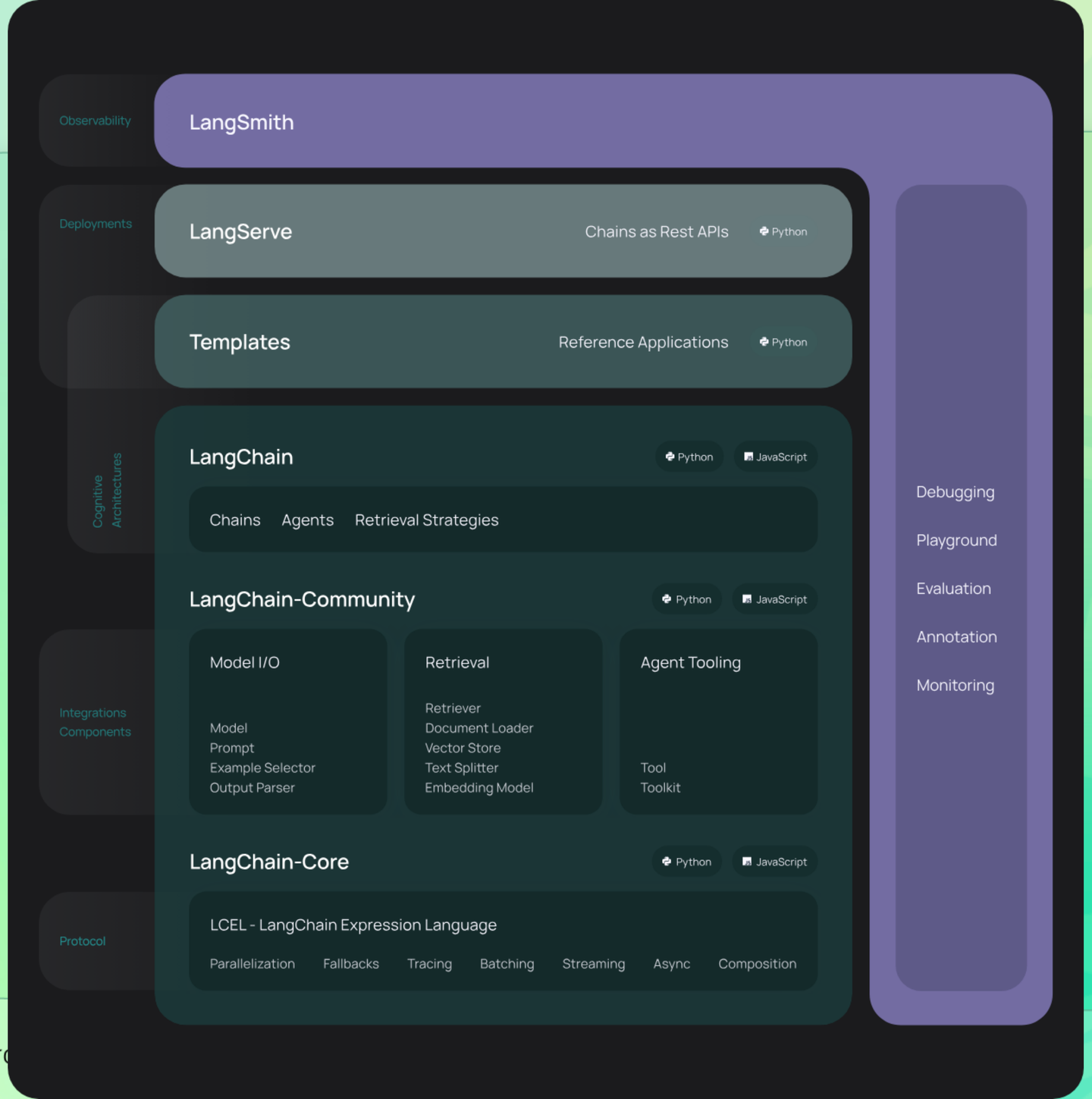


https://pub.dev/documentation/langchain/latest/

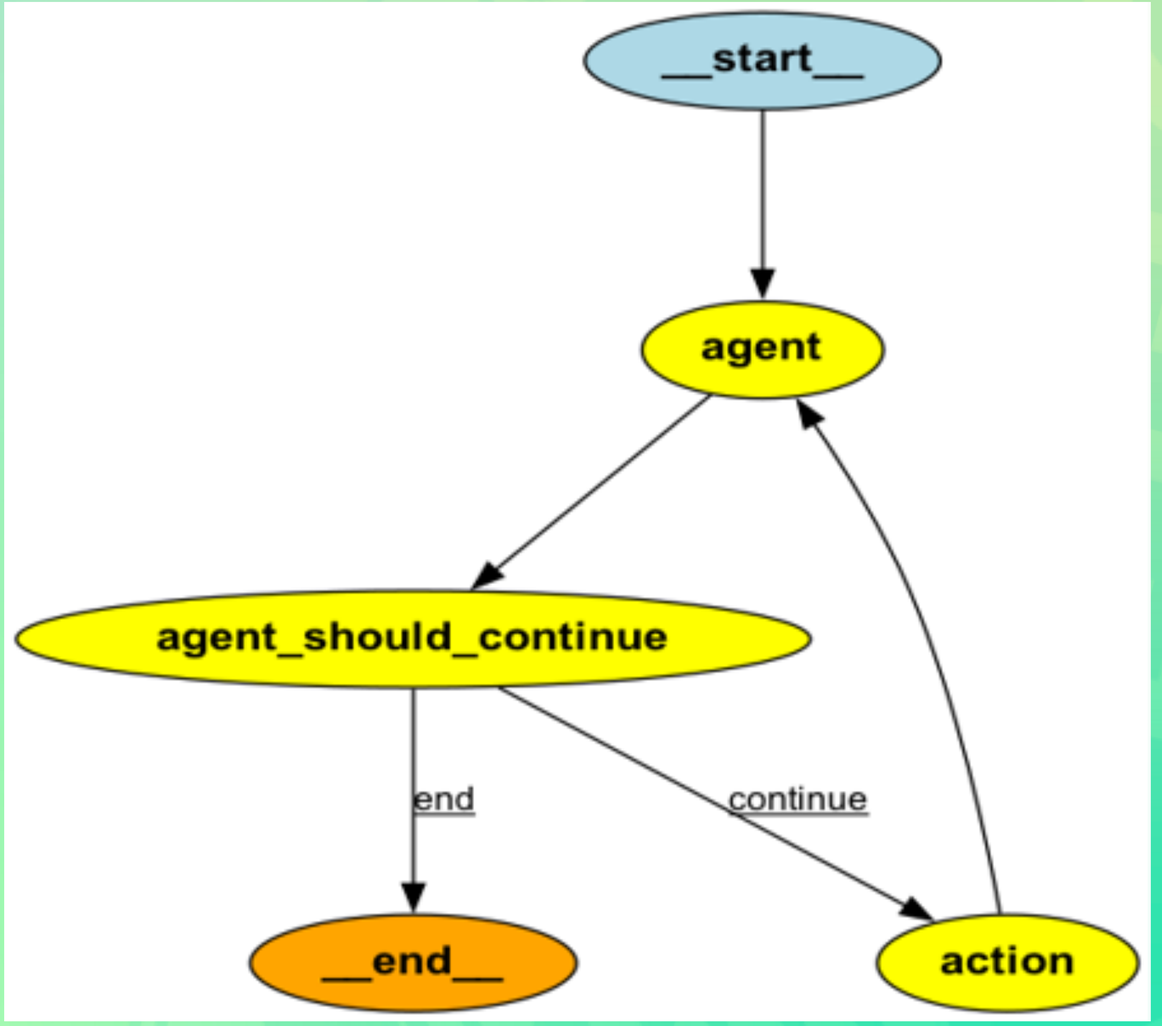
What is LangChain?

LangChain

https://python.langchain.com/v0.2/docs/introduction/



What is LangGraph?



https://blog.gopenai.com/conditional-edge-and-cycle-in-langgraph-explained-da4a112bf1ea

What is LangGraph?

```python
class AgentSate(TypedDict):
    """The dictionary keeps track of the data required by the various nodes in the graph"""

    query: str
    chat_history:list[BaseMessage]
    generation: str
    documents: list[Document]
```