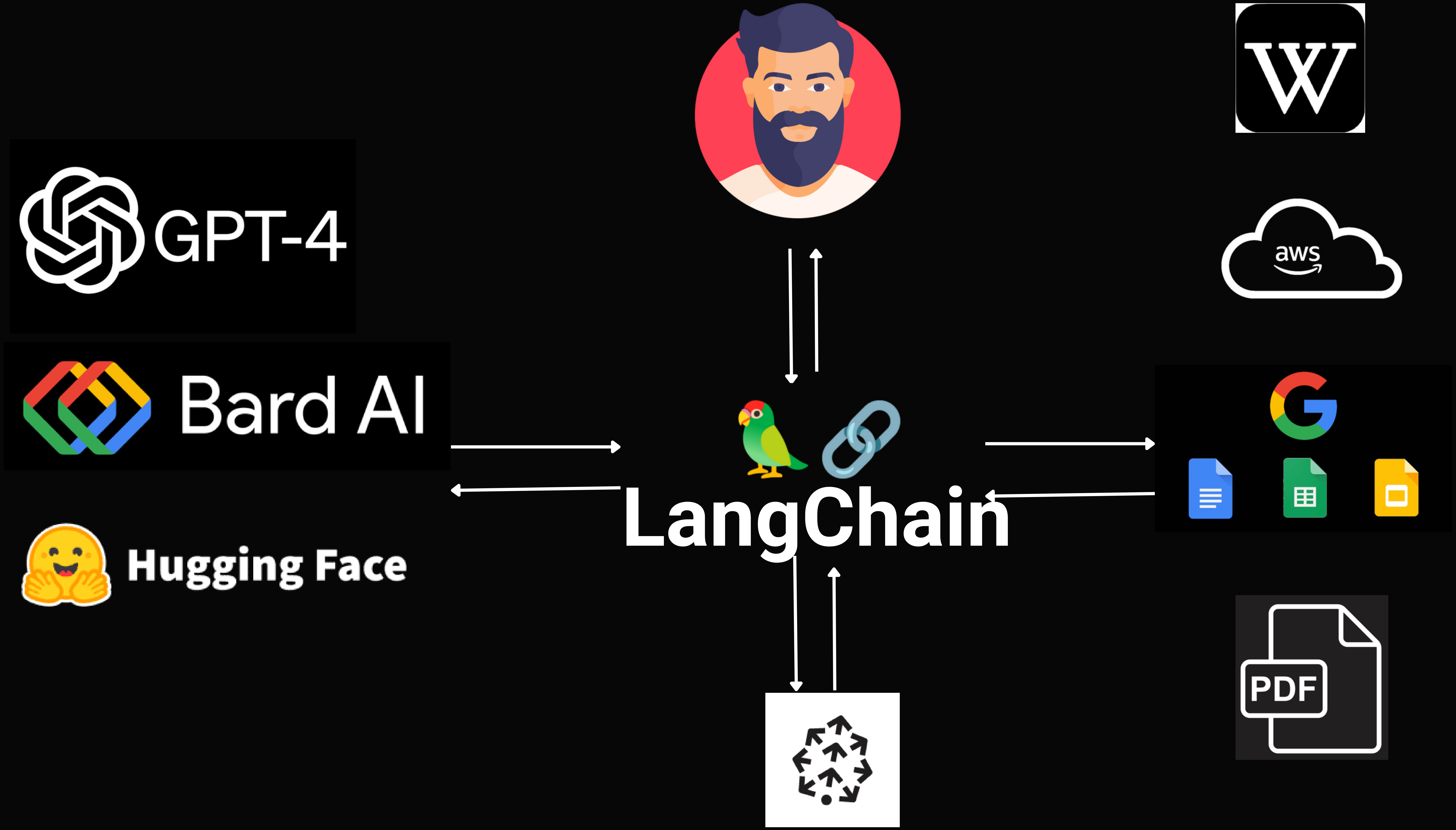
Q&A System on Private Documents Using OpenAI, Pinecone and LangChain (RAG)

GPT-4

🦜🔗 LangChain

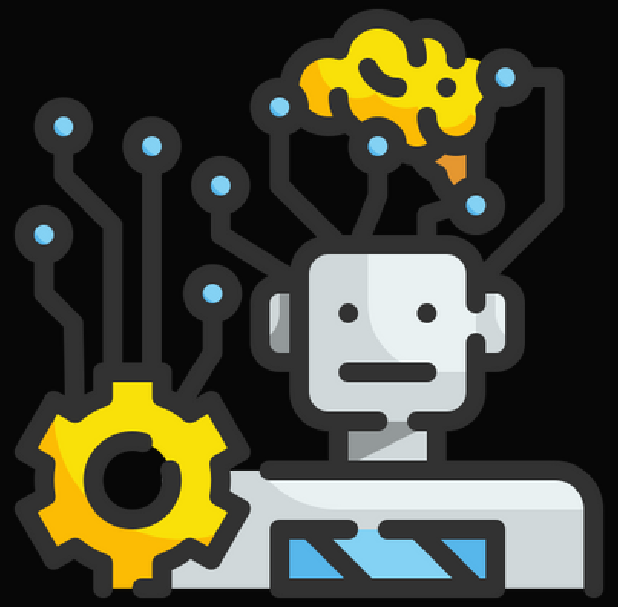
PDF



LangChain

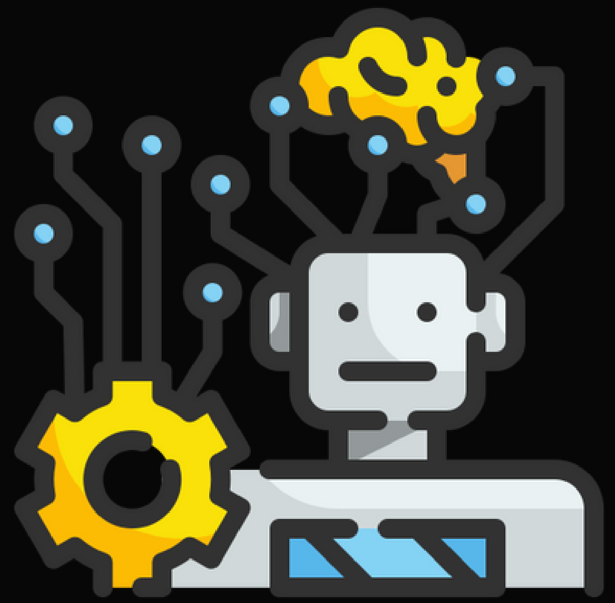
How can LLMs learn new knowledge?

1. Fine-tuning on a training set
2. Model inputs



How can LLMs learn new knowledge?

1. Fine-tuning on a training set
2. Model inputs



The recommended approach is to use model inputs with embedded-based search.

Question-Answering Pipeline

1. Prepare the document (once per document)

a) Load the data into LangChain Documents.

b) Split the documents into chunks.

c) Embed the chunks into numeric vectors.

d) Save the chunks and the embeddings to a vector database.

Question-Answering Pipeline

2. Search (once per query)

a) Embed the user's question.

b) Using the question's embedding and the chunk embeddings, rank the vectors by similarity to the question's embedding. The nearest vectors represent chunks similar to the question.

Question-Answering Pipeline

3. Ask (once per query)

a) Insert the question and the most relevant chunks into a message to a GPT model.

b) Return GPT's answer.

Summarization using **refine**

Step 1

summarize(chunk #1) => summary #1

Step 2

summarize(summary #1 + chunk #2) => summary #2

Step 3

summarize(summary #2 + chunk #3) => summary #3

....

Step n

summarize(summary #n-1 + chunk #n) => **final summary**

Summarization using refine

Prons:

- uses a more relevant context (better summarization)
- less lossy than map_reduce

Cons:

- it requires many more calls to the LLM
- the calls are not independent and can not be parallelized

Chunking is the process of breaking down large pieces of text into smaller segments.

Chunking is the process of breaking down large pieces of text into smaller segments.

It’s an essential technique that helps optimize the relevance of the content we get back from a vector database.

As a rule of thumb, if a chunk of text makes sense without the surrounding context to a human, it will make sense to the language model as well.

As a rule of thumb, if a chunk of text makes sense without the surrounding context to a human, it will make sense to the language model as well.

Finding the optimal chunk size for the documents in the corpus is crucial to ensure that the search results are accurate and relevant.

RAG - Retrieval Augmented Generation

🦜🔗

LangChain

RAG helps overcome knowledge limits, makes answers more factual, and lets the model handle complex questions.